12-2020

Computational behavioral analytics: estimating psychological traits in foreign languages.

Kristopher Wayne Reese
University of Louisville

COMPUTATIONAL BEHAVIORAL ANALYTICS: ESTIMATING PSYCHOLOGICAL TRAITS IN FOREIGN LANGUAGES

By
Kristopher Wayne Reese
M.S., University of Louisville, 2011
B.S., Hood College, 2009
B.A., Hood College, 2009

A Dissertation
Submitted to the Faculty of the
J.B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Computer Science and Engineering

Computer Science and Engineering
University of Louisville
Louisville, Kentucky

December 2020

COMPUTATIONAL BEHAVIORAL ANALYTICS: ESTIMATING PSYCHOLOGICAL TRAITS IN FOREIGN LANGUAGES

By

Kristopher Wayne Reese
M.S., University of Louisville, 2011
B.S., Hood College, 2009
B.A., Hood College, 2009

Dissertation approved on

October 23, 2020

by the following dissertation Committee:

Dissertation Director
Adel Elmaghraby

Roman Yampolskiy, CSE

Dar-Jen Chang, CSE

Adrian Lauf, CSE

Michael Losavio, Criminal Justice

DEDICATION

For my wife, Danielle, and my son, Ezra. Thank you for all of your support and love through all of the work that went into this dissertation.

ACKNOWLEDGMENTS

I'd like to thank all the people who have helped me throughout the dissertation process including my advisor, Dr. Adel Elmaghraby, who stood with me through the long process. Thank to the various Government colleagues who provided assistance in algorithm development and for helping to teach me various statistical techniques used in analysis. I'd also like to acknowledge my colleagues at the University of Florida, especially Dr. Damon Woodard, who helped with aspects of the authorship attribution experiments explored in this dissertation.

This dissertation makes use of various Free and Open Source Python 3 libraries for Natural Language Processing, Machine Learning algorithm development, and data analysis.

The research made use of data provided by the Linguistic Data Consortium, the University of Cambridge, and North Carolina Agricultural & Technical State University.

ABSTRACT

COMPUTATIONAL BEHAVIORAL ANALYTICS: ESTIMATING PSYCHOLOGICAL TRAITS IN FOREIGN LANGUAGES

Kristopher Wayne Reese

December 12, 2020

The rise of technology proliferating into the workplace has increased the threat of loss of intellectual property, classified, and proprietary information for companies, governments, and academics. This can cause economic damage to the creators of new IP, companies, and whole economies. This technology proliferation has also assisted terror groups and lone wolf actors in pushing their message to a larger audience or finding similar tribal groups that share common, sometimes flawed, beliefs across various social media platforms. These types of challenges have created numerous studies in psycholinguistics, as well as commercial tools, that look to assist in identifying potential threats before they have an opportunity to conduct malicious acts. This has led to an area of study that this dissertation defines as "Computational Behavioral Analytics."

A common practice espoused in various Natural Language Processing studies (both commercial and academic) conducted on foreign language text is the use of Machine Translation (MT) systems before conducting NLP tasks. In this dissertation, we explore three psycholinguistic traits conducted on foreign language text. We explore the effects (and failures) of MT systems in these types of psycholinguistic tasks in order to help push the field of study into a direction that will greatly improve the efficacy of such systems.

Given the results of the experimentation in this dissertation, it is highly recommended to avoid the use of translations whenever the greatest levels of accuracy are necessary, such as for National Security and Law Enforcement purposes. If translations must be used for any reason, scientist should conduct a full analysis of the impact of their chosen translation system on their estimates to determine which traits are more significantly affected. This will help ensure that analysts and scientists are better informed of the potential inaccuracies and change any resulting decisions from the data accordingly.

This dissertation introduces psycholinguistics and the benefits of using Machine Learning technologies in estimating various psychological traits, and provides a brief discussion on the potential privacy and legal issues that should be addressed in order to avoid the abuse of such systems in Chapter I. Chapter II outlines the datasets that are used during the experimentation and evaluation of the algorithms. Chapter III discusses each of the various implementations of the algorithms used in the three psycholinguistic tasks - Affect Analysis, Authorship Attribution, and Personality Estimation. Chapter IV discusses the experiments that were run in order to understand the effects of MT on the psycholinguistic tasks, and to understand how these tasks can be accomplished in the face of MT limitations, including rationale on the selection of the MT system used in this study. The dissertation concludes with Chapter V, providing a discussion and speculating on the findings and future experimentation that should be done.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

"Computational Behavioral Analytics" (CBA) is a subset of computational social sciences focused on using Machine Learning and Artificial Intelligence to estimate psychological and cognitive features to understand and add context to the behaviors, activities, and motivations of the human behind observed data. Our daily communications have increasingly shifted to written text on technological tools - text messaging on cell phones, online messaging services, email services, or social media platforms. Understanding the people behind these tools is a modern challenge that businesses and government agencies are tasked with to identify potential indicators of threats to their intellectual property and their task to protect their people.

The increase of technology proliferating in the workplace has increased the threat of intellectual property loss, classified information, and proprietary information for companies, governments, and academics. This rise in technologies can cause economic damage to the creators of new Intellectual Properties, companies, and whole economies. This technology proliferation throughout our culture has also assisted terror groups, lone-wolf actors, and nation-states in pushing their messages to broader audiences, the spreading of misinformation, or finding similar tribal groups that share common, often intolerant, beliefs across various social media platforms. As law enforcement, intelligence agencies, and businesses seek out possible threats, the amount of data available for psychologists to review grows beyond the number of psychologists

available to each organization. Psychologists provide methods and techniques for understanding the behaviors of malicious actors. These psychologists provide methods for understanding the behaviors of malicious actors. Research and Development of new tools have provided commercial applications that estimate psychological traits and monitor for potential threats, such as IBM's QRadar [1], for detecting insider threats; IBM's Personality Insights [2], for estimating Personality Traits; and Crystal Knows [3], for estimating group dynamics using DISC profiles.

As our economies and national security threats become globalized, the use of foreign language threats to businesses and countries rises, many of the tools are reliant on Machine Translations to help in understanding the messages conveyed by people. Organizations have done little or no research on the impact of Machine Translations on these psychological tools when attempting to understand human behavior. This dissertation looks at several cutting edge machine learning models for estimating emotional intensity, personality traits, and authorship styles and determine how these models are affected by machine translations.

In this chapter of the thesis, we introduce the rich history between Machine Learning and Cognitive and Psychological Sciences, as well as existing studies in psycholinguistics attempting to understand psychological traits. The chapter also delves into the various reasons such a field exists, focusing on the needs of businesses and government agencies. It mainly focuses on the impact such a field could have on Insider Threats to organizations, Counterproductive work behaviors, and identification of potential radicalization methods. A discussion of some of the potential privacy issues that could arise in "Computational Behavioral Analytics" concludes the chapter. Chapter III explores each of the machine learning models in greater detail, giving an introduction to each of the psychological theories underpinning the work. It also discusses each of the machine learning models employed in greater detail. Chapter II explores the datasets used in training in greater detail and discuss methods for

expanding the datasets in the future, offering methods for collection used in future studies. Chapter IV discusses the experimental design and results of the experiments to understand the impact of Machine Translation on the algorithms to determine the viability of using translations for estimating psychological traits. The dissertation concludes by discussing the results of the experiments and future research.

1 Computational Behavioral Analytics

The field is rife with possibilities to open up an understanding of people through the use of psycholinguistics, cognitive studies that mathematically model cognitive functions, or the use of computer vision for estimating psychological or cognitive functions hidden in collected data. While there has been some work in bringing Machine Learning into psychology, the field is still relatively nascent. These studies often suffer from a relatively low number of properly ground-truthed datasets. At the same time, these studies challenge researchers to build sufficiently large human-derived datasets to make Machine Learning viable in the field. Many datasets rely on psychologists' expertise to train models, but this method is untenable due to the limited availability of psychologists and the amount of time needed to train large datasets sufficiently. Ideally, the field would rely on subjects taking various validated psychological or cognitive assessments, but this requires significant effort to collect data. However, building datasets using these validated instruments disperses the time needed by experts to the subjects themselves, making the collection of more massive datasets viable for use in Machine Learning models.

Using either form of dataset introduced some form of bias into a machine learning system, making the understanding of an individual more challenging. Using datasets created by experts, the Machine Learning models determine the traits based on how the experts analyzed the persona that a subject in portraying to the experts. The use of psychological instruments can suffer from unintentional or intentional gaming

of the instruments in an attempt to hide their real personality. However, psychologists have long dealt with the problem by adding inversions of questions to identify those subjects who may be attempting to game the instruments. Using psychological instruments for building datasets allows Machine Learning models to estimate traits based on how the subject answers the questions.

Despite the challenges and the current lack of sizeable datasets, studies into the field of "Computational Behavioral Analytics" continue to occur, and continued progress adds to the understanding of various aspects of human psychology. However, significant effort needs to occur to make the field as cross-disciplinary as it ought to be. Many studies often occur without consulting researchers outside of their fields of study. Various studies exist in computer science that delves into psychological traits, but these often lack any consultation of psychologists in helping to interpret the data. In the same vein, the use of Machine Learning in Psychology has taken off, often limited to older models that might be inferior to the tasks at hand. Consultation of Computer Scientist and Mathematicians might increase model efficiency and accuracy.

Machine Learning & Psychology

Machine learning and cognitive science have a rich history together. The fields grew from a postwar movement away from the behaviorist paradigm that dominated cognition research during the first half of the twentieth-century [4]. Some of the preeminent names in Artificial Intelligence, such as Allen Newell and Herbert Simon, were cognitive scientists who developed new algorithms in AI in an attempt to model human processes. Up until the 1970s, Artificial Intelligence and Cognitive sciences often shared conferences where the two paths began to diverge. Since this schism, cognitive scientists and artificial intelligence researchers have shared methodologies, but often with very different goals. Jerome Feldman [4] identifies the differences between

the two fields as AI's search for an understanding of intelligent behaviors, whether applied to animals or artificial systems, whereas cognitive science has a more narrow goal of understanding human intelligence.

Because of the differences, Cognitive Sciences tend to be much more interdisciplinary, incorporating linguistics, psychology, and artificial intelligence using computation as the universal language that could assist in the understanding of the human mind [5]. It is only recently that artificial intelligence and machine learning have reached a point where it can begin to contribute as a common language for many different fields, including in the study of human behaviors. The recent advancement of Artificial Intelligence and Machine Learning is needed because of the increase in the use of technologies for communications.

According to Raconteur [6], humans will produce 463 Exabytes of data every day by 2025. Even in 2019, society is producing massive amounts of data per day: nearly 500 million tweets, sending 294 billion emails, 65 billion WhatsApp messages, and generating over 4 Petabytes of Facebook data, consisting of 350 million photos, 100 million hours of video content, and other social media postings on the platform [6]. Sifting through this data would require a massive number of psychologists to identify potential threats. In 2012, the American Psychological Association estimated that there are 106,500 licensed clinical psychologists in the United States [7]. Without increasing the number of psychologists, by 2025, each psychologist would have to go through 4.3 Petabytes of data per day in search of potential threats. This amount of data also assumes that all licensed clinical psychologists have dedicated their time to identify threats, which would take them away from their medical duties.

Despite this need for more clinical psychologists, there is a shortage of psychologists around the world, even within the medical facilities of Malaysia and Australia [8]. Artificial intelligence and machine learning can help psychologists and play a more significant role in the behaviorist paradigm for helping psychologists understand

the motivations, activities, and behaviors of people, especially in an ever-increasing virtual world.

2 Incipient and Enduring Threats

The expansion of technologies as a communications medium has raised incipient risks and threats to consumers. Cyber-Psychologists have identified exposure to sexually explicit material and child pornography [9], victimization through harassment and bullying [10], and internet addiction [11, 12, 13, 14] as some of the risks that modern consumers face on these technologies. The concept of internet-induced anonymity has also given rise to issues of deception in online environments, often assisted by the inability to verify information [15]. While many online platforms have taken steps to minimize deception that leads to significant financial or social losses, deception of identity is a growing social problem, even in online dating platforms [16]. While the magnitude of online dating deception is usually minor [17], online deception can sometimes lead to more extreme issues such as homicide, financial losses, or online predation.

Insider Threats

While technological developments create new and emerging threats, it has also reshaped ongoing and enduring threats posed against society. Both Psychology and Computer Scientists have identified Insider Threats as an issue for businesses. The term "Insider Threats" is a term often used by Computer Emergency Response Teams (CERTs). Carnegie Mellon's Insider Threat Team [18] has defined an insider threat as, "a current or former employee, contractor, or other business partners who has or had authorized access to an organization's network, system, or data and intentionally misused that access to negatively affect the confidentiality, integrity, or availability of the organization's information or information systems." [19] Psychologists have taken

a much broader definition and include the CERT definition of Insider Threats under the concept of "Counterproductive Work Behaviors." According to [20], "Counterproductive Work Behaviors (CWBs) include actions that harm organizations such as absenteeism, substance abuse on the job, and theft (i.e., organization CWBs), as well as harassment and aggressive acts directed towards coworkers (i.e., interpersonal CWBs)."

Insider threats have almost certainly always existed in companies; it is with new technologies that insider threats have been more able to steal Intellectual Property, trade secrets, or large quantities of classified information from government agencies. These acts damage businesses and whole economies and potentially threaten the financial and physical safety of citizens. Another enduring threat is terrorism. Today's Internet and communication tools have created a new landscape for groups to radicalize vulnerable people and incite violence towards those who share different beliefs.

The Commission on the Theft of American Intellectual Property [21] conservatively estimates the damages to the U.S. economy, caused by intellectual property theft, to be between $225 billion and $600 billion. These numbers are conservative because estimating the damages caused is challenging due to the ease of copying and distributing intellectual property because of technological advancements. In 2012, the U.S. Congress estimated the revenue loss to a company to be, on average, about $101.9 million a year[22].

According to findings of a 2018 Threat Report from Cybersecurity Insiders [23], most businesses feel vulnerable to insider threats, while 53% of surveyed businesses have confirmed insider attacks against their organization. In 2020, 68% of businesses have claimed that insider attacks are occurring more frequently, and the transition to cloud infrastructures has made identifying insider threats more difficult.

The Software Engineering Institute at Carnegie Mellon has created an Insider Threats team dedicated to researching Insider Threats at government agencies and

businesses. A 2012 study from the group [24], found several motivations that drive people to commit insider attacks, especially in the financial sector: revenge against the organization for perceived injustice done to the attacker [25]; those who believe that they own the Intellectual Property because of the work that they have put into it [26]; those who can recruit insiders to steal information that they may not have authorized access to, often for some larger purpose[26]; and financial gain caused by financial stress of the employee [25]. Despite the identification of these potential motivations, the group has not considered those individuals who may have similar motivations but never act on it during this study.

Numerous psychological experiments have explored the relationship between personality traits and counterproductive workplace behaviors (CWBs) [27, 28, 29, 30, 31, 20, 32]. In psychological studies, the organizational perspective (CWB-O) [33], defining CWB as acts that go against an organization's interests, including those activities that may involve verbal and physical attacks against coworkers. Other literature takes a more employee-centric perspective (CWB-I), often looking at CWBs as behaviors that harm or intend to harm others or organizations [34].

Studies often explore the relations between CWBs and the Five-Factor Model (FFM), though additional models, such as the HEXACO model, are not uncommon in literature. The FFM classifies personality traits into five categories: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness (the Five-Factor Model is discussed further in Chapter II. The primary difference between the FFM and the HEXACO model is the inclusion of a humility trait within the HEXACO model. Broader studies are still needed, but several studies have identified low Agreeableness and high neuroticism as correlated with Organizational CWBs (such as stealing and absenteeism), and low Agreeableness with Interpersonal CWBs (such as harassment and aggression) [30, 35]. Berry et al. [36] also found a strong correlation between low Agreeableness and low Conscientiousness related to overall CWBs. The correlation

of personality traits to CWBs appears to remain constant over several years. Le et al. [20] found that assessments of personality traits in adolescence appear to be predictive of CWBs in adulthood still. Over time, this stability of personality traits makes the psychological paradigm a potentially useful tool to help add additional context in understanding potential insider threats and other CWBs.

Psychologists studied the relationship between FFM traits and various forms of CWB. Bolton, Becker, and Barber's study [37] confirmed previous findings with the relationship of Agreeableness with CWB-I and Conscientiousness with CWB-O. However, this same study found that Conscientiousness is only predictive of behaviors of sabotage and theft. They also found that extraversion is predictive of theft, and Openness can help to predict production deviance. Some studies have shown that it is the interaction of traits that allow for more accurate predictions of CWBs [38].

Computer Scientists and Electrical Engineers have only recently begun to explore the predictive ability of personality traits to determine potential insider threats to an organization. However, due to a lack of data, situations are often simulated. A 2019 study from Singapore used games to simulate a world allowing them to put a subject in an environment that allows them to commit insider attacks while collecting personality traits, facial expressions, and linguistic features to identify Intellectual Property theft behaviors [39]. This type of simulation can cause issues when the reward structure does not closely match the real-world; this can cause inconsistencies in identifying potential insider threats to an organization.

Recent studies have attempted to bridge the divide between Computer Scientists and Psychologists. Maasberg et al. [40] looked at correlations of Dark Triad traits and Insider threats. Other studies have looked at Information-Theoretic approaches for measuring the "Trustworthiness" of employees [41] or the creation of Ontological methods for determining insider threats [42]. In 2018, Yang et al. [43] began to explore the use of a more comprehensive system utilizing various psychological models.

All of these psychological and computer science studies into the problem contribute to our understanding of insider threats, but the systems developed are still heavily reliant on linguistic processing. Many of these tools have not considered how large multi-national corporations might attempt to identify insider threats when non-English languages are present.

Terrorism, Radicalization, and Extremism

Terrorism is another enduring threat that faces the national security and safety of many people around the world. The U.S. National Intelligence Council says that terrorism is likely to increase over the next 5 - 20 years due to alienation, ethnic bonds in their networks, the loss of connection with their community of origin among immigrants, and ethnic and religious tensions [44]. Technological advancements allow terrorist groups to mask their identities and activities, recruit new members, finance their operations, and disseminate their messages to other countries [44].

Terrorism is a broad term and consisting of a highly diverse set of actions. Martha Crenshaw [45] describes terrorist acts as ranging from kidnapping of individuals to pressure governments to comply with their political demands indiscriminate bombings. However, the term is often political in nature and subjective, often stretching the definition of the term to new areas (e.g., "cyberterrorism," "ecoterrorism,") to elicit emotions [45].

There are limited studies in computer science literature that seek to understand terrorism, radicalization, and extremism in Computer Science literature; it is often looking for signals of attacks to prevent these attacks, rather than looking at understanding what might cause an individual to radicalize [46, 47, 48]. There have been studies that have used psychological assessments, such as the HCR-20V3 [49], as the basis for building tools to identify the risks of radicalization using social media platforms [50, 51, 52, 53, 54]. Many of these tools focus on identifies risk factors,

such as socioeconomic and demographic conditions of individuals [51], and attempt to build automated tools to identify those risk factors. These authors have also proposed analyzing and studying the relationships between users to measure the risk for radicalization.

Psychology has explored radicalization, terrorism, and extremism beyond just looking for economic and demographic factors. Instead, the field has looked at clinical or personality traits, cognitive factors, and emotions to understand what might cause someone to radicalize. Along with this, there is a rich history of studies that look at violence and the psychological traits. Computer Scientists often overlook these studies, but they can help add additional context for understanding why an individual might radicalize or even determine how de-escalation might occur for the individual.

Milan Obaidi et al. [55] identified that much of the prior research in the area of violent extremism has focused on explanations using clinical dispositions. Some psychologists have indicated a belief that group dynamics, especially in-group dynamics, may play a more significant role in violent tendencies over pathologies, personality, education, income, or any other demographic factor [56, 57]. Others have looked at psycho-pathological qualities, such as suicidal motives or mental health problems [58, 59]. The New York Police Department (NYPD) has utilized its models to help identify radicalization, focusing on "demographic, social, and psychological factors that make the individuals more vulnerable to the radical message." [60]

Sageman [61] suggested that it was personality traits that can predispose individuals towards the path of Jihad. Many early studies on the psychology of terrorism focusing on personality traits portrayed terrorists as having psychological problems [62, 63]. More recent studies have looked at terrorism through the lens of personality traits. Obaidi et al. [55] have shown that non-clinical personality traits, using the HEXACO model, may correlate more highly with tendencies towards violent extremism, mainly appearing in those that are more dogmatic, less empathetic, and less

emotional.

Extremism also includes non-violent extremism, including political extremism and socio-political attitudes. Using the HEXACO model, it [64] shows that Honesty-Humility negatively correlated with Social Dominance Orientation, and Openness to Experiences negatively correlated with Right-Wing Authoritarianism. In both cases, the correlations show that those who tend towards extremist views tend to lack humility, reflecting a hierarchical preference (one person ruling over others) in relations.

A separate study [65] using the Linguistic Inquiry and Word Count package and IBM Watson's Personality Insights found that both Left-Wing Extremists and Right-Wing Extremists tend to be less agreeable, less neurotic, and more open than non-extremists. Despite these correlations, Bell et al. [66] have shown that personality tests alone cannot distinguish those who would commit militant terrorist acts from non-terrorists. Another study looking at adolescents and extremism found that personality factors such as low intellect/Imagination, low Extraversion, and high Agreeableness indicate potential vulnerabilities to extremist ideologies. [67]

[65] also looked at emotions and whether they could be used to determine political extremism. It found that Left-Wing Extremists tend towards negative emotion words, while Right-Wing Extremists tend toward positive emotion words. However, the authors did note that these affective differences reflect the ideological direction of political extremists, but not their militancy.

While personality may help to add context, Other psychological models have also been applied, such as the Dark Triad / Dark Tetrad, looking at how they might identify the radicalization of extremists. The Dark Triad consists of Psychopathic, Narcissistic, and Machiavellian traits; the Dark Tetrad adds a fourth category for Sadistic traits. One study showed that this Dark Tetrad of traits could indicate radicalized behaviors in women, showing that a significant portion of non-clinical

french college women are at risk of religious radicalization [68].

We see that while Computer Scientists have focused on identifying existing radicals and extremists, psychology has worked to develop new models in understanding the process of radicalization. Martha Crenshaw [45] highlights that research into terrorism needs to go beyond the focus on current events or speculations about the future developments of the phenomena over time. Researchers can and should continue to focus on understanding the basis for rationalizing terrorism, but psychological research can also help in understanding what causes groups to end violence [45]. While the tasks that computer scientists have focused on are no less important tasks, using Machine Learning techniques to measure various psychological traits could help in more substantial studies to understand those people that might become militant extremists, perhaps helping to prevent radicalization or deescalate extremists in the future.

Deanonymization of Threats

Methods for computer-mediated communications have increased in popularity since the 1990s with the invention of the World-Wide Web. This increase in the use of computer-mediated communications has given rise to new Social Media platforms that play a large role in economies around the world. In 2010, the Canadian Library of Parliament claimed that Social media had changed the way content is both created and consumed, and because of that, the platforms have changed the information and communications technologies sectors [69]. Between 2002 and 2011, the economic growth in Canada related to Social Media platforms grew at twice the rate of Canada's overall economy [69]. More recently, the use of social media has begun impacting presidential elections, allowing the spread of information more broadly, and creating new concerns around the dissemination of misinformation and "fake news." [70]

Anonymity on the internet allows for the creation, updating, and distribution of

content to people around the world. This anonymity is essential in democratizing information by encouraging active engagement, particularly from members who may feel threatened by sharing information [71]. While this anonymity is vital in ensuring the privacy of those sharing information, anonymity can also create an environment that allows for much of the extant negative behaviors online [72]. These negative behaviors have real-world impacts, often affecting the lives of victims of attacks. The concept of "doxing" - the internet practice of researching and publicly broadcasting personal information [73] - and "swatting" - deceiving emergency services into sending police response teams to an individual's address [74] - have threatened the lives of people and on occasion caused unnecessary deaths. Other forms of cyberbullying have a significant psychological and social impact on both the victims and the bullies, and is a contributing factor for depression and suicide of victims [75]. More extreme forms of protecting identities online, in what is known as the "Dark Web," has allowed for the spread of hate speech, terrorism planning, and the exploitation of children [76].

Much of these harmful acts attribute to the ease with which individuals can commit acts of online deception, often aided by limited strategies to verify information conveyed to people [15]. In cases of anonymity, this seems to fit under the concept of identity deception - the willful intent to provide false information due to a lack of verification methods [77]. This form of deception can range from seemingly innocuous misrepresentations of gender, age, ethnicity, or physical appearance for online dating profiles [16] to more extreme cases, such as child pornography, sexual predators, and pedophiles [78].

While personality itself is not a good indicator of an individual's identity, it can help in understanding various potential aspects of one's identity. Things such as an individual's political leanings have correlated with personality - conservatives tend to be higher in Conscientiousness, but lower in Openness to Experiences than liberal counterparts [79, 80]. Another benefit is the use of the constancy of personality traits

from adolescence to adulthood [20]. This constancy can help validate the person one is talking to has remained the same throughout the communication.

Stylometry is a technique of analysis of documents used to determine the authorship of given documents. While this can be traced back to some of the early attempts to identify the authorship of Shakespearean plays, it was the work of Frederick Mosteller and David Wallace in the 1960s that laid the foundation for computer drive stylometry [81, 82, 83]. Enhancements in communications devices, including the internet, has created new challenges in the field allowing the techniques to assist in countering identity deception problems, such as email misuse [84], continuous, active authentication [85], presidential speechwriters [86], deceptive social media postings [87], and text messages from the short messaging service (SMS) [88]. Researchers have defined author stylometry as a behavioral biometric since authors can change their styles according to their desires, the genre, the topic, the author's emotional state, and a variety of other factors [89]. For this reason, various psychological models might also help identify the authorship of a document, post, or SMS.

Because of the impact of psychology on these various economic and social areas of interest, this dissertation explores a handful of behavioral models, including Emotional Intensity Analysis, Estimation of Personality traits, and Authorship stylometry. Many of these threats are not inherently in English; therefore, the impact of translation on psychological estimation systems needs to be analyzed. Chapter III highlights the models that were developed and conduct a literature review that looks at similar work in the three domains. Chapter II and IV explores the datasets used, the translation engine that chosen for experimentation, and conducts a statistical analysis to determine the impacts of translation on the systems, and potential methods for increasing the metrics of the systems while working in various languages.

3 Privacy and Legal Issues

While tools developed for identification of potential radicalization, insider threats, or the deanonymization of dangerous posting on the dark web or social media may be well-intentioned, there are potential privacy and legal concerns that might prevent the broad application of psychological tools. While companies may desire to log, monitor, and audit employee interactions on company devices, Carly Huth [90] highlights the challenges for companies establishing these policies on work devices, especially when companies may have vague personal-use policies. In these cases, these monitoring tools may lead to legal issues when used for employee termination. Carly Huth [90] gives an excellent survey of the various laws and regulations that are in place to maintain the privacy of employees and the further demands of government employers.

U.S. law enforcement and U.S. federal government agencies also have to abide by the constitutional and legal rights afforded to citizens and legal permanent residents. Many of the tools discussed may be most beneficial when used very broadly, but the Fourth Amendment of the U.S. Constitution's Bill of Rights protects people from unreasonable searches and seizures; this amendment should prevent tools such as these from being applied very broadly. While this protects U.S. citizens from many potential abuses, the "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001" (USA PATRIOT Act) modified various laws allowing federal agencies access to various tools to combat terrorism through the use of electronic surveillance [91]. Even under this congressional act, Presidential Executive Order 12333 provides further limitations on various federal agencies that might attempt to use such tools broadly [92].

Other countries may have more relaxed legal environments allowing them to use such tools more broadly. While these types of tools may be beneficial applied broadly, companies and governments ought to approach the use of such tools sparingly to avoid potential abuses. Understanding someone's personality alone would not make them

a potential terrorist. Understanding that an individual is emotionally angry does not make them a potential insider threat. These tools help provide additional context about a person and help inform analysts and law enforcement officers of the way an individual might react. The tools can help to understand what might have motivated the individual to commit an act. However, users of these tools should never make decisions about a specific individual without the additional context needed to make a final decision.

CHAPTER II

DATASETS

One of the limiting factors in the exploration of this field of study is in the creation of properly ground-truthed datasets. Computational Behavioral Analytics datasets can be created through a variety of different means depending on the task. In some cases, the exploration of these types of datasets was cost-prohibitive in this dissertation. Another complicating factor in this dissertation was in finding datasets in various foreign languages. Here we explore methods used in the creation of datasets for Affect (or Emotion) analysis; a discussion of the dataset used for personality trait estimation, along with the limitations in this dissertation that prevented the creation of a foreign language dataset with the potential methodology for other scientists to conduct future studies in foreign languages; and the datasets used in Authorship Attribution.

While there are several different possible tasks under Computational Behavioral Analytics, we focus on three specific tasks in this dissertation in order to show methods for conducting these psycho-linguistic assessments that makeup Computational Behavioral Analytics and could have some potential benefit in various areas of understanding the person behind the text.

The first task is affect (or Emotion) analysis. Unlike Sentiment Analysis, affect analysis attempts to understand the intensities of the primary six emotions being expressed in the writing sample, regardless of the topic discussed in the text. When used

in combination with topic analysis and sentiment analysis, this tool can potentially identify threats before a malicious actor can take escalating action.

The second task is personality trait estimation, or rather the exploration of estimating the traits that make up the Five-Factor Model - Openness to new ideas, Conscientiousness, Agreeableness, Extroversion, and Neuroticism (or Emotional Stability). While not necessarily indicative of whether someone will commit a malicious act (intentionally or unintentionally), it can give a better understanding of what might motivate a person and how this person might react in various situations. Other traits such as the Dark Triad or Locus of Control would be better indicators of whether someone might act maliciously.

The third task explored is Authorship Attribution. This task is a behavioral biometric, and while it is often not explored in psychology, it does rely on the author's emotional states, the author's personality, and other traits that can influence the author's word choice. This area of study has more impact in the identity sciences realm, and can be useful in identifying potential cyber-personae.

1 Affect Analysis

The primary purpose of Affect Analysis is to estimate the emotional intensities expressed in short textual messages, including SMS messaging or Tweets. This type of analysis is useful in estimating the emotional state of individuals who might commit malicious acts. The same analysis, when scaled up, can measure the general mood of a region as a possible indicator for social unrest within regions or groups of people, though further experimentation is needed to ensure that this marker is, in fact, useful.

Sentiment Analysis is a tangential task that is often better studied in literature than affect analysis. Sentiment Analysis generally focuses on the estimation of positive or negative valence toward specific topics, whereas affect analysis generally focuses on the emotional intensities expressed in the text (i.e., how angry is this per-

son overall) [93]. Many studies have used lexicons, whether manually generated or generated with the help of machine learning, for the development of their systems [94, 95, 96, 97, 98]. Other studies have also looked into lower-level semantic and syntactic structures, utilizing word n-grams [99] or parts-of-speech n-grams [100]. The use of lexicons remains one of the most common techniques for sentiment and affect analysis, even among recent studies looking at foreign languages [101]. Abbasi et al. [102] explored the use of many features within the linguistic space to estimate emotional intensity within short text samples.

Scientists have studied emotions for nearly a century in literature. Early studies of emotions looked at the physiological reaction of emotions [103]. This theory was later expanded on by Cannon and Bard, attempting to explain the physiological reactions as a concurrent signal being sent to the brain to evoke both an emotional and physiological reaction [104]. Further cognitive studies took on the Two-Factor Theory of Emotions proposed by Schachter and Singer [105]. This theory proposes a physiological arousal occurs first, followed by the individual reasoning about the cause of the arousal to experience an emotion.

While these theories attempt to explain emotions as an evolutionary process, various appraisal theories have also attempted to explain emotions. Appraisal theories explain the process of emotions as linked to the immediate evaluation of the individual's circumstances [106, 107, 108, 109]. This model explains the variations of emotions that an individual experiences under similar circumstances. While an individual may become saddened by the failure of a school assignment, another individual might experience anger or fear in a similar circumstance.

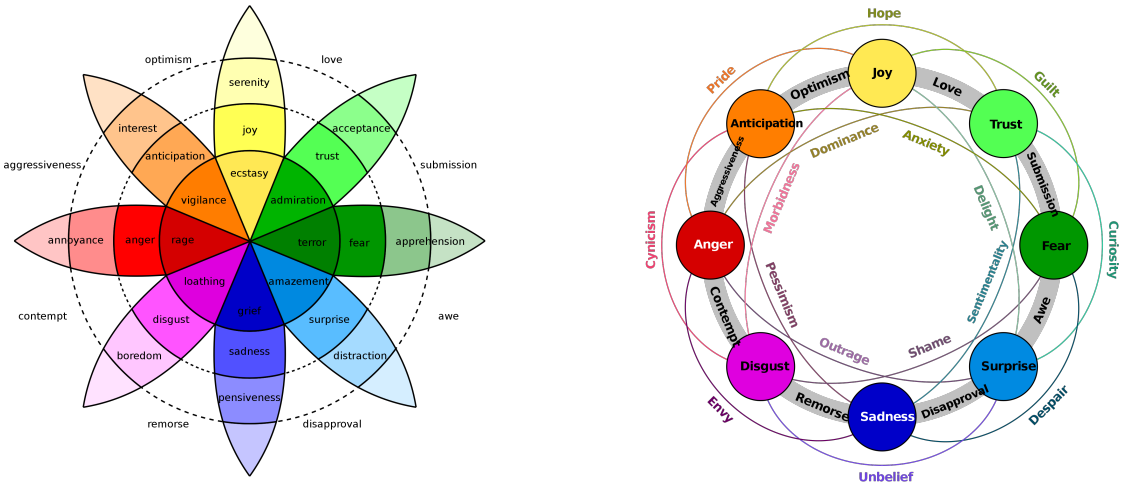
Psychological literature often cites two prominent theories of emotions: the emotional aspects of Plutchik's theory of actions as an adaptive problem [110, 111], and the basic emotions described by Paul Ekman [112, 113]. We utilize aspects of both of these theories for the development of the algorithm used in experimentation. Plutchik

identifies eight basic emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. Paul Ekman initially identifies six basic emotions: anger, disgust, fear, sadness, happiness, and surprise.

Paul Ekman's basic emotions were determined by looking at facial expressions to identify potential deceit from individuals. The basic emotions are anger, happiness, surprise, disgust, sadness, and fear. Expression of these emotions occur through similar facial muscular movements across both Eastern and Western cultures [114]. Later studies by Paul Ekman determined that other universal emotions exist cross-culturally, though they are not always identifiable through facial expressions. These emotions include: Amusement, Contempt, Contentment, Guilt, Relief, Shame, and others [112, 115].

Plutchik's wheel of emotions, shown in Figure 1a, shows each of the emotions that are identified by Plutchik. The emotions without colors shown between the colored, basic emotions are combinations of emotions. Plutchik theorizes that remorse is a combination of both disgust and sadness; contempt is a combination of anger and disgust; and so forth. Plutchik further expands this concept by looking at dyads of emotions, an image shown in Figure 1b. Plutchik defines some emotions as being opposites of one another - fear is the opposite of anger; sadness is the opposite of joy; and so forth. Each emotion that is shown 180 degrees around the wheel shows each of the opposite emotions. In Plutchik's Dyads, these opposites never occur in combination. Every other emotion is connected to highlight more complex emotions: Outrage is a combination of Anger and Surprise; Guilt is a combination of Joy and Fear; and so forth. Since emotions are often complicated, and it is often difficult to determine when a feeling is a combination of two or more feelings.

Plutchik also introduces the concept of emotional intensity. As each of the basic emotions becomes increasingly closer to the center of the wheel, the emotions become more intense: anger becomes the emotion of rage, fear becomes terror, joy



(a) Plutchik's Emotions **(b)** Plutchik's Dyads

Figure 1. Diagrams showing Plutchik's Wheel of Emotions (a) and emotional dyads (b) that were proposed by Plutchik [110, 111].

Image (a) is in the public domain [116] and Image (b) was used under Creative Commons Attribution-Share Alike 4.0 International License [117]; No modifications to the images were made. Both images were downloaded from Wikimedia Commons.

becomes ecstasy. As the emotions move away from the center, the emotions become less intense: less intense anger is annoyance, trust becomes acceptance, anticipation becomes interest. The algorithm discussed utilizes some aspects of Plutchik's theory. While the algorithm is not explicitly trying to identify specific combinations of emotions, the algorithm does allow for the possibility of multiple emotions conveyed in text. The algorithm also incorporates the concept of emotional intensities, allowing the algorithm to determine the intensity of each emotion's expression. For simplicity in creating an Arabic dataset, we utilize the emotions described by Ekman and incorporate the concept of intensities of emotions on a normalized scale.

English Language Datasets

The Support Vector Regression Correlation Ensemble methodology explored by Abbasi et al. [102, 46] appears to be one of the most flexible algorithms for understanding the intensities of emotions present in a text. For that reason, this dissertation repli-

cates portions of their work and expanded on to determine the viability of use in foreign languages.

For the research from Abbasi and Chen [46, 102, 93], two dark web forums, Al-Firdaws and Montada, were selected for exploration. For their studies, a subset of the collected forum posts were examined and coded by a domain expert. Five hundred sentences from each forum were selected and scored on a continuous, normalized scale by experts for the intensities of sentiments and affects expressed. The affects annotated in the dataset included: violence, anger, hate, and racism [93]. This dataset results in a relatively small dataset for each potential affect, and the selection by an expert could potentially bias a dataset. However, the methodology of their algorithm is potentially interesting. This dataset did not appear to be available to other researchers, so other datasets were obtained to test comparable systems.

To build the algorithm, we identified two English language datasets for testing the emotion analysis system. SemEval (Semantic Evaluation) is a competition to evaluate computational semantic analysis tools and algorithms hosted by the Associate for Computational Linguistics. In both 2007 and 2018, SemEval explored the evaluation of emotional intensities in both headlines from news sources (2007 Task 14) and Twitter posts (2018 Task 1). Both of these datasets have limitations that are not ideal in developing the final product, but they are sufficient for the needs of building and testing the algorithms in a language that could be understood by the author of this dissertation.

SemEval 2007 Task 14 is composed of news headlines from the New York Times, CNN, BCC News, and other news sources from the Google News search engine. The dataset was annotated from an online tool that displayed single headlines with six slides for each of the emotions, along with a seventh slide bar for valence. This system allowed annotators to score these headlines between [0, 100], where 0 is the emotion not being present in the headline, and 100 being the maximum intensity

of the emotion. The valence score was annotated on a [-100, 100] scale, where -100 is a highly negative news headline and 100 was a highly positive headline [118]. This study was one of the first datasets found where the authors explore fine-grained details on emotional intensities. Before this, many studies simply relied on binary decisions of the emotion being present or not being present. Fine-grained details allow the authors to capture joint emotions that might be present in the headlines. Further analysis of their dataset is found in [118]. This study ignored the valence scores and reformatted the data into a Comma Separated Value format, containing the news headline and the annotated emotional intensities from the SemEval dataset. The emotions captured in this dataset follow the Ekman six emotions, and since all emotions align with the headline, the correlation ensemble outlined by Abbasi et al. [102, 46] can be explored with this dataset.

The creation of the SemEval 2018 Task 1 dataset is a bit more complicated. In this, the authors created a lexicon of words for each of the four emotional categories explored (anger, fear, joy, sadness). Using this lexicon, the authors polled Twitter's feed for two months to identify tweets containing words in the lexicon. From this, 1,400 tweets were randomly chosen from the joy category and annotated through a crowd-sourcing platform. For the remaining three categories that the authors call "negative emotions", 600 tweets (200 from each of the emotional categories) were randomly chosen to be annotated for all three of the emotional categories. Another 800 tweets were selected for additional annotation for each of the emotions. Because of this, each emotional category contains 1,400 tweets. Except for the 600 "negative emotion" tweets, each tweet was annotated for a single emotion [119]. The capturing of only single emotions limits the ability to use this algorithm to create correlation ensembles. Each tweet is simply binned into their emotional category with an intensity score, allowing for a regression analysis of the text to estimate the intensity of that emotional category.

There are limitations that system architects need to be aware of when using these datasets for any specific purpose. While the SemEval 2007 Task 14 dataset is perhaps one of the more robust in capturing all six of the emotions for each headline, the use of news headlines limits the ability to use this outside of this category of text. Experimenting with angry tweets discussing topics such as "death" and "killing," it was found that news headlines rarely evoked anger. Instead, these words often were intended to evoke the emotion of sadness.

The lack of annotation of tweets for all emotions in the SemEval 2018 Task 1 prevents the use of correlations between emotions to improve the system's accuracy across these hidden correlations. This dataset makes many assumptions in its creation, and the use of a lexicon limits the potential to capture hidden features that might indicate a specific emotion that the author may not have taken into consideration.

Ideally, a future English language dataset would be collected to fix the limitations of these current datasets, but this is outside of this dissertation's scope. These limitations do not prevent the development of the algorithms for testing in foreign languages (which is the primary focus of this dissertation.) Instead, the SemEval 2018 Task 1 dataset allows for a Support Vector Machine for Regression (SVR) for each of the emotional categories explored, while the SemEval 2007 Task 14 dataset assists in building both the SVRs and in testing the correlation ensemble.

Foreign Language Dataset

A foreign language dataset is needed to test the efficacy of translations on the algorithms. Unfortunately, many of the currently existing datasets only focus on English. A dataset was created in order to test the efficacy of translations in emotion analysis. For our experimentation, we follow a similar data structure as the SemEval 2007 challenge dataset, choosing to use the Ekman basic emotions (anger, disgust, sadness,

fear, happiness, and surprise) [118]. These emotions provided a good baseline in experimentation and seemed to provide the most culturally agnostic set of emotions for building such a dataset.

The first task in building the foreign language dataset was the collection of, in this case, Arabic language tweets. While Modern Standard Arabic (MSA) is a standardized literary Arabic used in formal writings, the use of twitter often entails more colloquial terminology, though the Arabic script remains the same. The Egyptian Arabic dialect was chosen as a focus to account for potential colloquialisms that might occur in social media postings. A dataset was created using the 1% feed using Twitter's API and further filtered to account for the target language using Twitter's language tagging algorithm. We then looked at the geotag for tweets around a target area (in this case, geotags from Egypt) to maximize the collection in the targeted Egyptian Arabic dialect. All tweets were post-processed for anonymization, removing proper names and identifying information using Python's NLTK libraries.

Once the approximate threshold of 75,000 tweets was achieved, a series of three cascading tasks were given to annotators on the Amazon Mechanical Turk (AMT) platform. An initial test was given to potential annotators to ensure that the annotators spoke the Egyptian dialect of Arabic well enough to understand the presented tweets. Even using this method of identifying Egyptian dialect speakers, the method is far from perfect. There are significant overlaps between different varieties of Arabic, and many (though not all) of the dialects are mutually intelligible. Further complications arise because many Arabic speakers also have at least a passing familiarity with the Egyptian dialect, since Egypt can arguably be called the entertainment capital of the Arabic-speaking world.

Despite the challenges, the creation of this dataset used a test created by linguists and outlined in a technical report [120]. Annotators were given a list of Arabic sentences containing colloquial terms and phrases that were not highly intelligible in

TASK #	Median Agreement
Task 1: Is there emotion?	91%
Task 2: Is there (specific emotion?	94%

Table 1. Median Agreement between Annotators during tasks 1 and 2.

other Arabic dialects. The initial test asked the annotators to translate the Arabic sentence to mimic the task closely. While the task is slightly more challenging than a reading comprehension test, it is closer to the intended Human Intelligence Task (HIT) on AMT. An example test question is in Figure 2.

Once an annotator was approved for conducting the tests, a series of cascading tasks would occur. During the first two tasks, a series of control questions measured the agreement between annotators. Table 1 lists the median agreement between annotators on all control items in tasks 1 and 2.

The first task verified that the tweet shown was, in fact, the correct dialect, and whether the tweet expressed any emotions. Annotators read the tweet and indicated one of the following: Tweet expresses an emotion; Tweet contains no emotion; Cannot tell if the tweet contains an emotion; or that the tweet contains some language or dialect other than Egyptian Arabic.

In the instructions to the annotators, annotators were told to use the "wrong dialect" liberally. Annotators marked the tweet as the wrong dialect even if an emotion was present; if there was a mix of languages in the tweet (English and

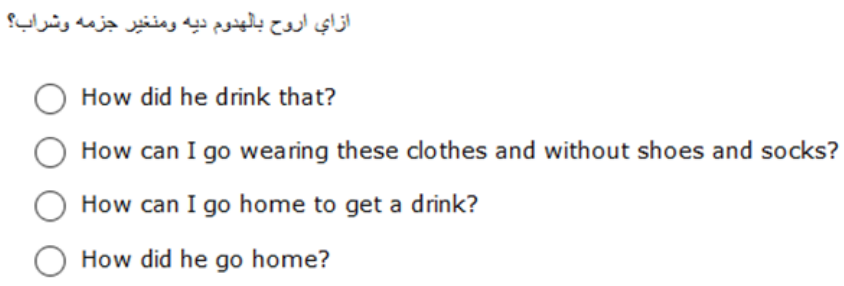


Figure 2. Example test question used for qualification to participate in annotation of the Egyptian Arabic dataset.

Arabic); or anything written in MSA or formal literary Arabic, such as passages from the Qur'an or Bible, formal invocations, prayers, greetings, or rituals. When a tweet was the "wrong dialect", the tweets were not evaluated further, and the tweets were excluded in the dataset. The annotators had to decide whether the tweet contained an emotion even if the tweets were too short to determine the dialect. Any loan words or Latin-scripted twitter handles or hashtags were not marked as wrong dialects either.

Once these tasks were completed, any tweets of the wrong dialect or expressed no emotions based on annotators consensus, were discarded from the dataset. All of the remaining tweets that contained some emotion were passed on to the second task. In this second task, tweets were bucketed based on the emotions that annotators believed that tweets contain. The annotators were shown a tweet and asked whether the tweet contained a specific emotion (e.g., anger) by selecting one of the following options: Has anger, No anger, Cannot tell, and not Egyptian Arabic. The option "Not Egyptian Arabic" remains in the task if any tweets managed to slip through the previous filters, though this option was infrequently selected after task 1.

Each tweet was shown for the six emotions, and to at least five annotators for consensus on each emotion. Therefore, each tweet was shown at least thirty times – five times for anger, five for happiness, continuing for each emotion. Since each tweet was tested for each emotion, a tweet was allowed to be binned into multiple emotional categories. These multiple categories of emotions allow the algorithms to identify correlations between emotions that might not otherwise be captured, further improving results and allowing experimentation against the algorithm(s) described in Abbasi et al. [102, 46].

The final task was the task intended to estimate relative intensities of emotions. For this task, minimization of cognitively demanding tasks was desired, such as rating how intense an emotion was using Likert scales. Using cognitively challenging tasks can result in more significant disagreements between annotators since each would cog-

nitively anchor their responses based on individual experiences. These disagreements can become problematic when training an algorithm, especially on subjective targets such as the emotional intensities explored in the study.

Instead, a simple pairwise judgment task was given to annotators for measuring the intensities. For each of the emotional buckets, two tweets were selected; the annotators were asked for comparative judgments on which of the tweets contained the most intense emotion. While, ideally, every tweet would be compared against one another, this would increase the costs and require significant amounts of time for annotators to complete. Instead, each tweet was involved in at least ten comparisons using random selections. Using these comparisons, we were able to induce the emotional intensities from the pairwise judgments using linear programming.

For this process, we define a set of *tweets* as the tweets from a specific emotion (A):

$$tweets = \{t_1, t_2, ..., t_n\} \tag{1}$$

$f_a(t_i)$ is defined as the function that determines the intensity of the emotion (A) from tweet (t_i). In order to create a linear programming optimization problem, we assume that we would like to computer $f(t)$ such that we respect as many of the annotator's judgements involved each tweet as possible, and where we assume "t_1 is angrier than t_2" implies $f_{\text{anger}}(t_1) > f_{\text{anger}}(t_2)$. Using these assumptions, we can encode the judgements as an inequality constraint:

$$\begin{aligned} f_A(t_1) &> f_A(t_2) + 1 - \lambda_1 \\ f_A(t_3) &> f_A(t_2) + 1 - \lambda_2 \\ &\text{etc.} \end{aligned} \tag{2}$$

Where λs are slack variables that can be increased or decreased to account for inconsistent judgments, the linear programming task becomes a minimization of the

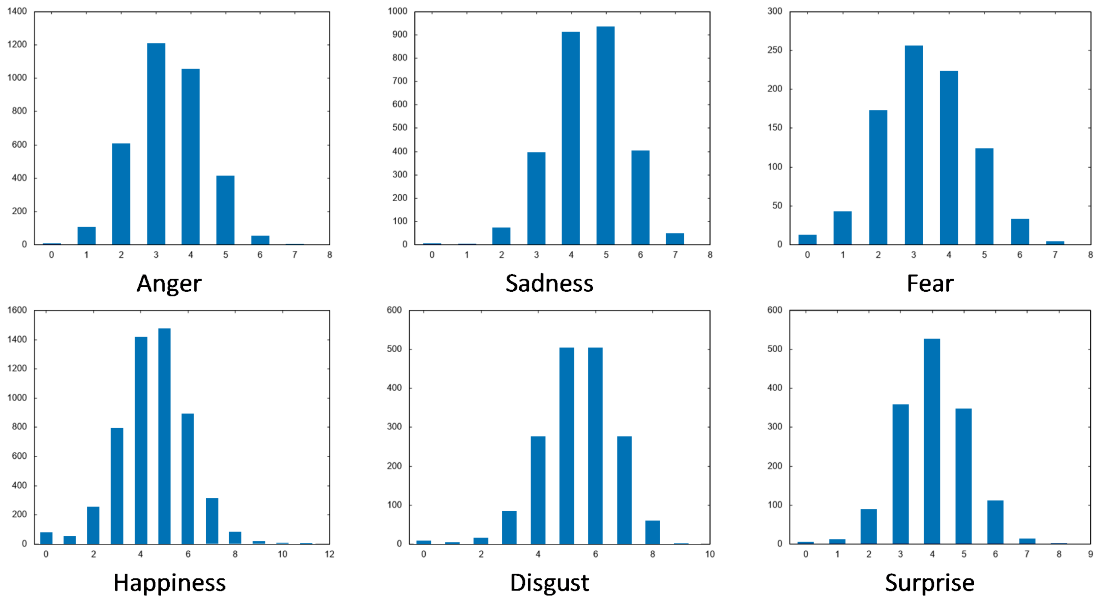


Figure 3. Equivalency classes for intensity scores for each of the six emotions.

Emotions	Tweets	Yield (%)
All tweets	75,019	-
Any emotions	28,165	38%
Anger	3,440	12%
Sadness	2,766	10%
Fear	870	3%
Happiness	5,388	19%
Disgust	1,727	6%
Surprise	1,461	5%

Table 2. All rows show the total number of Egyptian Arabic tweets collected for that class; "Any emotions" shows the filtering results in the first Task, the "Any emotions" yield shows the percentage based on "All tweets"; Each of the emotions list the total number of tweets in that emotional class, the yield percentage is based on the "Any emotions" tweets.

sum of lambdas ($\min \sum \lambda$), or minimization of the total effect of the inconsistencies. Equivalency classes for intensity scores were induced for each of the six emotions using this method. A histogram showing the number of tweets is in figure 3.

The final Arabic corpus collected contained 3,440 anger tweets; 2,766 sadness tweets; 870 fear tweets; 5,388 happiness tweets; 1,277 disgust tweets; and 1,461 surprise tweets. Of the initial 75k Egyptian Arabic tweet dataset, only about 38% of the tweets contained any emotions. Within those, only a small percentage yielded tweets with the targeted emotions. The yield percentages and fine yield totals are summed up in table 2.

Only the tweet text was kept in the collection, and no other demographic information was kept during the collection. This process was followed for the anonymity of the users. Since this was a random sampling of twitter users, we expect that the demographics should be similar to those of the usual twitter users.

2 Personality Trait Estimation

Datasets that are relevant to any psychological assessment of users are harder to acquire, whether through text, audio, or video. These datasets need to collect the media and also administer relevant psychological assessments, such as the International Personality Item Pool (IPIP), Revised NEO Personality Inventory (NEO-PI-R), short dark triad (SD3), or another psychometric tool commonly employed by psychologists. While the psychological instruments and text from the users are easy to acquire individually, it is rare to find datasets that acquire both social media and psychometric instruments.

In Personality Trait Estimation, algorithms attempt to estimate each of the tested traits against a psychological personality instrument. Various instruments could help build algorithms, computational scientists and psychologists have explored the use of some instruments for various Human Resources related tasks. Some of the most

prominent instruments include the DiSC profile assessment, Big Five Inventory, HEXACO, the Minnesota Multiphasic Personality Inventory, among others. Other psychological instruments also exist that may allow for correlations to written text or other forms of multimedia. Instruments such as the Dark Triad or Dark Tetrad might be relevant and exciting to explore, but data limitations prevented the exploration of these instruments. Many people are also familiar with tests such as the Myers-Briggs Type Indicator assessment. While this test has become a standard in some workplaces, its categories are not very reliable, not scientifically valid, not independent, and not as comprehensive as other scientifically valid experimentation [121]. Instead, this dissertation focuses on scientifically valid instruments.

Industrial psychologist Walter Clarke created the DiSC assessment, publishing his first version in 1956 [122]. Walter Clarke based his model on William Marston's earlier works that theorized that the expression of emotions categorizes into four primary types: Dominance (D), Inducement (I), Submission (S), and Compliance (C) [123]. Wiley, the test publisher, has continued developing the DiSC profiles attempting to make the tests more accessible to users of the test. Human Resources related companies have explored the use of the DiSC profile as a potential for identification of team fit, dynamics, and for hiring purposes for companies [3].

The Minnesota Multiphasic Personality Inventory (MMPI) is a more clinical test intended for testing personality and psychopathology in adults [124]. The test was created by Schielem Baker and Hathaway in 1943 and published by the University of Minnesota Press [125]. Since that time, the test has taken on several revisions, allowing the test to be given to adolescents (MMPI-A) [126], and adopting newer theoretical approaches to personality test development (MMPI-2-RF) [127]. It has been used by Psychologists and other Mental Health professionals to assist in the differential diagnosis and the development of treatment plans. While it has its basis in clinical domains, it is also often used in forensic psychology and candidate screening

[128].

The DiSC and MMPI are instruments for consideration in further studies for correlations to multimedia, but Psychological literature commonly focuses on the use of the Five-Factor Model or the HEXACO models for personality and behavioral studies. The models came about as a result of factor analyses of both survey data and lexical features from several independent researchers. Early lexical studies in the 1940s by Cattell [129] found 16 personality traits. Tupes and Christal [130] were the first to propose just five factors in 1961, but this analysis did not catch on until independent researchers conducted similar studies through the 1980s and 1990s, including the work from Goldberg [131], and Costa and McCrae [132].

The Five-Factor Model (FFM) is a taxonomy that identifies five underlying dimensions used to describe personality:

- Openness to experience
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism (or, when inverted, Emotional Stability)

Openness to Experience consists of two major sub-components: intellectual dispositions and the other related to aesthetic appreciation and sensory experiences [133]. McCrae and Costa [134] highlight that while Openness may be related to Intellect, it is an indicator of personality and not intellectual ability. They continue by saying that people who score high in Openness do not necessarily have a corresponding high Intelligence Quotient (IQ). Those who score highly in Openness have a higher intellectual capacity; often enjoy art, music, or other aesthetic impressions; or have

a wide variety of interests. Those with a low openness score are often described as more conservative and often repress anxieties. [134]

Conscientiousness determines how someone takes obligations and tasks. Highly Conscientious people tend to show self-discipline, and often plan behaviors rather than act spontaneously. They are often neater and more systematic than their counterparts. Hogan [135] also views Conscientiousness as inhibiting impulsive behaviors. People who score lower on the Conscientiousness scale tend to be more laid back, less goal-oriented, and maybe more likely to engage in antisocial and criminal behaviors [136].

Extraversion tends to be representative of how outgoing or reserved a person is in situations. Watson and Clark [137] identify seven components that make up Extraversion: venturesomeness, affiliation, positive affectivity, energy, ascendance, and ambition; McCrae and Costa view of Extraversion assigns ambition to Conscientiousness [134] and breaks affiliation into two categories, warmth and gregariousness. Those who score low on Extraversion tend to be more quiet, reserved, silent, or withdrawn.

Agreeableness reflects a general concern for social harmony and tend to value getting along with others. Those who are highly agreeable tend to be more considerate, kind, generous, trustworthy, and willing to compromise their interest with other individuals [138]. Those who score low on agreeableness tend to have a less optimistic view of human nature and tend to place self-interest above other individual's interests. They are often competitive, argumentative, or seen as less trustworthy [139].

Neuroticism represents the tendency to express emotional distress. Highly Neurotic scorers tend to experience negative emotions and are more prone to develop a variety of psychiatric conditions [140]. These distressful experiences are often associated with irrational thinking, low self-esteem, and ineffective coping [132]. While low Neurotic scorers are not necessarily mentally healthy, it does mean that they are

usually even-tempered, calmer, and more relaxed than their counterparts.

The HEXACO model used the same factor analysis methods that helped in the discovery of the FFM. The HEXACO model adds a single personality trait, Honesty-Humility, while renaming Neuroticism as Emotionality. The discovery of the Honesty-Humility trait occurred while conducting comparable analyses in foreign languages. Analysis in foreign languages showed that a sixth trait existed. It was not until modern computing power could further analyze the English language that the sixth trait was found in English language studies. [141] The Honesty-Humility trait is associated with sincerity, fairness, pretentiousness, and greed [142].

Personality Dataset

For experimentation, and due to data limitations, studies in this dissertation focus on the FFM traits. Between 2007 and 2012, the MyPersonality dataset* was created at the University of Cambridge, collecting both social media posts and various psychological instruments [143]. This dataset was collected with the consent of its users through a Facebook application. They targeted subjects through a variety of means having each user provide permission to download demographic and status updates from the Facebook platform. Subjects provided a self-report using an implementation of both the 100- and 300-questionnaire of the IPIP proxy for NEO-PI-R, receiving a reliability reported in their respective manuals or standardization samples.

The dataset had numerous tests available for subjects to take, and not all subjects gave consent to collect their Facebook status updates. This mismatch of subjects leads to some inconsistencies in the number of viable subjects for training and testing any machine learning algorithms. When looking only at those subjects who took the

*This dataset is no longer available and was obtained by the University of Louisville before the dataset's closure for research purposes. The dataset was removed by the University of Cambridge to cooperate with changes in Facebook's Terms of Use, but the dataset remains scientifically valid. Future research into personality estimation, by this author or others, will require building comparable datasets for scientific purposes.

Statistic	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Mean	3.79	3.49	3.58	3.55	2.77
Median	3.80	3.50	3.67	3.58	2.75
Std. Dev.	0.68	0.73	0.81	0.70	0.80
Minimum	1.00	1.00	1.00	1.00	1.00
Maximum	5.00	5.00	5.00	5.00	5.00

Table 3. Basic Statistics for the FFM traits found within the Dataset for all subjects. Numbers are normalized between [1,5]

NEO-PI-R for collection, we have a total of 3,029,503 total users.

For each of the FFM traits, we expect to see a normal distribution across the normalized scale from [1,5]. Histograms showing the number of subjects which fall within the scale is in Figure 5. Each of these traits is skewed slightly to the right (with Neuroticism being inverted, and skewed slightly to the left. The mean of each of these generally fall around 3.5 - the exceptions being Openness, which is at 3.8, and Neuroticism, at 2.7. Each of the standard deviations is between 0.68 and 0.8 on the scale. Complete statistics for the data within each trait is in Table 3.

While the NEO-PI-R data contains over 3 million unique users, there are far fewer users who consented to having their Facebook status posts collected. In this case, there were only 153,727 unique users in the Status Updates data. This subset of users consists of over 22 Million unique status updates that were collected. By limiting the FFM Trait data to only this subset of users, we can see that the statistics do not change significantly, as shown in Table 4. This subset also does not significantly alter the distribution of the user scores, as shown in Figure 6.

Statistic	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Mean	3.85	3.47	3.52	3.57	2.73
Median	3.95	3.50	3.5	3.65	2.75
Std. Dev.	0.67	0.73	0.81	0.70	0.80
Minimum	1.00	1.00	1.00	1.00	1.00
Maximum	5.00	5.00	5.00	5.00	5.00

Table 4. Basic Statistics for the FFM traits found within the Dataset for only those users who provided status updates. Numbers are normalized between [1,5]

Further exploration in the data reveals that the number of posts for each user follows an exponential distribution, as shown in Figure 4. This data shows that the median number of posts provided by users was 93 status updates per user, the maximum for a user was 2,441 status updates. Users within this subset should provide enough data for training and testing any algorithms.

Further demographic information helped to determine if this data would be useful in identifying whether any foreign language data exists in the dataset. Much of this data was not useful as many people appeared to obfuscate information about their country. We found a total of 77 different locales that were provided by user demographic information. While en_US and en_GB were the two largest populations, the other locales had far fewer and include countries with which the data did not correspond to any foreign language data found. This mismatch of language and locale indicates that the locale is not a useful tool for identifying foreign language data. It also appears that the country that the authors had listed would not be useful in identifying potential foreign language data. The data had 218 countries represented for users in the subset, including 20 users in "Antarctica."

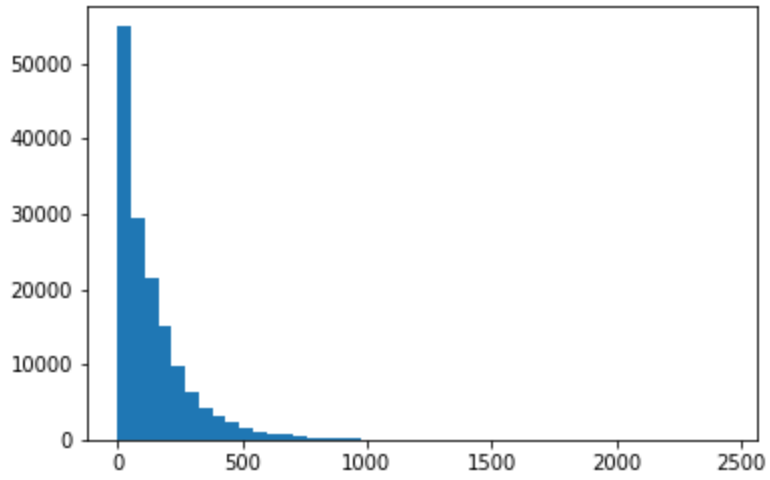
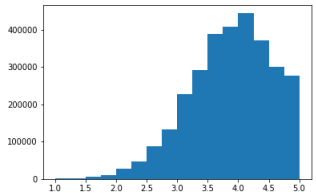
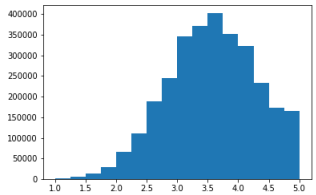


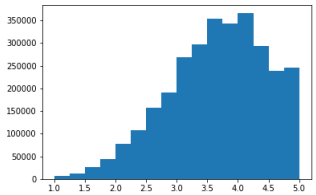
Figure 4. Normally distributed Posts per user within the subset of data.



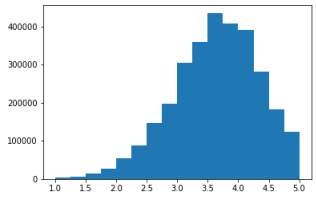
(a) Openness



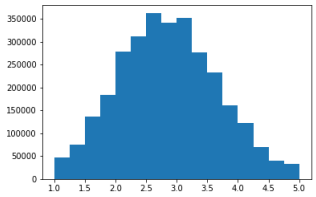
(b) Conscientiousness



(c) Extraversion

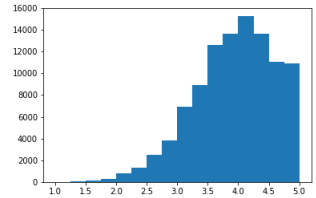


(d) Agreeableness

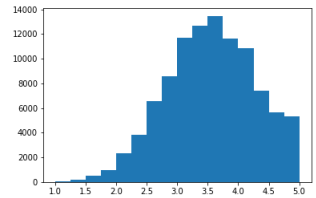


(e) Neuroticism

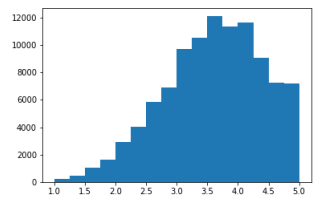
Figure 5. OCEAN Traits for each of the 5 personality traits for all subjects that took the personality instrument.



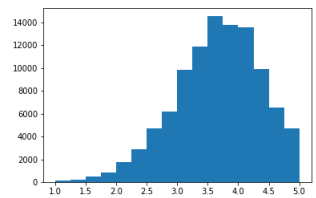
(a) Openness



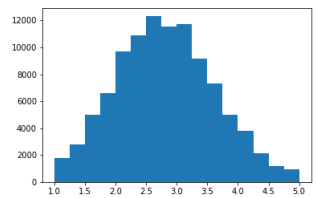
(b) Conscientiousness



(c) Extraversion



(d) Agreeableness



(e) Neuroticism

Figure 6. OCEAN Traits for each of the 5 personality traits for only the subset of users who have status posts.

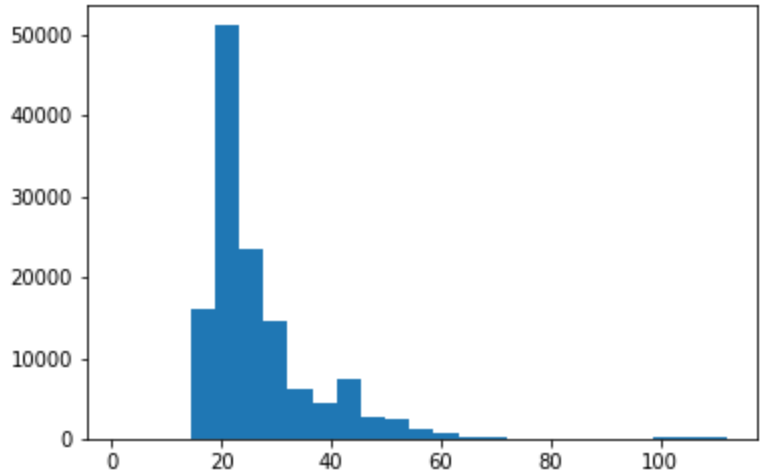


Figure 7. Normally distributed Age of users within the subset of data.

Demographic information such as gender and age helped to determine potential bias in the datasets. The data appears to have a slight over-representation of women in the dataset, containing 86,319 female subjects and 66,470 male subjects, who provided status updates. The users' age distribution appears representative of social media users, and is shown in Figure 7. The age data shows a normal distribution cut off at the age of 18, which is expected, since children are a protected class in research studies. Most of the users in the study fall between the ages of 20 and 30. However, the collection does include older users, and may also contain noise created by users hiding their age, as some ages include users over the age of 110.

Identifying Non-English users

From the demographic information explored in the subset of data, the demographic data would not be useful in identifying potential foreign language data in the dataset, so other methods had to be used to determine the viability of this data for experimentation. The method identified users who spoke primarily another language by looking at the text in their status updates.

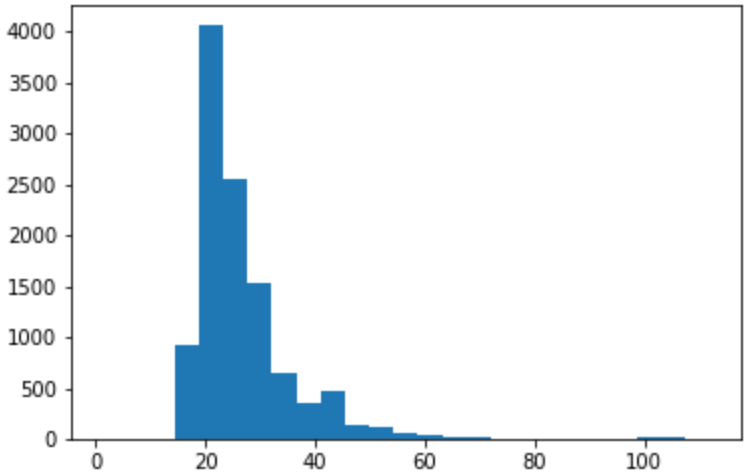


Figure 8. Normally distributed Age of users within the Foreign Language subset of data.

Every status post in the dataset was looped through so that each posts' language could be established using python's LangDetect library, a direct port of Nakatani Shuyo's language detection library for Java [144]. Going through each post in the data allowed each post's language to be determined. This library is not entirely accurate, and the potential for one-off status posts by users who may not speak the language posted, identification of foreign language data did not solely rely on the LangDetect library.

Each post was sorted to their respective authors once each post's language was determined. Each language found for the authors was summed. If most of the posts from an author was determined to be English, we assumed that the author's primary language was English. While this may remove some that might be fluent in multiple languages, we took a conservative approach during this downsizing to ensure that we identified authors with another primary language. A total of 11,829 authors were found whose primary language was not English. During this stage, a file containing each of the authors' posts was created to simplify translations in the next stage.

Further sub-sampling of authors to only those who spoke a non-English primary

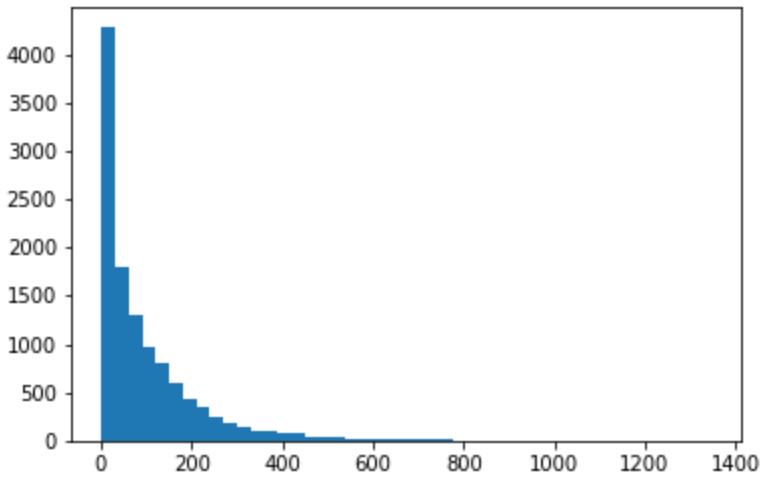
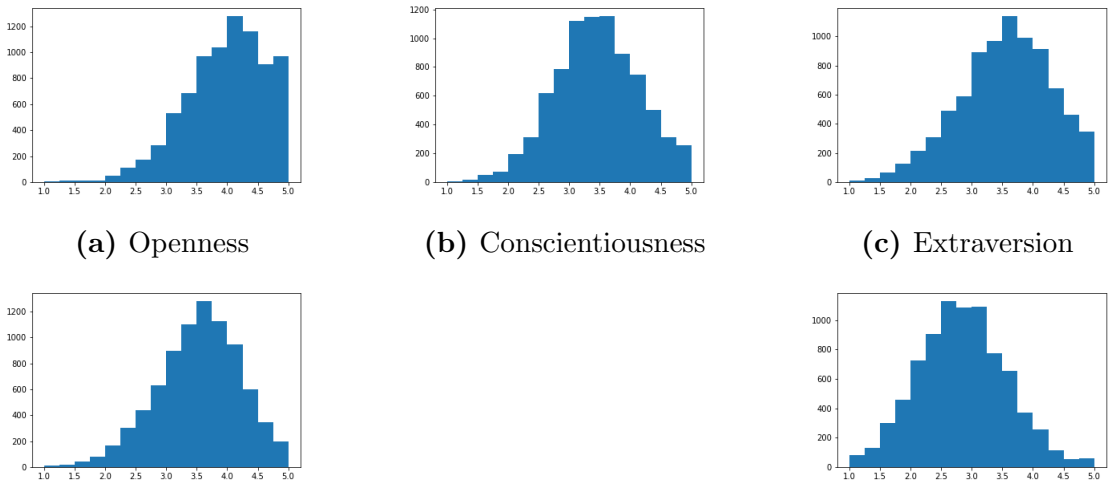


Figure 9. Normally distributed Posts per user within the Foreign Language subset of data.



(a) Openness **(b)** Conscientiousness **(c)** Extraversion

(d) Agreeableness **(e)** Neuroticism

Figure 10. OCEAN Traits for each of the 5 personality traits for users who speak a foreign language.

Statistic	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Mean	3.88	3.36	3.45	3.44	2.76
Median	4.00	3.33	3.5	3.5	2.75
Std. Dev.	0.67	0.69	0.75	0.68	0.72
Minimum	1.00	1.00	1.00	1.00	1.00
Maximum	5.00	5.00	5.00	5.00	5.00

Table 5. Basic Statistics for the FFM traits found within the Dataset for only those users who speak a Non-English Primary Language. Numbers are normalized between [1,5]

language did not significantly alter the demographics. The authors within the subset of foreign-language authors had a total of 1,214,565 status updates. The distribution of posts per user is in Figure 9. Age distributions were not significantly altered within the subset either, and is in Figure 8.

Distributions were also not significantly altered by sub-sampling to Foreign Language authors. We see the basic statistics of the FFM traits in Table 5, and the histograms for each of the five traits in Figure 10. Based on this subset, we should be able to reliably use this dataset in our algorithm to determine whether the use of translations of foreign languages will have any significant impact on the estimation of FFM traits.

3 Authorship Attribution

As mentioned in Chapter I, authorship attribution and verification is not necessarily a psychological trait, but is a behavioral biometric that can be informed, and may rely on various psychological traits for decisions in word choices. Thus, the exploration of authorship attribution and verification is of interest in this dissertation as well. Instead of creating new algorithms, in this task, we explore five prominent algorithms in the authorship attribution domain. These five algorithms are, notably, the best-performing algorithms found by Neal et al. [89].

We verify the results of Neal et al. by exploring Five authorship attribution algo-

rithms in Chapter III against the CASIS dataset. The CASIS dataset consists of 4,000 blog samples from 1,000 English speakers, totaling four samples per author. Each sample averages 13 sentences, 304 words, and 1,634 characters. North Carolina Agricultural and Technical University collected the dataset by the Center for Advanced Studies of Identity Science for authorship attribution experimentation. The dataset has helped in the testing of Authorship Attribution algorithms and the creation of tools for adversarial authorship [145, 146, 147].

BOLT Datasets

The BOLT datasets were created by the Linguistic Data Consortium for the DARPA BOLT (Broad Operational Language Translation) program [148]. This program focused on developing machine translation and information retrieval systems with informal genres, primarily user-generated content. While the BOLT datasets focus on machine translations, samples of the data show that LDC retained author usernames from the forums, making the dataset potentially useful in experimentation of authorship attribution against foreign languages.

Two of the collected datasets proved to be useful in this dissertation's experimentation, the Arabic Discussion Forums [149] and the Chinese Discussion Forums [150]. Both datasets were collected in the same manner. Native speaker annotators seeded the collection efforts by scouting web pages for specific content. Scouts sought content in a specific dialect: Egyptian Arabic for the Arabic dataset, and Mandarin Chinese in the Chinese dataset, and were attempting to find original, interactive, and informal conversations on posts. Once the scouts identified an appropriate thread, they would upload a URL and some information to a database. If multiple threads on a site were submitted, the entire forum was scrapped.

With the scouts' information, LDC harvested HTML files from each of the sources and converted them using custom scripts. Anonymization and cleaning occurred,

Number of Posts	Arabic User Count	Chinese User Count
> 25	53677	79101
> 50	36784	42568
> 75	28629	29022
> 100	23689	21803
> 125	20147	17194
> 150	17530	14072
> 175	15429	11928
> 200	13829	10264
> 225	12484	8943
> 250	11383	7887
> 275	10449	6992

Table 6. The Table shows counts of users for authors who posts more than the number of posts listed.

including attempted removal of quotes from the sources, to avoid text that might not be from the specific author. While LDC took measures to ensure that the data contains only the proper dialect and removes any quotes, the authors note that some quotes or other dialects may remain in the dataset [150, 149].

The original Arabic dataset contains 251,581 unique usernames and a total of 13,269,241 total posts across all users. Within this original dataset size, the average number of posts per user is about 52 posts with a standard deviation of 263 posts. We see in the histograms shown in Figure 11 that the distribution is exponential with a minimum of 1 post for a user and a maximum of 28,028 posts for a user. The median of this exponential distribution is 4 Posts per user. We see the histograms and medians of the distributions that users tend to fall between 1 and 10 posts per user. While this does not discount the dataset for authorship verification, it does show that much of the data is not ideal for authorship attribution or verification.

There is a sizable set of users who had a large number of posts within the dataset. The Counts of users with minimum numbers of posts are shown in Table 6. As the number of minimum posts increases, we see a drop-off in minimum users (as is expected in an exponential distribution), but there are 10,449 users with greater than

275 posts within the BOLT Arabic dataset.

We see similar effects occur in the BOLT Chinese dataset. The original Chinese dataset contains a total of 1,331,569 unique usernames, with 15,160,154 total posts. The average posts per user is significantly smaller than the Arabic dataset at 11 posts with a standard deviation of 204 posts. As shown in Figure 12, the distribution follows an exponential distribution with a median of 2 posts and a maximum of 191,651 posts for a single user.

The drop-off users occurs much quicker than the Arabic dataset, as shown in Table 6. The number of users with more than 25 posts has 26,000 more users than in the Arabic dataset. As we increase the number of posts, we see a sharper decline, requiring only 200 posts to get a list of 10,000 possible users. More data is collected in the BOLT Chinese dataset, but this dataset for authorship attribution or verification will require less overall data per author to be usable for the experimentation.

Choosing users

LDC provides the dataset in two formats, HTML and XML. While both are usable, the XML data provides a small size to work with while maintaining the author's original text. A script was run to loop through all of the files in each of the collections from LDC to obtain individual posts. The authors of posts were obtained from the XML files. Each post was sorted by their author to determine the number of posts each author had in the dataset. This method also allowed for the determination of the dataset's value for potential authorship attribution or verification experiments. A file was created containing posts of each author for analysis based on the outcomes of the counts of posts.

Instead of using the number of posts by the authors, it was decided that number of characters would give a better understanding of the potential useful samples in the dataset. Author files that were created were processed and sorted based on the

	Trans. Raw		Cleaned		Subsets	
	Chinese	Arabic	Chinese	Arabic	Chinese	Arabic
Authors	1,206	2,120	1,173	2,083	1,003	492
Posts (or Samples)	903,857	2,179,716	801,001	1,904,106	25,075	12,300
Avg. Words per Post	245.86	58.60	263.13	64.20	1,904.53	1,167.06
Avg. Sentences per Post	13.97	2.23	14.88	2.34	104.82	40.80
Avg. Posts per Author	749.46	1,028.17	682.87	900.29	25	25
Avg. Word per Post per Author	655.97	104.05	678.27	109.54		
Avg. Sentences per Posts per Author	36.12	3.77	37.29	3.93		

Table 7. Statistics for the translated portions of the BOLT datasets for possible use in authorship attribution experimentation.

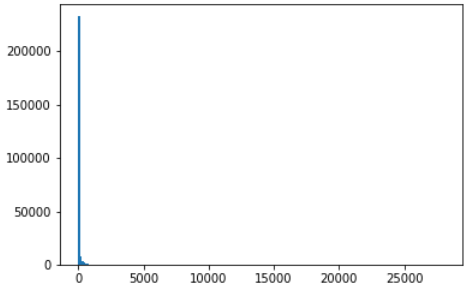
number of characters for every post of the author. From this sorted list, only the top 10,000 users were selected for potential translation. This translation phase will be described further in Chapter IV.

The translation system was run for a smaller subset of those users until more than 1,200 users completed. A more significant subset was not completed due to the length of time needed to complete the translations. Running 1,200 users took multiple weeks to complete, and running in both Chinese and Arabic took several months. Due to a desired sample size large enough to determine whether translations affected authorship attribution, we did not need to complete the translations on all 10,000 users.

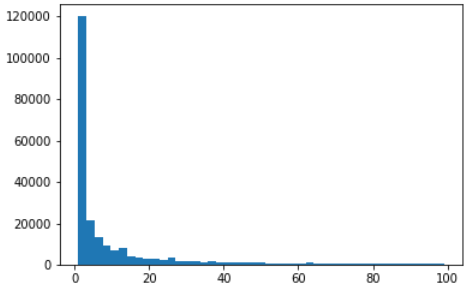
After translation, and with the assistance of colleagues at the University of Florida, the translated dataset was further analyzed, cleaned, and compiled for further experimentation. Some of the data was not properly translated because of the method of translation that was used. These users were thrown out. Within the dataset, a total of 1,202 out of 1,206 translated Chinese users were found to be usable, and 2,118 out of 2,120 translated Arabic users were found to be usable. Further analysis of the translated data is in Table 7.

Further cleaning occurred on the data to remove duplicates in the original text and their corresponding indices in the translated text, any HTML and CSS style tags that remained in the datasets, any standard text or phrases, and any null posts or those

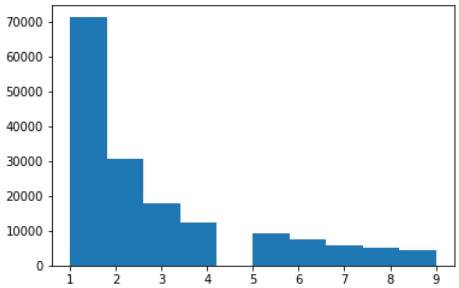
with any foreign characters. We do see a slight improvement in the statistics because of the removal of some of the authors. These improved statistics include the average words per post, average sentences per post, average words per post per author, and average sentences per post per author. However, many of these remain low. Based on prior experimentation, a script was used to further break the dataset into 25 samples per author to allow for increased performance of the algorithms by increasing the number of sentences per sample. Statistics for both cleaning and creating the subsets is in Table 7.



(a) All Posts

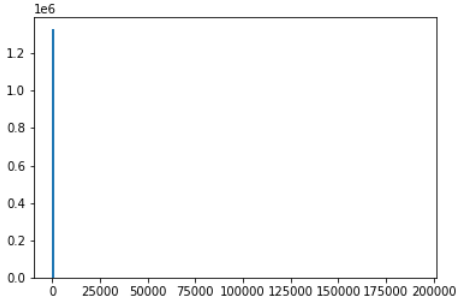


(b) Filtered to 100

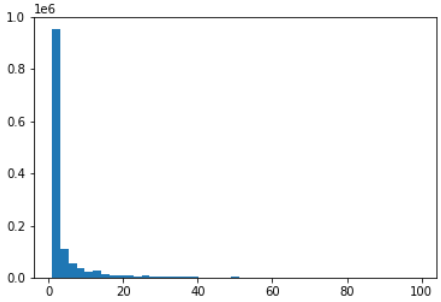


(c) Filtered to 10

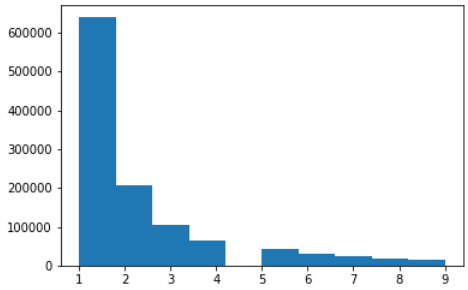
Figure 11. Histogram showing the Posts per user. This shows a histogram for all users (a), and zooms in towards the from of the exponential distribution by filtering authors with less than 100 posts (b) and 10 posts (c)



(a) All Posts



(b) Filtered to 100



(c) Filtered to 10

Figure 12. Histogram showing the Posts per user. This shows a histogram for all users (a), and zooms in towards the from of the exponential distribution by filtering authors with less than 100 posts (b) and 10 posts (c)

CHAPTER III

LINGUISTIC TASKS

Algorithm development for each psychological and behavioral biometric needs to be done before any experimentation on translations can occur. A literature review of techniques and methods used by other authors is conducted for each of the tasks, though Personality Trait Estimation has been explored far less than the others. For both Affect and Personality Trait estimation, we use specific literature that appeared to be most promising and further explored those systems in an attempt to improve the English language system. Ablation testing helped determine some of the potential machine learning hyperparameters and features used in the systems.

Several existing research studies were used for determining the Authorship attribution algorithms used in this dissertation. This section discussed the algorithms used, and the performance of those algorithms against the CASIS dataset.

For both affect analysis and authorship attribution, the performance of the models is tested within the native foreign language. Due to the limitations of the Personality estimation task, exploration into a foreign language using this dataset is impossible. In future research efforts, conducting personality trait estimation in a native language should be explored - especially a non-Latin based language to determine the efficacy of the algorithm in various rooted languages.

1 Affect Analysis

As mentioned in Chapter II, Affect and Sentiment Analysis are tangential tasks that seek to measure specific states of the person being analyzed. The distinguishing factors between the two tasks are the intended measurements. In Sentiment Analysis, the user seeks to identify the target's valence (whether they feel positively or negatively) towards a topic. Affect analysis, on the other hand, attempts to measure the intensities of the emotions expressed by an author as a gauge of their overall emotional state, regardless of the topic being discussed [93]. This gauge of emotional state can lead to a more significant understanding of the person at any given time and help understand how their emotional states might lead to further, potentially malicious actions.

Sentiment Analysis has been more widely researched. The most common method that has been used in Sentiment Analysis involved the use of fixed lexicons. Grefenstette et al. [94], a manually generated lexicon created by Subasic et al. [97] was used to identify words and their associations to emotions. Words like 'gleeful' could be associated with multiple emotions. In their example, they associate the word to both 'Happiness' and 'Excitement.' The words are then weighted according to which emotion the word is most likely to express. The same study also looked at emotive patterns to help in the creation of a seed list. Phrases such as, "looking extremely ...", would be added to the list. The phrases would then be looked for within internet sources to identify the following word used as part of the lexicon. Ma et al. [95] utilized the WordNet-Affect Database [151, 152] to help them study and understand emotions within conversations. Studies from Chuang et al. [98] manually created emotional descriptors based on a Chinese lexicon of 65,620 words. These were then manually mapped to their specific emotions.

While lexicons provide an easy way to map text to emotions, it does not afford the ability for semantic change to be captured. Semantic change is a form of language

change to the evolution of the meaning of words. According to Leonard Bloomfield [153], word meanings can shift over time. This shift can occur for many reasons: when meanings are narrowed or widened, changed by another meaning that is similar to the original meaning, by an elevation or lowering of the status of the original meaning, when the word for part of something begins to be used to reference the whole of the system, when nouns are replaced by something close to the meaning, or when exaggerations and understatements change the meaning of the words. While the most common words rarely change, the meanings of various English words have changed over time. This form of semantic change is common in modern colloquialisms and slang [154]. This limitation in capturing semantic change within lexicons means that lexicons should be manually updated with modern colloquialisms every few years to capture emotion within the new contexts.

Other studies in the area of Sentiment and Emotion Analysis have utilized Word n-grams and parts-of-speech n-grams. Mishne initially focused on the use of frequency counts for word n-grams and parts-of-speech n-grams within a Blog corpus [100]. Later experimentation from Mishne looked at additional Natural Language Processing features, including frequency of special characters and post lengths. While this process would have to be updated for consideration of semantic changes as well, the use of these Natural Language processing features can assist in creating a more automated approach for continually updating the model and automatically capturing semantic change.

Abbasi and Chen [46, 102, 93] utilized many of these features in building their models, utilizing both lexicons, word n-grams, part-of-speech n-grams, character n-grams, and counts of hapax and dis legomina. With these features, the Support Vector Regression model determined the best features for each of the identified emotions. This SVR model was then run against a dataset created from forum posts to identify potential terrorist activity. It is unclear whether their model utilized translations of

the Arabic text or worked within the original Arabic script. In either case, the model proposed by Abbasi and Chen provides the basis for the development of the affect system used in further experimentation in this dissertation.

Algorithm Design

The algorithm proposed by Abbasi et al. [46, 102, 93] appeared to be one of the most promising algorithms for further experimentation. Their proposed algorithm was adapted for the SemEval-2018 dataset discussed in Chapter II. The use of Correlation Ensembles was not possible with the SemEval dataset, so the Correlation Ensemble aspect is not used in the analysis from the analysis. Instead, an Ensemble of Support Vector Regression (SVR) models are used to calculate each of the emotions within the SemEval-2018 dataset.

Abbasi et al. [46, 102, 93] also discuss the features that are used in their algorithm. Abbasi et al. [102] use various types of n-grams, including Character, Word, and Parts-of-Speech n-grams for their analysis, and only up to trigrams (3-grams). In their studies, they also include "hapax legomina" and "dis legomina" collocations. Removal of these collocation occurrences removes words that only happen once or twice, respectively, in the entirety of the dataset, removing sparsely occurring words allows the n-grams to generalize more over the entire dataset.

Before fitting the data to the SVR model for each emotion, numerous preprocessing steps are conducted for the word n-grams and parts-of-speech n-grams. Each tweet is tokenized with a Twitter-aware tokenizer. All tokens are converted to lowercase, contractions are expanded, and all slang words are replaced. All contractions and slang replacements are retokenized to avoid cases where the expansion or replacement introduces additional word tokens into the text. Any tokens consisting of punctuation or that are empty are removed from the token list. Once all of the tokens are obtained, stemming and lemmatization of the words occurs in order to

group inflected forms of words (e.g. "Better" and "Good" share the same lemma, "Running" is stemmed so that it shares a lemma with "run"). Once preprocessing is completed, the tokens can be treated as words. These words are used to calculate the parts-of-speech for use in training the system.

Support Vector Regression uses many of the same principles as Support Vector Machines, with minimal differences. Since the output of SVR models are real numbers, making prediction more difficult since this introduces an infinite possibility space. To resolve this, SVR introduces a margin of tolerance (ϵ). This ϵ is used to help the algorithm only consider those points within the set margin, ignoring those data points with a more significant error rate. The margin of tolerance allows the algorithm to create a better fitting model using whichever kernel. SVR models are created for each of the emotions within the tested datasets.

A correlation ensemble algorithm is used when emotions are correlated. Correlations of emotions means that a single tweet or message contains a list of all emotions calculated in the model. In both the SemEval-2007 and Arabic datasets that are tested, this means that every tweet or news headline contains values for the entire list of Ekman emotions. The algorithm for correlation ensembles is outlined in [102] and rewritten in Algorithm 1.

English Performance

The Kernel, Regularization Parameters, and ϵ for each of the SVR models is optimized using a Grid Search method, an exhaustive search over the hyperparameter space. These parameters include a Linear kernel, Polynomial kernel, and Radial Basis Function kernel in the kernel space; $[0.1, 1, 1.5, 2]$ in the regularization parameter, 'C' space; and $[0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5]$ in the ϵ space. These features are ranked using the Root Mean Squared Error obtained from the regression model.

This grid search is also conducted on the powerset of features tested: character n-

Algorithm 1: Support Vector Regression Correlation Ensemble for use in comparing the emotions of tweets where the training data includes correlations of emotions

Input:

$E :=$ set of emotions

$M :=$ set of training instances

$c \in E$;

$a \in E$;

$c \neq a$;

$m \in M$;

$$Corr(c,a) := \frac{\sum_{x=1}^{m}(c_x - \overline{c})(a_x - \overline{a})}{\sqrt{\sum_{x=1}^{m}(c_x - \overline{c})^2 \sum_{x=1}^{m}(a_x - \overline{a})^2}}$$

$$K = \begin{cases} 1, & \text{if } Corr(c,a) > 0 \\ -1, & \text{Otherwise} \end{cases}$$

$$SVRCE_c(i) := SVR_c(i) + \sum_{a;a\neq c}^{E} (Corr(c,a)^2 (SVR_a(i) - SVR_c(i))K) \ ;$$

grams, word n-grams, parts-of-speech n-grams, and unprocessed word n-grams. Due to Python's limitations, the hyperparameters are selected programmatically, but the feature selection is made manually. For each emotion and feature, a grid search optimization for the hyperparameters is run. The RMSE values are then calculated using a 10-fold cross-validation on the same dataset using a 10% split for the test set. Table 8 shows the results of the grid search optimization. We see that both Anger and Joy provide the lowest RMSE when both character- and word- n-grams are selected; fear is best optimized when using character-, word-, and the vanilla n-grams; and sadness is best optimized when parts-of-speech-, character-, and word- n-grams are used. Table 9 shows the results of the Grid Search optimization of the SemEval-2018 dataset for each hyperparameter and feature optimized.

With the optimized hyperparameters and feature sets, various metrics are calculated again using a 10-fold cross-validation of the algorithm on the dataset with a 10% split for the test set. The Error metrics selected include R^2, when the kernel is

linear; Mean Absolute Error (MAE); Mean Squared Error (MSE); and Root Mean Squared Error (RMSE). R^2 is not reported for non-linear kernels because the method of calculation can result in high values, but, due to the non-linear function used for the regression model, the value could be meaningless in determining the goodness of fit when the model has wide confidence intervals. Table 10 shows the calculated metrics for each of the optimized emotion models.

Based on the metrics in Table 10, we see that the R^2 scores are consistent with many psychological studies at slightly 0.5. Human behaviors are harder to predict, and often result in a lower coefficient of determination. The scores around 0.5 for Anger, Fear, and Joy are consistent with many human behavior studies and show a goodness-of-fit for the linear models that are good with estimations of human behaviors. The R^2 for the Sadness trait is extremely good for predicting sadness within the system as it is nearly a value of one. This high metric may be caused by an over-fitting of the data or indicate how sadness is expressed through twitter. This expression may be shared among a large population of users. Further research into the linguistic and psychological understanding of this identified phenomenon is necessary to ascertain this model's generalizability to datasets outside of the acquired datasets.

Both the Mean Absolute Error (MAE), Mean Squared Errors (MSE), and Root Mean Square Errors (RMSE) are methods that summarize the performance of the system without regard to whether the algorithm over or underpredicted the outcome. The MAE shows that the estimates of the traits falls within that range most of the time. The scales for the emotions are normalized from [0,1] inclusive. An MAE for Anger of 0.1034 means most of the errors observed in the system will be off by plus or minus .10. RMSE provides the same scale that the system operates in - in this case, the RMSE should be a value between [0,1]. RMSE weighs the higher magnitude error rates in the system more than MAE, providing a sense of how large the system's error

	Anger	Fear	Joy	Sadness
{'word'}	0.1364	0.1845	0.2011	0.1593
{'pos'}	0.1739	0.1939	0.2045	0.1958
{'vanilla'}	0.1680	0.1859	0.1971	0.1618
{'char'}	0.1352	0.1426	0.1493	0.1927
{'pos', 'word'}	0.1370	0.1880	0.1572	0.1790
{'word', 'vanilla'}	0.1347	0.1777	0.1522	0.1599
{'pos', 'vanilla'}	0.1395	0.1887	0.1640	0.1683
{'char', 'word'}	0.1284	0.1379	0.1461	0.1900
{'pos', 'char'}	0.1373	0.1502	0.1561	0.0455
{'char', 'vanilla'}	0.1310	0.1389	0.1487	0.1904
{'pos', 'word', 'vanilla'}	0.1349	0.1871	0.1546	0.1641
{'pos', 'char', 'word'}	0.1311	0.1402	0.1489	0.0452
{'char', 'word', 'vanilla'}	0.1302	0.1370	0.1463	0.1883
{'pos', 'char', 'vanilla'}	0.1329	0.1412	0.1515	0.0455
{'pos', 'char', 'word', 'vanilla'}	0.1312	0.1393	0.1484	0.0450

Table 8. This table shows the RMSE values for each of the emotions and the features within the SemEval-2018 dataset. Blue highlighted cells in each column represent the feature combinations that provided the lowest RMSE rate. Other features in the Gridsearch - Regularization parameter, ϵ, and kernel are not shown and were chosen programmatically.

	Features	C	ϵ	Kernel
Anger	{'char', 'word'}	0.10	0.05	Linear
Fear	{'char', 'word', 'vanilla'}	0.10	0.05	Linear
Joy	{'char', 'word'}	0.10	0.05	Linear
Sadness	{'pos', 'char', 'word'}	10.00	0.05	Linear

Table 9. The selected features from a Grid Search Optimization against the SemEval-2018 data. These parameters provide the lowest error rates against twitter data.

	R^2	MAE	MSE	RMSE
Anger	0.4594	0.1034	0.0167	0.1284
Fear	0.4965	0.1090	0.0188	0.1370
Joy	0.4944	0.1163	0.0214	0.1461
Sadness	0.9464	0.0432	0.002	0.0452

Table 10. Performance metrics calculated for each of the optimized English emotion models. Metrics include: R^2 for linear models (N/A is shown when a non-linear model is selected), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

rates are. We see that the RMSE of all of the system increases over the MAE, meaning that while we can expect the errors to average around the RMSE, we can expect that we will have some significantly larger errors in the system from some of the input text. However, these metrics show that the system used for our experimentation is competitive with other Algorithms used against the SemEval-2018 dataset [119] and Abbasi et al.'s algorithm performance against short Fifty-word short story text [102].

Arabic Performance

Since the system built for this experiment is competitive with other emotion estimation systems, a similar methodology was used to determine the effectiveness of these algorithms against the Arabic Emotion dataset. Each hyperparameter and power-set of features were again optimized against the dataset using a grid search method. These parameters again include a Linear kernel, Polynomial kernel, and Radial Basis Function kernel in the kernel space; $[0.1, 1, 2, 5, 10]$ in the regularization parameter, C space; and $[0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5]$ in the ϵ space. This grid search was ranked using the Root Mean Squared Error obtained from the regression model.

There are far fewer tools for experimentation in the Arabic space. Stanford University's CoreNLP [155] is one of the few research tools available for Natural Language Processing in Arabic. CoreNLP implements lemmatization in the language, Parts-of-speech taggers, Named-Entity Recognition, and other tools that Stanford's Natural Language Processing Group has designed. Columbia University has also developed MADAMIRA [156], which improved Morphological Arabic analysis. However, this tool is more complicated to set up and requires access to the Linguistic Data Consortium's Standard Arabic Morphological Analyzer (SAMA) [157]. The University of Maryland [158] also created an improved tool for lemmatization, but this tool requires MADAMIRA as a dependency. For simplicity, analysis in this section is completed using Stanford's CoreNLP. Future research should analyze potential improvements

using the University of Maryland's Arabic Toolkit, combined with Columbia University's MADAMIRA.

Analysis of the Arabic data used character n-grams, word n-grams, and parts-of-speech n-grams. The Stanford CoreNLP tool was used to tokenize words and determine the parts-of-speech in the language. Since other tools used in the English preprocessing were not designed for analysis of foreign languages, we rely solely on the tokenization conducted by CoreNLP. Due to limitations in Python, the hyperparameters are selected programmatically, but the feature selection is made manually. For each emotion and feature, a grid search optimization for the hyperparameters is run. The RMSE values are then calculated using a 10-fold cross-validation on the same dataset using a 10% split for the test set. Table 11 shows the results of the grid search optimization. We see that both Anger and Joy provide the lowest RMSE when both character- and word- n-grams are selected; fear is best optimized when using character-, word-, and the vanilla n-grams; and sadness is best optimized when parts-of-speech-, character-, and word- n-grams are used. Table 12 shows the results of Grid Search optimization against the SemEval-2018 dataset for each hyperparameter and features that were optimized. Performance metrics were again calculated using a 10-fold cross-validation with a 10% holdout and are shown in Table 13.

Based on the metrics in Table 13, we do see a degradation in performance over the English language data. Except for Joy, the R^2 values of Anger, Disgust and Fear now sit in the lower- to mid- 0.3s. These scores indicate that the methods used here are not quite to the same level of the English language preprocessing or feature sets. Future work should explore other methods for preprocessing and feature extraction. The University of Maryland toolkit appears to handle various preprocessing steps similar to the NLP tools in English. This toolkit should provide improvements to the algorithms.

The MAE shows some degradation in emotions as well. The algorithm simply has

	Anger	Disgust	Fear	Joy	Sadness	Surprise
{'word'}	0.2742	0.1763	0.1258	0.3274	0.2470	0.1837
{'pos'}	0.2656	0.1996	0.1448	0.2957	0.2420	0.1839
{'char'}	0.2760	0.1995	0.1455	0.3418	0.2483	0.1838
{'pos', 'word'}	0.2246	0.1770	0.1279	0.3327	0.2477	0.1838
{'char', 'word'}	0.2757	0.1623	0.1191	0.3398	0.2482	0.1838
{'pos', 'char'}	0.2121	0.1643	0.1233	0.2219	0.2484	0.1838
{'pos', 'char', 'word'}	0.2112	0.1637	0.1209	0.3397	0.2482	0.1838

Table 11. This table shows the RMSE values for each of the emotions and the features within the Arabic dataset. Blue highlighted cells in each column represent the feature combinations that provided the lowest RMSE rate. Other features in the Gridsearch - Regularization parameter, ϵ, and kernel are not shown and were chosen programmatically.

	Features	C	ϵ	Kernel
Anger	{'pos', 'char', 'word'}	0.10	0.10	Linear
Disgust	{'char', 'word'}	0.10	0.05	Linear
Fear	{'char', 'word'}	0.10	0.10	Linear
Joy	{'pos', 'char'}	0.10	0.10	Linear
Sadness	{'pos'}	0.10	0.10	RBF
Surprise	{'word'}	1.00	0.05	RBF

Table 12. The selected features from a Grid Search Optimization against the Arabic data. These parameters provide the lowest error rates against arabic twitter data.

	R^2	MAE	MSE	RMSE
Anger	0.3573	0.1448	0.0447	0.2112
Disgust	0.3339	0.0831	0.0264	0.1623
Fear	0.3193	0.0539	0.0143	0.1191
Joy	0.4704	0.1624	0.0492	0.2219
Sadness	N/A	0.1654	0.0586	0.2420
Surprise	N/A	0.0954	0.0339	0.1837

Table 13. Performance metrics calculated for each of the optimized Arabic emotion models. Metrics include: R^2 for linear models (N/A is shown when a non-linear model is selected), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

a higher magnitude error rates over the English algorithms; however, the significant increase in the RMSE shows that the errors that are being obtained by the algorithm tend to be heavily weighted toward higher magnitude errors. In the English language analysis, we see that the MAE and RMSE values were relatively close, indicating that there were few higher magnitude errors to weight the RMSE higher. The larger RMSE on the Arabic dataset indicates that there are likely some much higher magnitude errors occurring in the data.

While this algorithm is perhaps not quite at the same level as the same algorithm on English language text, further experiments in this dissertation show that the use of translations significantly impacts results. This finding indicates that further exploration of Emotion analysis in the native language is necessary to improve foreign languages' performance.

2 Personality Trait Estimation

Automatic Personality Recognition, or the inference of the self-reported personality of an individual from behavioral evidence, has slowly been studied, but with the increased usage of social media, psycholinguistic analysis of the text has become a common practice. Tausczik and Pennebaker [159] state it most concisely: "Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. [...] [Words and language] are the medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings."

Youyou's 2015 study [160] found that Computer judgments of personality traits exceed human judgments on the same tasks. Computer judgments correlated (Pearson's r of 0.56) more strongly with participants self-ratings over human judgments (Pearson's r of 0.49). Their study also found that computer models were better at predicting life outcomes and other behaviorally related traits than human judgments.

The use of computer models to estimate personality traits provides not only superior judgments, according to Youyou, but can also help to assist psychologists in analyzing more copious amounts of data, as discussed in Chapter I.

The use of text as a medium for estimating personality traits has been studied since 2005. In 2005, Argamon et al. [161] used written essays and a constructed lexicon using attribute-value taxonomies to estimate both Extraversion and Neuroticism, achieving a 58% accuracy in their Support Vector Machine classifier. Mairesse et al. [162] also utilized written essays and the Linguistic Inquiry and Word Count (LIWC) [159] and the Medical Research Council (MRC) Psycholinguistics Database [163] to estimate all five traits within the Five-Factor Model, achieving Mid-50% accuracy on Extraversion, Agreeableness, Conscientiousness, and Neuroticism, and 62% accuracy on Openness. While both of these methods relied on the use of fixed Lexicons, Oberlander and Nowson [164, 165] used blog posts and word n-grams to achieve slightly higher accuracy over the Lexicon methods, achieving mid-50% accuracy for Extraversion and Neuroticism, while achieving 61% accuracy on Agreeableness and 65% accuracy for Conscientiousness.

Social media platforms have been more commonly studied since at least 2011, as social media provides unique platforms for the study of social sciences. Golbeck, Robles, and Turner [166] used a small sample of Facebook Profiles to estimate the Big 5 personality traits. The features include: profile images, whether their profile was set to block anonymous users; activities of the users; favorite books; the number of friends; Egocentric Network Density; and various linguistic features. With these features, the authors were able to achieve Mean Absolute Errors (MAE) of 0.10 for Openness; 0.10 for Conscientiousness; 0.14 for Extraversion; 0.11 for Agreeableness; and 0.13 for Neuroticism. MAE often underweights higher magnitude errors that might be observed in the system, providing inaccurate results for what to expect. However, the use of other non-linguistic features might also play a role in the perfor-

mance of their estimation of the Big 5 traits.

Golbeck et al. [167] and Quercia et al. [168] also looked at the use of Twitter as a mechanism for estimating personality traits in 2011. Golbeck et al. used a set of 2,000 tweets per their 279 authors. Their features included profile information provided by Twitter and the use of LIWC and MRC lexicons for estimating personality traits. Their system achieved MAEs of 0.12 for Openness; 0.14 for Conscientiousness; 0.16 for Extraversion; 0.13 for Agreeableness; and 0.18 for Neuroticism. Quercia et al. estimated personality traits just based on 335 twitter profiles, using features such as the number of followers, number of people following, and listed counts. No linguistic analysis was conducted during this study since the authors were attempting to estimate the personality of private twitter users. They achieved a Root-Mean-Squared-Error (RMSE) of 0.69 for Openness; 0.76 for Conscientiousness; 0.88 for Extraversion; 0.79 for Agreeableness; and 0.85 for Neuroticism. The use of RMSE provides better insight into the overall performance of the human prediction task since it emphasizes higher magnitude errors.

A Computational Personality Recognition Workshop was held at a 2013 AAAI conference [169]. Several interesting findings came out of this workshop: feature selection was most effective when a ranking algorithm was used over the initial feature space; use solely on word n-grams was not very useful; use of lexical resources (lexicons) in general seemed to be the most useful; ensemble methods seemed to perform the best in the prediction of personality traits; that cross-domain learning (e.g., using essays to train against Facebook status) is possible. These findings played an essential role in selecting features in the algorithm used in experimentation in this dissertation.

The algorithm used in this dissertation is meant to estimate personality solely based on a linguistic analysis. While it has been shown that other profile features may provide additional insight into personality, one cannot assume that the profile in-

formation is always available. This assumption that a profile would be provided is especially wrong on various dark web platforms where users commonly seek anonymity for their activities. However, understanding their personality is still an important aspect.

A proprietary algorithm is used for further experimentation around the estimation of personality traits. This section will discuss details about the algorithm's creation, and provide some insight into the performance of the system when tested against the dataset described in Chapter II. RMSE was the primary measure for the task to ensure that we consider the overall performance and emphasize higher magnitude errors.

Algorithm Design

The algorithm used in our experimentation is based on the work by Farnadi et al. [170], Park et al. [171], and Schwartz et al. [172]. Schwartz et al. [172] used Latent Dirichlet Allocation [173] to generate open-vocabulary topics for all of the documents compared against, allowing for correlations between the topics and the personality traits. The studies of [171] and [172] looked at other features, such as LIWC, but there seemed to be minimal impact overall when LIWC was used in conjunction with topic modeling. The proprietary algorithm used in our experimentation avoids the use of LIWC due to the current costs of the most recent version.

Farnadi et al. [170] provide state of the art RMSE for a personality estimation system. The ensemble created in their system uses various feature sets against each of the Big 5 personality Traits. Their system achieves an RMSE of 0.651 for Openness using Facebook activity, demographic information, LIWC, MRC, SentiStrength Lexicons, and the Structure Programming for Linguistic Cue Extraction (SPLICE) features. A RMSE of 0.717 for Conscientiousness using the same features as Openness. A RMSE of 0.784 for Extraversion using only Facebook Activity and demographic

Algorithm	RMSE
Linear Regression	0.82
Support Vector Regression (RBF Kernel)	0.82
Decision Tree	0.78
K Nearest Neighbors	0.76
Radial Basis Neural Network	0.75

Table 14. Root Mean Squared Error for each of the tested algorithms for selection. Features were the 2000 topics identified in earlier testing.

information. A RMSE of 0.692 for Agreeableness using the same features as Openness. Lastly, a RMSE of 0.768 for Neuroticism using only Facebook Activity data and demographic information. We compare the developed algorithm to Farnadi et al.'s system to compare the algorithm against state of the art.

Various methods that have been found in the literature were tested to determine the optimum. Ablation testing was conducted to determine the optimal number of topics and which algorithms performed the best. Table 14 shows the RMSE values of various algorithms against a 2,000 topic model to determine the most optimal algorithm to use in experimentation. We see through observation that the use of Radial Basis Neural Networks provided the best mean RMSE for all Regression models. This model was chosen to optimize further.

While LDA was used to generate a 2,000 topic feature set, optimization was conducted on the Radial Basis Neural Network to determine the number of topics that provide the optimal number of topics for the training set. Table 15 shows the RMSE values for topics from 200 to 5,000 topics. We see that we start to achieve optimal results between 1,000 and 2,000 topics. While further optimization could produce slightly improved results, 2,000 topics were selected to avoid any potential over-fitting of the model and simplify the selection of the features. These topics were generated using an optimized version of collapsed Gibbs Sampling within the LDA model. While Variational Bayes is a computationally faster approach, the Variational Bayes model often results in a less accurate model.

# Topics	RMSE
200	0.89
500	0.95
750	1.1
1000	0.78
2000	0.75
5000	0.84

Table 15. Root Mean Squared Error for set numbers of topics used to determine the number of topics to use for personality estimation. RMSE given is the average RMSE for all five personality traits.

By running LDA with 2,000 topics against the dataset described in Chapter II, we see that we can achieve a RMSE error comparable to the results in Farnadi et al. Achieving a mean RMSE between the five personality traits of 0.75, utilizing only the linguistic data in the dataset. Further ablation tests of other Natural Language Processing features occurred, and the following features were found to provide additional benefits to the algorithm:

- Age
- Gender
- Total number of words
- Count of Hapax and Dis Legomina
- Yule's K Measure
- M1 and M2 values computed during the calculation of Yule's K measure
- Sichel's Measure
- Brunet's Measure
- Honore's Measure
- Count of Negative Sentiment Words

Trait	RMSE
Openness	0.469
Conscientiousness	0.628
Extraversion	0.721
Agreeableness	0.557
Neuroticism	0.736
Mean All 5	.630

Table 16. Root Mean Squared Error for each personality trait with 2,000 topics and the additional NLP features.

With these 2,011 different features, Singular Value Decomposition is used to reduce the overall feature set to 100 features. These decomposed features become the inputs to the Radial Basis Network, with the outputs of the Network expected as the estimate of the personality trait being tested. Each personality trait is a separate model, creating an ensemble of five algorithms for estimating each of the personality traits. Table 16 shows each of the final RMSE scores for each of the five traits using the described model.

These RMSE values show significant improvements over Farnadi et al., indicating that further optimization of their algorithms could increase scores. While this is interesting, the algorithm described here provides a platform for further testing of the system against translations in Chapter IV. However, further sensitivity analysis is conducted to determine the optimal number. It provides additional insight into the requirements for the dataset to ensure that we can limit the users to some minimal criteria that will provide the more considerable insight into the effect of translations on the system.

Using the optimized algorithm, we test the documents of varying word lengths against each of the personality traits to determine the impact to the RMSE of smaller numbers of words. Figure 13 shows the word sensitivity at different document word lengths. We see that using fewer than 500 words results in a significant increase in the RMSE of each personality trait. We see that each of the lines appears to begin

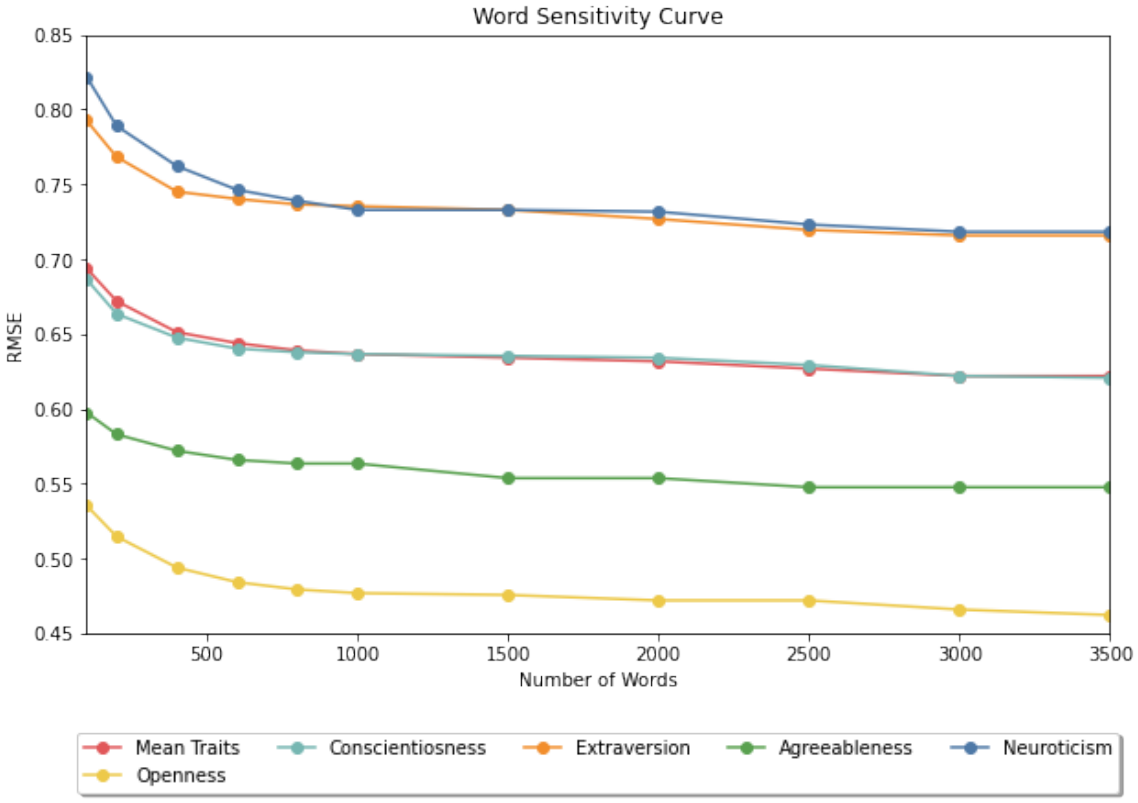


Figure 13. Word Sensitivity for the RMSE of the personality trait estimation algorithm.

to level off at around 1,500 words with only minor gains in RMSE as more words are added. For this reason, a minimum threshold of 1,500 words per document is required to test the system for optimal performance properly. This document word length is used for further experimentation.

3 Authorship Attribution

Authorship Attribution and Verification is a behavioral biometric intended to identify or verify the user based on a stylometric analysis of the author's text. In 2017, Sundararajan et al. [174] and Neal et al. [89] conducted a survey analysis of fourteen prominent authorship attribution algorithms based on the work of Potthast et al. [175]. The University of Florida study was intended to determine the applicability of the fourteen algorithms to large datasets, compare sentence lengths, and outline

the challenges of authorship attribution at a massive scale. These studies found six algorithms that perform well across large datasets.

Algorithms

This section of the dissertation replicates the analysis conducted in [174] to estimate the performance metrics of 4 prominent algorithms in the study. Due to computational resources available in this study, the dissertation then chooses a single algorithm to measure various metrics in English and conduct further studies to determine the effect of translations on the algorithm. A more extensive study based on this work is currently ongoing between this author and the University of Florida. It will measure the effects of translations on all four of the prominent algorithms.

Four algorithms were selected from the study. These include the Keselj [176], Koppel [177], Stamatatos [178], and Teahan [179] algorithms. Each of these algorithms uses different techniques for determining authorship.

The Keselj algorithm [176] uses pairs of character n-grams and the corresponding frequencies of those n-grams. The algorithm was designed to be language-agnostic and was tested by Keselj on English, Greek, and Chinese datasets. The decisions made by the authors were intentionally made to make the algorithm as language-independent as possible. The work was based on an algorithm proposed in 1976 by Bennett [180], which utilized a similarity metric requiring frequencies of bigram occurrences in "Standard English." Keselj utilized a metric that removes this language-dependent parameter and utilizing measured bigram frequencies. After measuring the frequency of bigram occurrences, two profiles are compared using Algorithm 2 proposed by Keselj. Once the dissimilarity measure is compared, the author that minimizes the dissimilarity measure is matched, and the anonymous text is aligned with that author's profile.

The Stamatatos algorithm [178] is very similar to the Keselj algorithm, using

Algorithm 2: Profile Dissimilary of two author profiles proposed by Keselj [176]

Result: *sum*

$sum \leftarrow 0$

forall *n-grams* $\in profile_1$ *or* $profile_2$ **do**

  Let f_1 be frequency of n-gram in $profile_1$, 0 if not present

  Let f_2 be frequency of n-gram in $profile_2$, 0 if not present

  $sum \leftarrow sum + (2 \times (f_1 - f_2)/(f_1 + f_2))^2$

end

Algorithm 3: Profile Dissimilarity of two author profiles proposed by Stamatatos [178]

Result: *sum*

$sum \leftarrow 0$

forall *n-grams* $\in profile_1$ *or* $profile_2$ **do**

  Let f_1 be frequency of n-gram in $profile_1$, 0 if not present

  Let f_2 be frequency of n-gram in $profile_2$, 0 if not present

  Let f_3 be frequency of n-grams across all documents in training set

  $sum \leftarrow sum + (2 \times (f1 - f2)/(f1 + f2))^2 \times (2 \times (f2 - f3)/(f2 + f3))^2$

end

pairs of n-grams and the corresponding frequencies of those n-grams. The difference between the two algorithms lies in the dissimilarity metric that is used. Stamatatos adds a measure based on the frequency norm of available training data by concatenating all of the training data from all of the authors in the dataset. By changing the distance measure, the algorithm can weight those n-grams that deviate the most from the norm to help determine the author of the system. If the n-gram exactly matches the norm, it would be weighted as a 0 since it does not contribute to identifying the author from other authors. This dissimilarity measure is shown in Algorithm 3.

The Teahan algorithm [179] uses compression based methods for measuring the information content of messages to determine authorship styles. Teahan uses cross-entropy based on the character n-grams of varying lengths, up to 5-grams, with respect to a category model. This cross-entropy method measures the average bits per coded symbol in the documents. Sundararajan [174] and Neal [89] found this algorithm,

Algorithm 4: Algorithm proposed by Moshe Koppel in 2011 [177]

Input: document of length $L1$
Data: known documents of length $L2$ for each candidate Author
while $k1$ *times* **do**
    Randomly choose some fraction of the full feature set $k2$
    Determine features in $L1$ matching feature set $k2$
    Find the top match using cosine similarity
end
foreach *candidate author as* A **do**
    $Score(A)$ = proportion of times A is top match
end
Result: $max_A\ Score(A)$ if $max\ Score(A) > \sigma^*$; else $Unknown$

among the fourteen tested, to perform the best when document lengths were fixed. Benedetto [181], another compression algorithm using the LZ77 algorithm to measure document similarity also performed well in the testing. Because of this, compression based methods are interesting, but more understanding into how these algorithms function would need to be gained to understand if they are actually capturing stylistic information of the documents or matching based on topic and genre specific words.

The Koppel algorithm [177] was designed based on Koppel's prior findings in [182]. This algorithm assumes based on that finding that the known text of the author is likely to be the text most similar to the document even as the feature sets are varied. While another author may match on a handful of feature sets, it is unlikely that a non-match would stay consistent of many different features sets. The idea of this algorithm is therefore to check if a given author proves to be the most similar to the test for many different randomly selected feature sets of a fixed size. Koppel therefore proposes the algorithm shown in Algorithm 4.

Overall the findings of [174] and [89] shows that character level features appear to perform the best in large scale authorship attribution systems. However, the concept of style in writing samples is not well understood. The University of Florida is continuing to perform studies in an attempt to better understand the concept of writing styles in short documents, such as social media [183]. Authorship Attribution

and Verification has been shown as a potential for continuous verification by the University of Florida [184], but in order to make such systems accurate enough to be used at a larger scale, a better understanding of the problem needs to occur.

English Performance

This dissertation reproduces the work conducted in [174] to determine the performance of the top-performing algorithms for selection in further experimentation. Due to computational limitations, the speed of the algorithms and memory efficiency of the algorithms are essential factors in the selection of the algorithms as well.

The CASIS dataset, discussed briefly in Chapter II, was used for testing. This dataset was also used in testing in both [174] and [89]. We follow the same processes outlined in their papers, breaking each author sample down into samples containing equal numbers of sentences, including 5, 10, and 20 sentences. While their original papers looked at two sentences per sample, it was shown that cutting the samples down this far degrades the algorithms' performance, likely due to some sentences being useless for stylometric analysis, which is explored in later studies [183]. This breakdown, based on the number of sentences, was achieved by ensuring that the new samples contained no overlapping sentences from the original documents.

With this breakdown, we have a total of 378 authors with 6,691 samples in the five-sentence dataset; 103 authors with a total of 1,597 samples in the ten-sentence dataset; and 19 authors with 229 Samples in the twenty-sentence dataset; the original dataset contains 1,000 authors with 4,000 samples. These breakdowns match the datasets outlines in [174]. For each of the four tested algorithms, a 5-fold cross-validation was performed using each of the datasets (5, 10, 20, and original). Recognition accuracy was measured as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

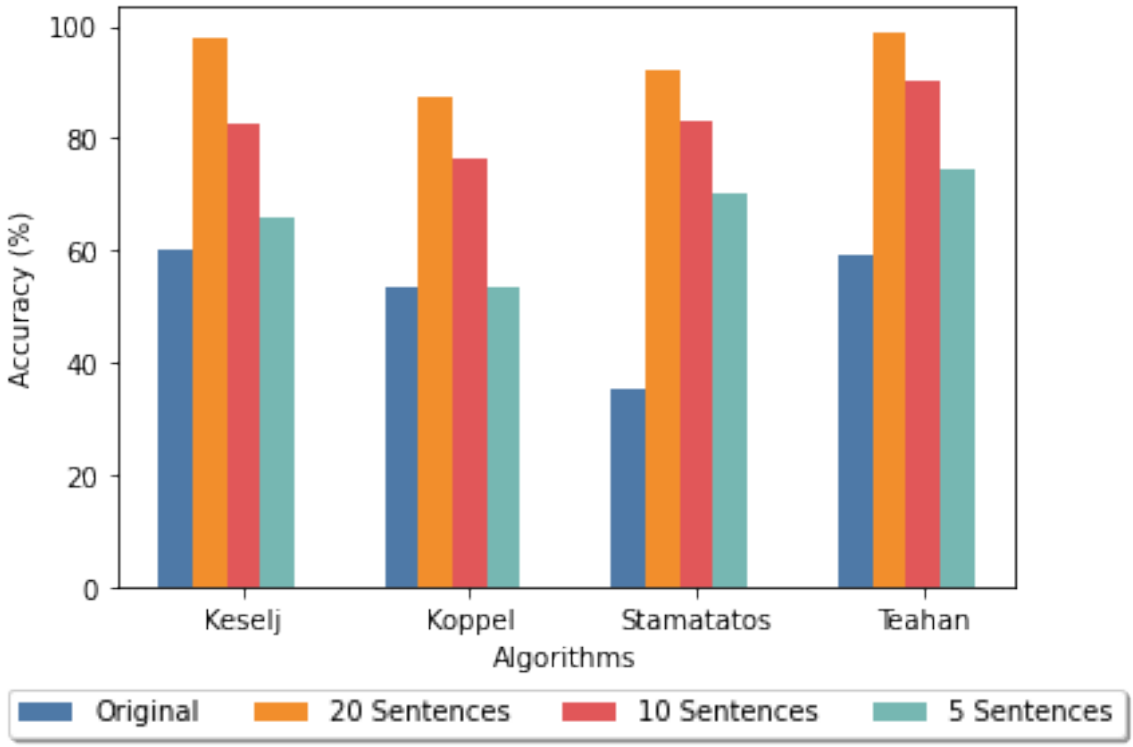


Figure 14. Algorithm accuracy against the CASIS dataset for the four primary algorithms of focus: Keselj, Koppel, Stamatatos, and Teahan algorithms.

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. The accuracy of the four algorithms is shown in Figure 14.

The accuracy of the algorithms in this testing is comparable to the results achieved in [174] on the same CASIS dataset. Teahan performs slightly better than other systems, though all of the systems perform well on the experiments. Due to computational resources, we down-selected two algorithms based on the computational complexity of the Koppel algorithm and the memory resources required for the Teahan algorithm. On more massive datasets, the Computational complexity of the algorithm (and thereby clock time required) was far greater than the other algorithms. Running the algorithm on the Arabic dataset of 12,300 samples took about 15 minutes of processing per iteration. Future experiments rely on a large number of iterations, and even parallelizing the iterations on commodity CPUs would require months of processing.

While the Teahan algorithm performs the best in the experiments, the amount of memory required far exceeds most commodity hardware and frequently crashed the python kernel with more massive datasets on commodity hardware, often exceeding the 24Gb of memory used in experimentation for this dissertation on the Arabic dataset.

The Keselj and Stamatatos algorithms perform well based on their accuracy and are speedy algorithms, even on commodity hardware. Based on the results, we see that the Keselj algorithm performs slightly better on the CASIS datasets with about 60.4% accuracy on the Original dataset, compared to the 35.4% accuracy of the Stamatatos algorithm on the same dataset. The Stamatatos algorithm does perform slightly better on the five-sentence dataset, but the accuracy on the ten-sentence dataset is comparable, and the Keselj algorithm performs slightly better on the twenty-sentence dataset (97.8% accuracy vs. 92.1% on the Stamatatos algorithm).

Because of the slightly better performance, further experimentation in this dissertation focuses only on the Keselj algorithm. Future research experiments will look at the performance impacts of translation on all four of the algorithms. Further analysis of the algorithms is performed. The False Acceptance Rate and False Rejection Rates are calculated and plotted on a ROC curve using Cosine similarity measure, shown in Figure 15. We see that based on the ROC curve, the algorithm does perform reasonably well, though the 20 sentence dataset does perform slightly better overall. Using these metrics, we can calculate the Equal Error Rates (EER) for each of the datasets. The average EERs for the datasets are: 23.8% for the original dataset, 25.7% for the twenty-sentence dataset, 27.4% for the ten-sentence dataset, and 33.0% for the five-sentence dataset.

Lastly, we can calculate the sensitivity of the keselj algorithm and plot the genuine vs. imposter distributions. Ideally, we would want to see that these distributions do not have significant overlaps. When there is little overlap, we can easily distinguish

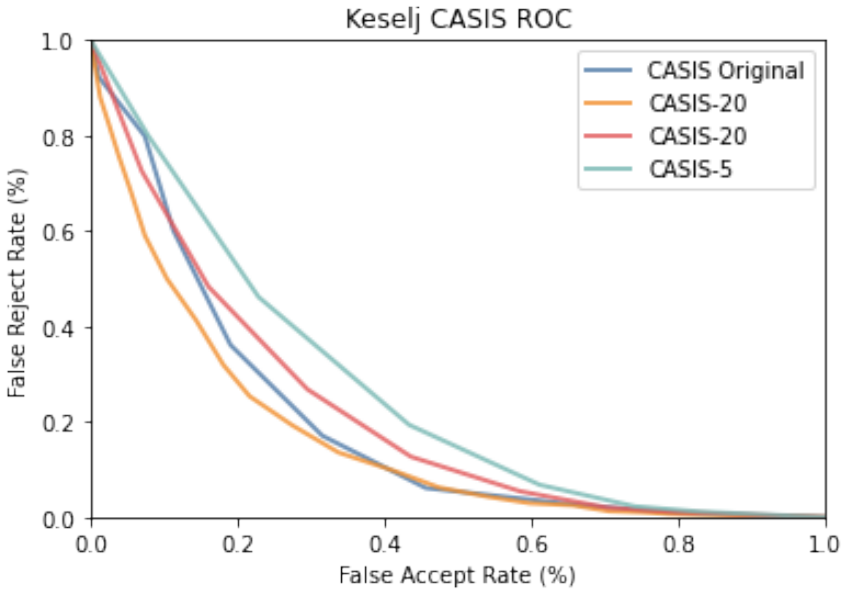
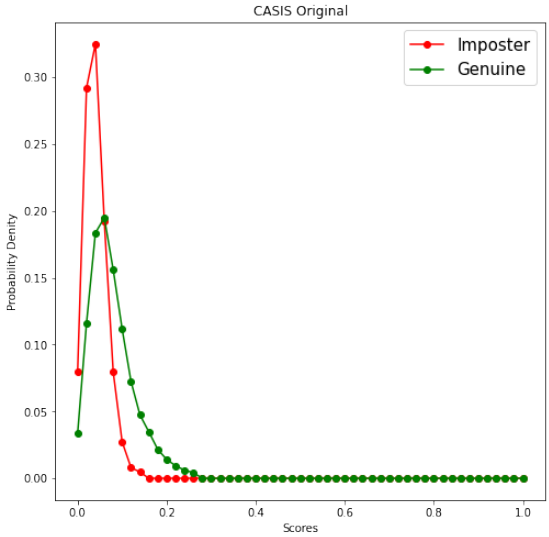
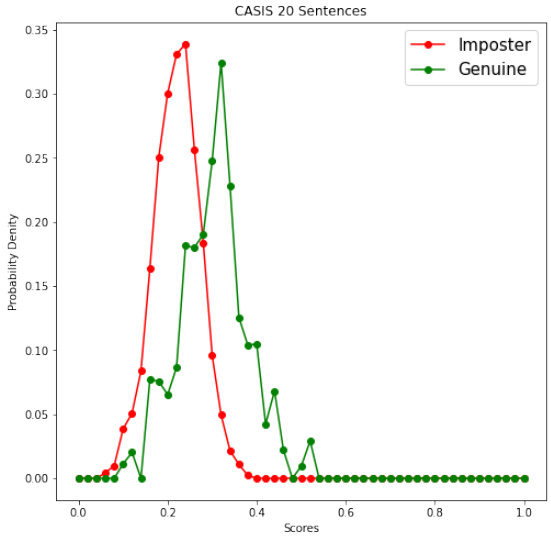


Figure 15. ROC Curve of FAR to FRR showing Curves for number of CASIS sentences.

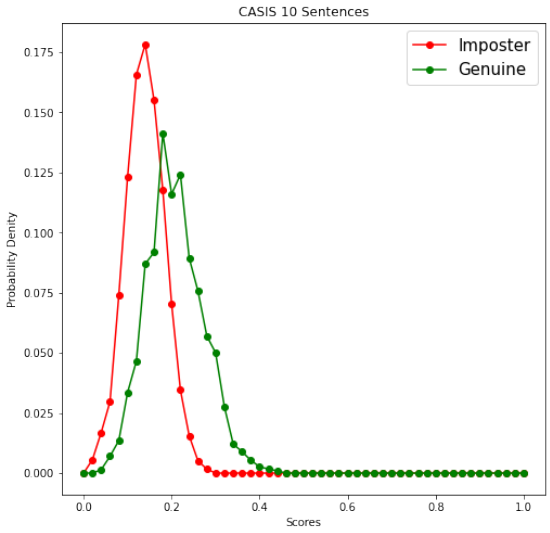
genuine authors from imposter authors. However, as shown in Figure 16, we see that the keselj algorithm has a significant overlap. This overlap indicates that authorship attribution is a challenging problem, even in the English language. Calculating the Sensitivity index (d') against the datasets is also done: 0.923 for the original dataset; 1.258 for the twenty-sentence dataset; 1.213 for the ten-sentence dataset; and 1.106 for the five-sentence dataset. Overall, amongst all of the algorithms tested, holding the number of sentences constant improves the sensitivity index, ensuring that the algorithms can distinguish the genuine authors from the imposter authors. This finding shows that the Arabic and Chinese datasets were fixed to a 25 sentence sample minimum for each of the authors.



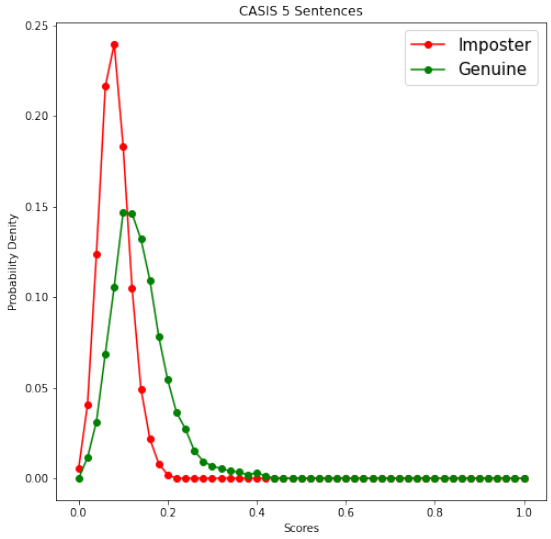
(a) CASIS Original



(b) 20 Sentences



(c) 10 Sentences



(d) 5 Sentences

Figure 16. The probability density charts for the Keselj algorithm showing the genuine vs imposter distributions within subsets of the CASIS dataset.

CHAPTER IV

EFFECTS OF TRANSLATION

The datasets and algorithms used in experimentation were discussed in Chapters II and III, respectively. These datasets allow for setting up the experimentation necessary to determine the effects of translations on the algorithms. To do this, we must first translate any of the foreign language datasets using a translation engine.

Both the Affect and Personality algorithms use regression models to estimate the scores from the text. We can use the error rates to test the null hypothesis for each emotion and personality trait. This null hypothesis is expressed in H_0.

> **H_0** (Null hypothesis)**:** There is no statistically significant difference between the means (or medians, in the case of any distribution except normal) of the distributions of the error rates for each of the emotions or five-factor model traits between the English language and translated text.

Authorship verification, on the other hand, uses different metrics as the system is not a regression model. We use the Equal Error Rate to determine the effect of translation on the behavioral biometric system. Future experimentation should look at the sensitivity index (d') to determine whether translations impact the sensitivity of the algorithms.

Languages			Multi-lingual dataset				Dataset (No EN)	
ISO 639-1	**Language**	**Family**	**# Authors**	**# Posts**	**# Posts (EN)**	**% EN**	**# Authors (No EN)**	**# Posts (No EN)**
es	Spanish	Indo-European	784	95456	25354	20.99%	598	73990
pt	Portugese	Indo-European	212	26107	5958	18.58%	113	13556
ro	Romanian	Indo-European	141	15807	2425	13.30%	68	12322
fr	French	Indo-European	174	20761	4150	16.66%	22	3830
it	Italian	Indo-European	103	17380	2687	13.39%	25	3822
no	Norwegian	Indo-European	29	4725	1267	21.14%	146	18941
cs	Czech	Indo-European	26	3961	620	13.53%	44	5219
pl	Polish	Indo-European	65	6934	1400	16.80%	111	15211
tl	Tigrinya	Afro-Asiatic	175	21614	12594	36.82%	119	24881
sv	Swedish	Indo-European	147	29433	2965	9.15%	82	13573
id	Indonesian	Austronesian	115	19130	5172	21.28%	21	2574
hu	Hungarian	Uralic	32	3844	879	18.61%	22	3015
de	German	Indo-European	54	7391	1911	20.54%	18	3393
nl	Dutch	Indo-European	66	8307	2164	20.67%	22	2408
sk	Slovak	Indo-European	18	3393	357	9.52%	39	7420
et	Estonian	Uralic	21	2287	344	13.07%	6	717
vi	Vietnamese	Austroasiatic	27	2912	970	24.99%	50	6166
fi	Finnish	Uralic	48	9193	1251	11.98%	15	1622
sl	Slovenian	Indo-European	9	1130	318	21.96%	41	4556
af	Afrikaans	Indo-European	58	6563	2326	26.17%	23	2472
hr	Croatian	Indo-European	35	3743	729	16.30%	41	5813
da	Danish	Indo-European	50	6692	1066	13.74%	15	1537
sq	Albanian	Indo-European	5	337	167	33.13%	2	150
so	Somali	Afro-Asiatic	2	82	77	48.43%	1	52

Table 17. Information about the languages that were identified in the dataset, along with a basic counts for each of the two created datasets, including a multi-lingual (containing some percentage of English posts) and a No English dataset. Since the original collection was targeted towards English speakers, many of the languages of authors in the dataset are Indo-European Family languages.

1 Data Translations

Google translate [185] was chosen as the primary translation engine for testing the algorithms. While other translation engines provide similar services, Google translate maintains some of the highest Bi-lingual Evaluation Understudy (BLEU) scores across translation engines on challenging foreign languages, like Arabic [186]. BLEU is an algorithm used for evaluating the quality of machine-translated text from one language to another. The BLEU algorithm is one of the first algorithms to claim a high correlation with human judgments, and has remained one of the most popular metrics for estimating the quality of translations [187].

Along with the high BLEU scores across numerous languages, Google translate also has far more bi-directional language translation models (e.g., English $\rightarrow$ Arabic $\rightarrow$ English, English $\rightarrow$ Chinese $\rightarrow$ English), ensuring that the data preprocessing for the experiment could capture a far greater number of possible languages. Google

translate also has a built-in system for detecting the language for translations. Both the large number of language models and the ability to automatically detect the language was especially important in the translation of the personality dataset, which included 24 unique languages.

Google has also built the tool into its Chrome browser to automatically translate foreign language pages to another language for their users. For cost savings, this feature was utilized by creating unique web pages for each of the author files and using JavaScript and a Python Flask server to automatically go through each web page, scroll the page to ensure chrome translated the entire web page to English, then captured the table and passed it back to the Flask server to save the file as XML. This XML file was then parsed and associated with each of the author's posts and respective metadata (including any associated scores).

Affect Dataset

The affect dataset was much smaller than the other datasets translated. Translation of this occurred directly through Google translate's cloud API, improving the likelihood of missing some of the translations using the web browser approach described. Since the entire dataset collected involved a single dialect, the API was easy to use and only cost $20 to run all 15,000+ collected tweets. Translation was accomplished by looping through each of the collected Arabic tweets and passing the text to Google translate. The response was recorded with the original text and the emotion scores to ensure that each translation was adequately associated with the annotated scores for further experimentation.

Personality Dataset

The 11,829 authors, who identified as having a non-English primary language, were separated from the rest of the dataset. Files created for these authors were trans-

lated using the Google translation script. Google's translation engine automatically detected the social media post's language and attempted to translate the text from the author's primary language to English. Once all posts finished translation, a final validation check, using python's LangDetect library, was conducted to ensure that the final text was English. This validation step is necessary, as translation engines may skip some of the authors' posts, either intentionally or unintentionally, if the system could not process the text fast enough. As described in Chapter III, 1,500 words is the optimum for the developed personality estimation algorithm. During this validation phase, each of the author's posts was counted for their number of words to ensure that the minimum word threshold was present. After removing any authors that failed the translation and had too few words, 2,396 authors met all criteria for experimentation.

Further preprocessing broke the authors down by their primary languages. A count of the number of authors for each language is in table 17. During the validation phase, a handful of posts remained that still classified as English posts. The percentage of English posts in this dataset are in the last column of table 17 for each language. Since the dataset described in section II was initially targeted towards English speaking countries, it is worth noting in the table that many of the foreign languages identified are of the Indo-European family of languages, especially Latin rooted languages.

Experiments were run on this multi-lingual dataset to test the impact on multi-language speakers. However, for more in-depth experimentation, English posts were removed to run similar experiments on author posts that contained no English. After removing the English posts from the identified authors and validating the minimum criteria was met, only 1,644 authors remained in the "No English" dataset. The remaining language author counts are in table 17 under the columns "# Authors (No EN)" and "# Posts (No EN)."

Authorship Dataset

The size of the authorship dataset was also cost-prohibitive and used the web page translation method described earlier in this section. Much like the personality data, an author file was created containing all of the social media posts associated with each author. This author file allowed for a dynamic web page to load the posts of each author. These web pages were then translated in the Chrome browser and an XML file written back to the Flask server. The XML files were processed and associated back to the original text and metadata for each of the authors' posts. This dataset was then further processed, as described in Chapter II.

2 Affect Analysis

Due to the smaller size of the English language emotion data sets, some challenges arose in analyzing the English language dataset's performance for comparison to the foreign language data set. Monte Carlo cross-validation (MCCV) helped to estimate the error distributions. MCCV is a common technique in statistical learning that splits the dataset into two subsets through a random sample without replacement. The model is then trained on one portion of the dataset and tested on the other dataset, commonly done in traditional cross-validation methods. This method is repeated to estimate the distributions [188, 189], minimally for N^2 iterations to get a distribution as close to cross-validation over the unique datasets [190]. This method improved results in chemometric domains [191] over the commonly used "Leave-one-out" method.

The algorithm was trained on the English Language tweets within the dataset, with a random holdout of 10 tweets. These ten tweets calculated the error rates for the algorithm. The number of runs was calculated using equation 4, and the total number of samples was calculated using equation 5. Only four of the six Ekman

emotions are tested due to the limitations of the SemEval-2018 dataset.

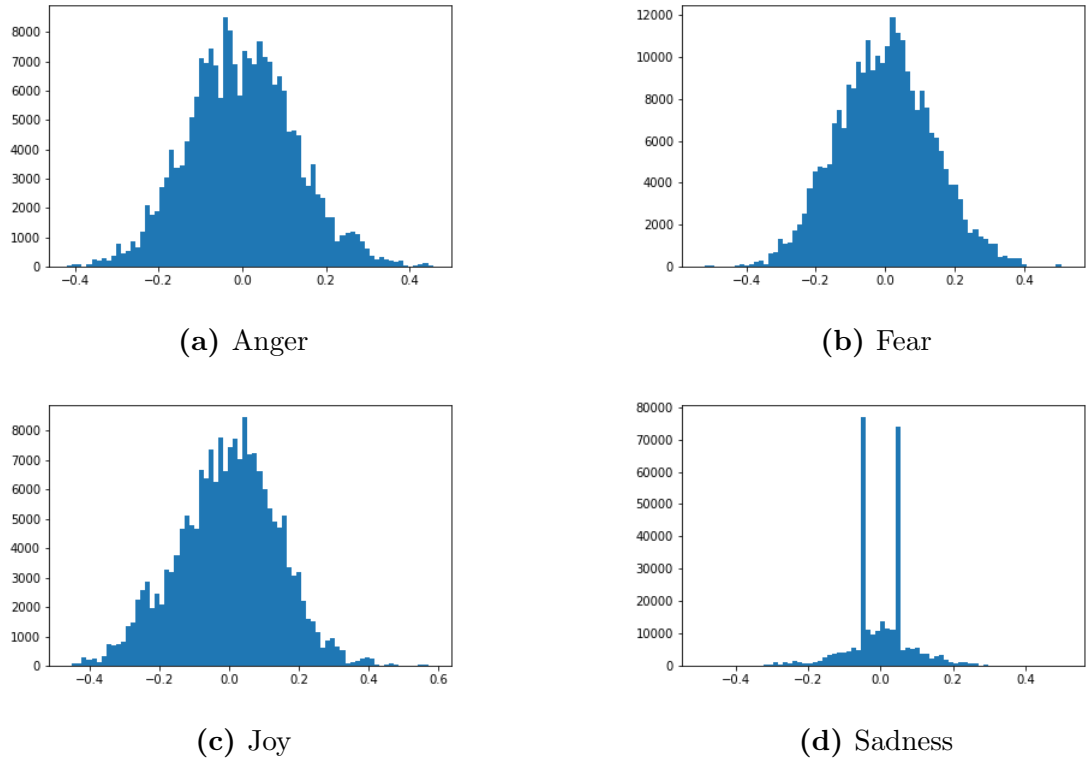
$$|R| = |P \setminus H| * |H| \tag{4}$$

$$|S| = |R| * |H| \tag{5}$$

Where R is the set of runs, P is the original dataset, H is the holdout set, and S is the set of samples. || represents the cardinality, or length of the sets. $\setminus$ represents the set difference; in this equation, the set difference removes all of the holdouts, H, from the original dataset, P. Using the SemEval-2018 dataset, we can calculate the number of samples acquired using this technique using the equations. The calculation for each emotion is shown in equation 6. Using this MCCV method allows us to sample the model's error rates using a smaller set of data.

$$\begin{aligned}
\text{Anger: } & 2089 * 10 = 20,890 \text{ runs} * 10 = 208,900 \text{ samples} \\
\text{Fear: } & 2641 * 10 = 26,410 \text{ runs} * 10 = 264,100 \text{ samples} \\
\text{Joy: } & 1906 * 10 = 19,060 \text{ runs} * 10 = 190,600 \text{ samples} \\
\text{Sadness: } & 3066 * 10 = 30,660 \text{ runs} * 10 = 306,600 \text{ samples}
\end{aligned} \tag{6}$$

This method was not necessary for the analysis of the Translated Foreign Language dataset. Since the dataset was collected similarly to the SemEval-2018 dataset, we treat this translation dataset as a subset of the English model trained on the entire SemEval-2018 dataset. This collection method means that each translated tweet can run through the English language model and the error rates observed without using the MCCV method. For testing purposes, any tweet within the Translation dataset with a value of zero for a specific emotion was omitted. This omission resulted in a total of 3,440 Anger tweets, 2,766 Sadness tweets, 870 Fear tweets, and 5,388 Joy



(a) Anger **(b)** Fear

(c) Joy **(d)** Sadness

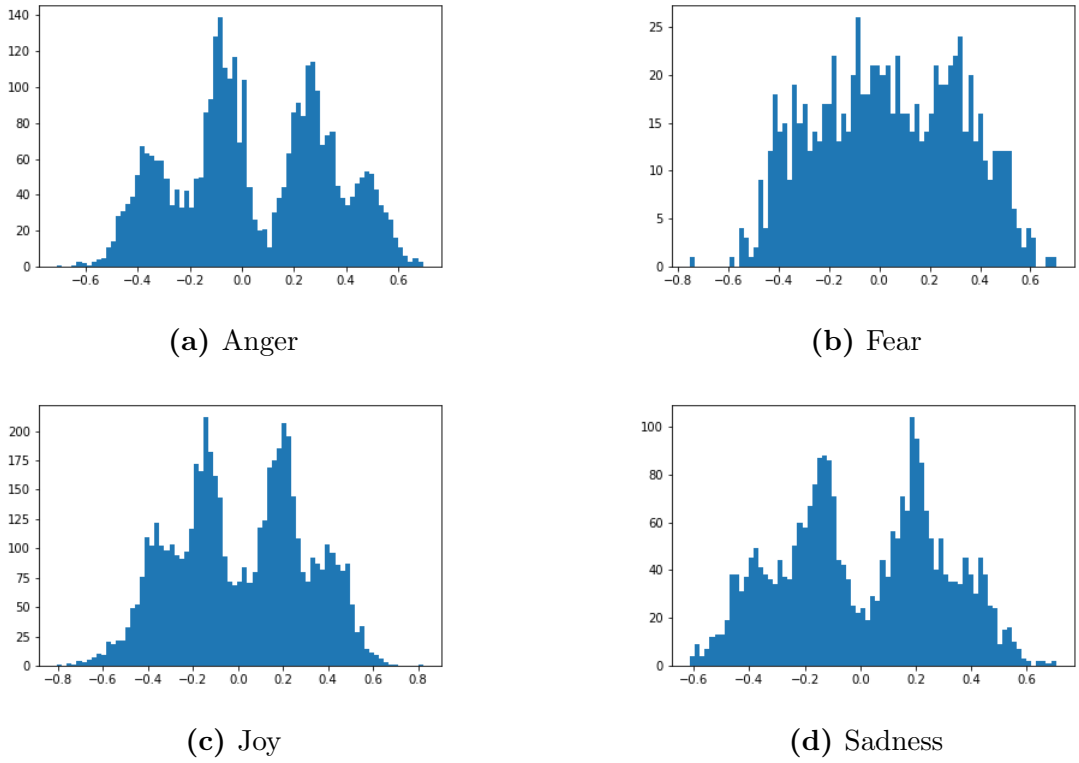
Figure 17. Error Distributions for each of the four tested emotions in English.

tweets within the translated dataset.

Error Analysis

The method for observing the English Dataset provides us with a sample of the Emotion algorithm's error distributions for each of the Four tested emotions. The squared error rate is calculated for each of the four tested emotions by subtracting the estimated value from the algorithm from the expected value of the annotations, then squaring the result. This error rate allows for the exploration of the impact of translations on the overall system.

Distributions of the English dataset for each of the emotions generally follows a normal distribution, except for the sadness distribution, which has unexplained peaks on both sides of the mean. These distributions are in Figure 17. The error squared distribution for the English dataset of all emotions shows an exponential distribution



(a) Anger **(b)** Fear

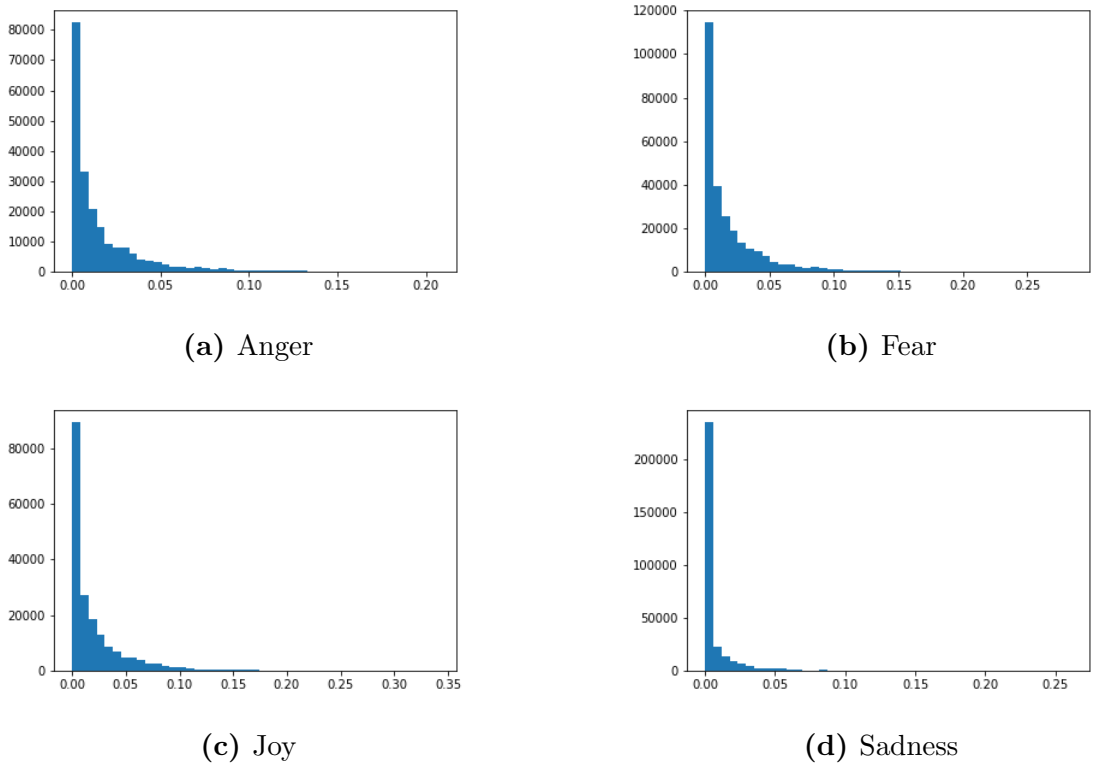
(c) Joy **(d)** Sadness

Figure 18. Error Distributions for each of the four tested emotions from translated Arabic.

19.

The distributions of the Translated dataset were significantly altered through translations. The error rates, in Figure 18, obtained in the translation dataset are no longer normally distributed, but show multiple peaks within the distribution. This new distribution indicates an impact that can be observed by the error distributions of the two sets. The error squared distributions, in Figure 20, show exponential distributions, though the decay of the distribution is visibly much weaker than the English language dataset.

For our experimentation, comparisons of the error distribution can show whether the translations significantly alter the mean of the distribution in any direction. The error squared distribution helps identify any potential changes in the magnitude of the errors.



(a) Anger **(b)** Fear

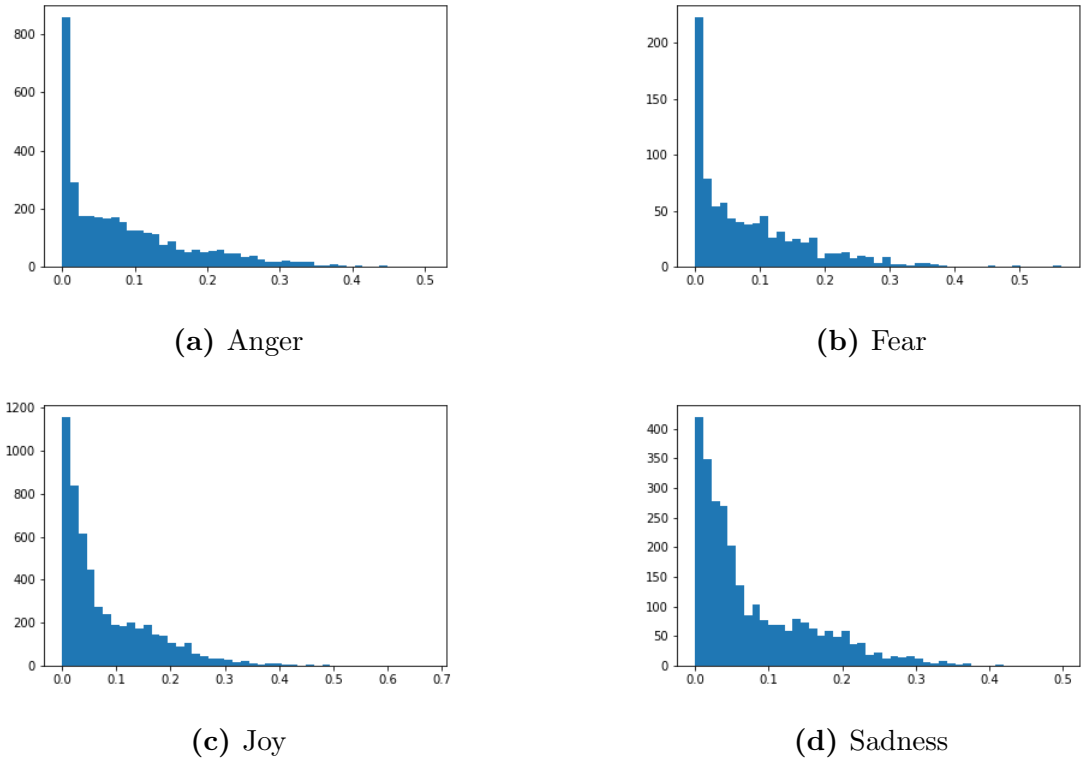
(c) Joy **(d)** Sadness

Figure 19. Error Squared Distributions for each of the four tested emotions in English.

Statistical Analysis

For affect analysis, we analyze both the error and the error-squared distribution to ensure that the observed histogram observations hold statistically. Each distribution for each of the emotions was tested using either a Student's t-test or the Mann-Whitney U-test. The error distributions for Anger, Fear, and joy are Normal distributions, so a Student's t-test is used to analyze the data. The sadness error distribution fails normality tests, so a Mann-Whitney U-test is used for comparisons.

The p-values of the t-tests for Anger, Fear, and Joy is significantly small. Analysis of the error distributions is shown in Table 18. Anger shows a p-value of $1.93 * 10^{-22}$, far below a threshold of 0.05. The Fear distribution has a p-value of 0.0001, and joy has a p-value of $2.64 * 10^{-5}$. All three of the emotions with normal distributions reject the null hypothesis meaning that there is a statistically significant difference



(a) Anger **(b)** Fear

(c) Joy **(d)** Sadness

Figure 20. Error Squared Distributions for each of the four tested emotions from translated Arabic.

between the translated error distributions and the English error distributions. The Mann-Whitney u-test shows a very low probability and results in the machine giving a p-value of 0. This error appears to be an issue with the way the computer handles floating-point numbers; since the p-value is very low, the floating-point numbers are not capable of storing the results, and the computer rounds the number to 0.

We also run correlation analyses between the distributions, shown in Table 19. We see that Pearson's r p-values are above a 0.05 threshold, indicating that no significant correlations were found for the Anger, Fear, and Joy emotions. The Spearman ρ correlation was also found to be insignificant for the sadness distribution. Given that the translation error distributions fail the normality test, the Cohen d metric for measuring the effect size is not appropriate.

The error squared distributions for all of the emotions follows an exponential

Feature	u/t-value	p-value	CI (mean [95% CI])	H_0
anger	-9.8128	1.9341e-22	-4.8495e-02 [-4.8505e-02 to -4.8485e-02]	Reject
fear	-3.8964	0.00010511	-3.7693e-02 [-3.7711e-02 to -3.7676e-02]	Reject
joy	-4.2065	2.6351e-05	-1.6425e-02 [-1.6434e-02 to -1.6416e-02]	Reject
sadness	8.22428e+07	0		Reject

Table 18. Student's t-test results and p-values for each of the Emotions tested to determine whether any of the emotions are statistically significantly different between the Arabic dataset and the English Dataset. Sadness uses the Mann-Whitney U-test, since the sadness distribution fails the normal distribution test.

Feature	Pear. r	Pear. p-val	Spear. Rho	Spear. p-val
anger	0.0106	0.5317	0.0127	0.4549
fear	0.0079	0.8166	0.0015	0.9655
joy	0.005	0.7136	0	0.999
sadness	0.0086	0.6511	0.0117	0.5383

Table 19. Correlation Analysis for each of the Emotions tested to determine whether any of the emotions are statistically correlated in translations between Arabic dataset and English.

distribution. We use the Mann-Whitney U-test to compare the distributions between the translation and English datasets. This analysis is shown in Table 20. We see in each of the emotions that the p-values are significantly low. Anger, Fear, and Sadness error squared distributions result in a 0. As was mentioned earlier, this appears to be caused by floating-point numbers being small enough that the computer rounds the value to 0. The fear distribution has a p-value of $1.18 * 10^{-152}$. The resulting p-values indicate that each error-squared distribution is statistically significantly different, and the null hypothesis is rejected. We also see that a correlation analysis using Spearman's ρ results in no statistically significant correlations, as all p-values fall well above a threshold of 0.05.

Since we know that the distributions are statistically significantly different, the medians show whether translations improve or degrade the algorithms' error. The median for each of the four emotions between the two datasets is compared in table 21. We see in this table that all medians increase when the translated dataset is compared to the English dataset. Since these are exponential distributions, the magnitude

Feature	u-value	p-value	Cohen d	Pear. r	Pear. p-val	Spear. Rho	Spear. p-val
anger	1.63827e+08	0	-1.0656	-0.0087	0.6104	-0.0188	0.2696
fear	5.55645e+07	1.1825e-152	-1.0469	0.0502	0.1393	0.0049	0.8844
joy	2.26379e+08	0	-0.9647	-0.0045	0.7401	-0.0061	0.6534
sadness	8.22428e+07	0	-1.2537	-0.0087	0.6463	-0.013	0.4922

Table 20. Mann-Whitney u-test results, p-values, and Correlation analysis for each of the Emotions' squared error tested to determine whether the magnitude of the errors are statistically significantly different between the Arabic dataset and the English Dataset.

Feature	English Median	Translated Median
anger	0.0075	0.0597
fear	0.0089	0.0557
joy	0.0090	0.0475
sadness	0.0025	0.0480

Table 21. Median of the Error Squared distributions for each of the four tested emotions, showing that the median of the distributions for the translated dataset increase when using translations.

of the errors statistically significantly increases when translations are used in the system. This increase means that one can expect that the translations will result in statistically significantly increased error rates when translations are given to the algorithm.

3 Personality Trait Estimation

Various subsets of the Foreign Language datasets were able to be created from the personality dataset, including a Multi-Lingual set containing only the author's primary language and a handful of subsets to look at specific languages, and a subgroup focused on Linguistic Families. In this section, an analysis of the impact that translations have on the error rate is conducted, testing the hypothesis that Translations have a statistically significant effect on the estimation of personality traits. An analysis further breaks this down by looking at the error rate impact on a Multi-lingual set, sets containing only the author's primary language, a handful of languages that included only a single language (e.g., Spanish, French, and so forth), and finally looking

at the linguistic family of the languages.

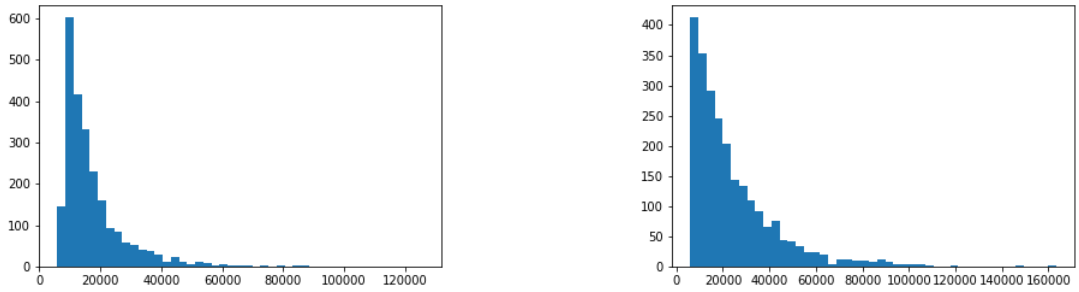
The section then digs into each of the 2,011 features used by the algorithm to determine whether translations have a statistically significant impact on each of the features' distributions. It finally explores the effect on the readability measures to examine whether the translated text has a statistically significant effect on the readability of text comparing English Language subsets and Foreign Language subsets.

A subset of the English data is used for experiments comparing the error rates for statistical significance against the foreign language datasets. This subset contains a list of primarily English speaking authors from the same dataset in Chapter II not used in training the personality estimation algorithm described in Chapter III. This subset allows us to compare sets of unobserved data to determine whether they likely come from the same distribution as one another.

Because the original dataset is large, a random sampling of the English authors is used to select the authors that most closely match the word count distribution of Non-English authors. A histogram showing the character counts of the Non-English speaking authors is in Figure 21a. Observations show that the number of characters the users use follows an exponential distribution with a median of 14,135 characters, a minimum character count of 5,923, and a maximum character count of 125,794.

English authors were randomly selected if they met the minimum character count matching the 5,923 characters observed in the non-English dataset. Figure 21b shows the distribution of the randomly sampled English dataset. This sampled dataset follows an exponential distribution with a median of 18,329 characters, a minimum character count of 5,295, and a maximum character count of 163,228.

With both a subset of English language speakers and the translations of the non-English speakers as described in the section, estimates of the personality traits for each author in each of the datasets were calculated using the algorithm described in Chapter III, and calculate the error rates based on the expected, ground-truthed



(a) Multi-Lingual **(b)** English

Figure 21. Character counts for both the Multi-lingual dataset (a) and English language subsets (b).

values. Each of the resulting features used for determining trait values was stored for later analysis to understand how translations might be impacting the feature sets.

Error Analysis

Following the estimates of the author's personality traits, the squared error rate is calculated for each of the Five-Factor model traits by subtracting the estimated value from the expected value and squaring the result. The error rate is calculated for each of the datasets that are tested, including the English subset. This error rate allows us to explore what impact translations have in the system.

The English dataset distributions for each of the five traits follow a normal distribution, as shown in Figure 22. The error distribution shows the range of the errors without regard to the magnitude of the errors. This distribution allows us to understand whether the translations result in an overall change in the mean error expected from the algorithm. The squared error distribution of the English dataset follows an exponential distribution, as shown in Figure 23. This distribution shows the magnitude of the error. The medians of the distributions help determine whether the magnitude of the errors increase or decrease overall.

The distributions of the Translated Foreign Language datasets show the same dis-

Feature	u-value	p-value	Cohen D	Pearson r	Pearson p-val.	Spearman ρ	Spearman p-val.
Error O	1.75578e+06	0.425248	-0.0447	-0.0063	0.791	0.0053	0.8216
Error C	1.7253e+06	0.134387	0.0434	0.0393	0.096	-0.0107	0.6515
Error E	1.72238e+06	0.116321	0.052	0.0014	0.9513	-0.0453	0.0549
Error A	1.67225e+06	0.003442	0.0281	-0.023	0.3292	-0.0234	0.3206
Error N	1.62227e+06	1.3e-05	0.1703	0.0138	0.5577	0.0109	0.6446

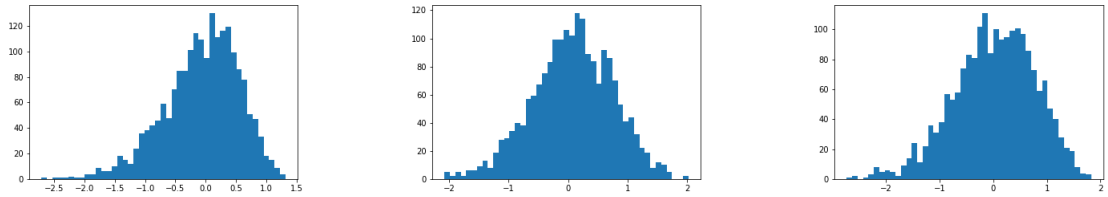
Table 22. P-values for each of the Five Factor Model traits testing whether any of the traits are statistically significantly different between the Multi-lingual dataset and the English Subset.

tributions. The error rates for the five personality traits within the translated dataset have a normal distribution, as shown in Figure 24. The error squared distribution follows an exponential distribution, as shown in Figure 25.

Since the error distribution and error squared distribution types are similar between the two datasets, further analysis has to occur to determine whether translations significantly affect the results obtained by the algorithm. Since we care about whether the magnitude of the error significantly changes, we analyze the error squared distribution. Mann-Whitney U-tests are used to compare the error squared rates obtained between the English and translated distributions to test the null hypothesis, H_0. The Mann-Whitney U-test is a non-parametric statistical method for testing whether two independent samples selected from populations share the same distribution [192, 193], much like the parametric Student's t-test is used for normal distributions.

Multi-lingual Data

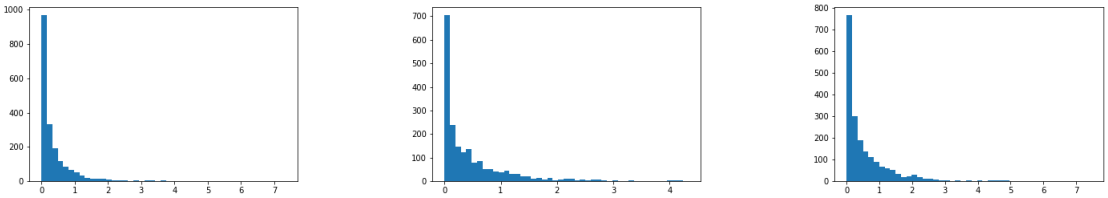
The first dataset used was the multi-lingual dataset created in Chapter II. In this dataset, the authors who spoke a primary language other than English were identified. However, English posts were not excluded from those authors' writing samples, which allowed us to capture some of their English language texts. All of the non-English posts were translated into English. Using the translations and the English posts, the prediction algorithm was run to estimate their psychological traits, comparing this



(a) Openness **(b)** Conscientiousness **(c)** Extraversion

(d) Agreeableness **(e)** Neuroticism

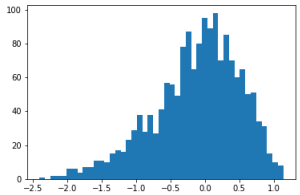
Figure 22. Error Distribution of the estimated OCEAN Traits for each of the 5 personality traits within the English language sample.



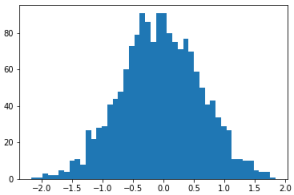
(a) Openness **(b)** Conscientiousness **(c)** Extraversion

(d) Agreeableness **(e)** Neuroticism

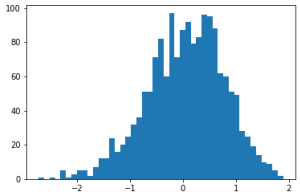
Figure 23. Error Squared Distribution of the estimated OCEAN Traits for each of the 5 personality traits within the English language sample.



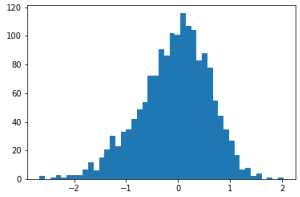
(a) Openness



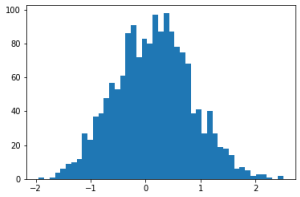
(b) Conscientiousness



(c) Extraversion

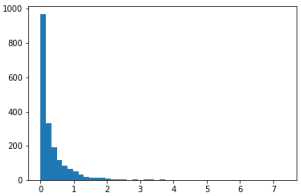


(d) Agreeableness

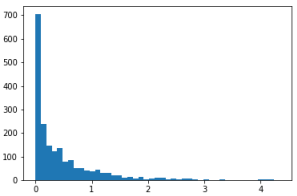


(e) Neuroticism

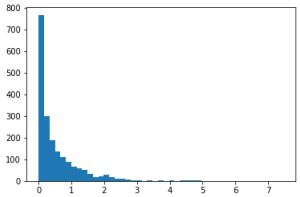
Figure 24. Error Distribution of the estimated OCEAN Traits for each of the 5 personality traits within the Translated sample.



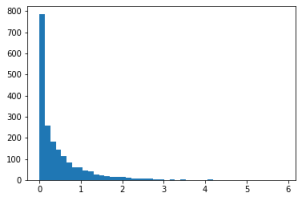
(a) Openness



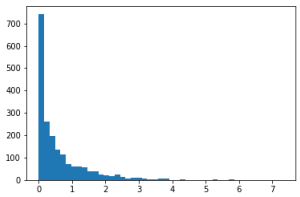
(b) Conscientiousness



(c) Extraversion



(d) Agreeableness



(e) Neuroticism

Figure 25. Error Squared Distribution of the estimated OCEAN Traits for each of the 5 personality traits within the Translated samples.

Feature	u-value	p-value	Cohen D	Pearson r	Pearson p-val.	Spearman ρ	Spearman p-val.
Error O	1.61004e+06	0.4861	-0.0594	0.0141	0.5678	-0.0253	0.3046
Error C	1.58432e+06	0.1945	0.0412	-0.0093	0.7059	-0.0147	0.5505
Error E	1.5772e+06	0.1378	0.0621	-0.006	0.8069	0.0089	0.7178
Error A	1.54009e+06	0.0112	0.0235	0.0044	0.8593	-0.0129	0.6025
Error N	1.49992e+06	0.0002	0.1687	0.0313	0.2048	-0.0152	0.5369

Table 23. P-values for each of the Five Factor Model traits testing whether any of the traits are statistically significantly different between the Non-English Primary Language dataset and the English Subset.

error rate to the expected values in the original dataset.

The results of the Mann-Whitney U-test are in table 22. The null hypothesis cannot be rejected for some of the psychological traits, such as Agreeableness and Neuroticism. Both of these values are far below a threshold of 0.05. Based on this fact, the use of translations to estimate certain psychological traits significantly changes either the sample population or the population distribution, indicating that translations should not be used when calculating some of the psychological traits, even when English posts are present.

Openness shows a very high p-value, likely indicating that translations have no impact on estimating the openness score of a subject. While the remaining two traits, Conscientiousness, and Extraversion, are low, but not small enough to reject the null hypothesis.

Primary Language

Some English social media posts in the data could impact the results obtained, potentially raising or lowering the p-values in the initial tests. Further experiments conducted on the dataset removed all English language posts from the translation dataset. Outcomes of this experiment are in table 23.

The results are consistent with the original test on the Multi-lingual dataset. The estimations for Agreeableness and Neuroticism remain statistically significantly impacted by the use of translations. In each of the remaining traits, the null hypothesis

cannot be rejected. Openness remains high, while the p-values for Conscientiousness and Extraversion are only marginally increased.

Per Language

Since the datasets are not normal distributions, it becomes easier to test smaller subsets. Smaller subsets can be tested to determine whether specific languages might have a statistically significant impact on both the Extraversion and Conscientiousness traits. For this portion of the testing, only languages with a minimum author count of at least 50 authors meet the criteria in the "No English" dataset.

Results for the analysis are in table 24. Eight languages met the criteria, and test results for each of these languages are found in the same table. Depending on the language, certain traits jump between being statistically significant and not. We speculate that the resultant translation for each of the languages might have a considerable impact on the feature sets, causing the significance test to give such drastically different results.

Spanish, French, and Italian had by far the most significant number of authors. The results show that Openness and Extraversion are statistically significantly different from those of the English subset with p-values near or below a 0.05 threshold. Agreeableness and Neuroticism both have low p-values, but in each, one of the languages appears to flip the language from being statistically significant. Conscientiousness was only statistically significant for the Spanish language. Based on this experiment, it appears that the translation's language has an impact on the ability to estimate psychological traits.

Per Linguistic Family

Most of the languages identified as an author's primary language were Indo-European languages. It is also worth looking at each of the linguistic families for each of the

languages. A clustering of those authors was completed based on linguistic families of their primary spoken language. For example, Indo-European includes Spanish, French, Italian, Portuguese, and Swedish. The Languages and their Linguistic families are in table 17.

The results of this portion of the experiment are in table 25. The analysis shows that the null hypothesis is not rejected for the Openness, Conscientiousness, and Extraversion traits (though it can be for Extraversion with Afro-Asiatic languages). The null hypothesis is rejected for Agreeableness and Neuroticism, except Austronesian languages and Neuroticism in Uralic languages.

Impact on Feature Distributions

The feature set for estimation is made up of 2,011 different features, as laid out in section III. Each feature is calculated for the "No English" and English datasets to determine the effect that translations have on features used by the algorithm. These features include the conditional probabilities of each of the 2,000 topics, along with the counts of negative sentiment, "hapax legomina", "dis legomina", and the Yule, Sichel, Brunet, and Honore measures of vocabulary richness. Each of these topics and the Natural Language Processing measures make up exponential distributions, so any testing on the impact that translations have on the feature sets must use the Mann-Whitney U-test.

The analysis of the effect on the feature distributions can be found in table A1. Only a handful of the tested features were acceptances of the null hypothesis, with many of the scores rejecting the null hypothesis having very low p-values, usually a smaller value than 10^{-100}, representing a very significant difference. In total, 141 of the features tested accepted the null hypothesis meaning that they were not statistically significantly different. The remaining 1,870 features rejected the null hypothesis. Based on the number of features that rejected the Null Hypothesis, it can be seen

that the use of translations significantly alters the usual topics in the text from the authors, as shown by the very low p-values. It also decreases the median scores for vocabulary richness in using all scoring measures - Yule's K-measure, Sichel's S-measure, Brunet's W-measure, and Honore's R-measure. This decrease can also be seen by observing the median values, which are significantly lower in the translation dataset.

Impact on Readability Measures

Along with vocabulary richness, further experiments examined how translations impact the readability measures of the text in English and foreign languages. While these features are not used in the algorithm to determine personality traits, they can help increase understanding of what impact translations might have on the text itself to determine the possibility of using translations for psycho-linguistic tasks in the future.

Table 26 shows six of the most common readability measures. In all cases, the translations significantly increase the difficulty in interpreting the text. The median score for English text requires only a 9th-grade reading level using the Flesch Kincade Grade Level metric. Using translations requires a college-level education (16th grade).

In nearly all cases, the p values are minimal (around 10^{-200}), indicating a significant increase in the readability measures' distributions. This result means that, beyond the 1,840 features that were statistically significantly altered, a significant difference is found in other common linguistic measurements. Therefore, translations are likely not very well suited for use when calculating psycho-linguistic traits, such as personality estimation.

4 Authorship Attribution

Limitations in the final BOLT Arabic and Chinese datasets pose some challenges in determining the effects of translations on the Authorship Attribution system. Similar to the Affect Analysis experiments, Monte Carlo Cross-Validation (MCCV) is utilized to determine the effects of translations on the algorithms' performance. Unlike the Affect Analysis experiments, the use of a behavioral biometric requires matching authors to one another. In this case, both the translations and the original text are utilized to determine the performance between them.

The experiment controls as many variables as possible to determine the metrics, relying solely on the algorithm's performance. Since the Keselj algorithm is meant to work on any language, it can be run on the Chinese, Arabic, and both translation datasets without any modifications. This feature of the algorithm allows the datasets to be a control, ensuring that the results are focused on the effects of translations rather than minor changes to the algorithm.

A random holdout of authors and documents is generated for each iteration. To control potential issues caused by random selection twice, the holdout sets and training sets contain the same authors and documents during each iteration between the translation and original text. Due to computational resources, only 100 authors are selected for testing. These authors remain the same between the translation and the original language datasets. The holdout set contains 10% of the authors (10 authors) and three documents per author, for a total of 30 documents in the holdout set. The remaining documents are used to train the Keselj algorithm during each iteration.

In our experiment, we calculate the number of iterations required using equation 7:

$$|R| = |A| \times \frac{|D_a|}{|D_a^h|} \times n \tag{7}$$

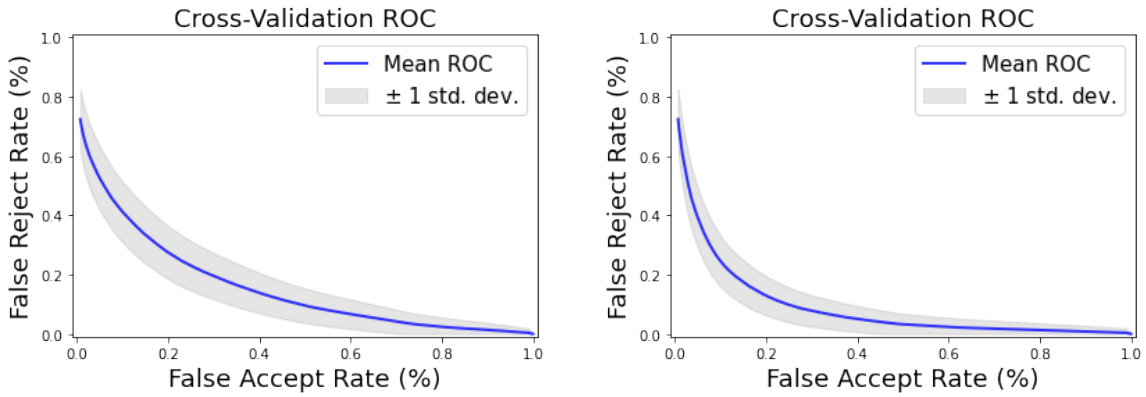
where R is the set of iterations, A is the author set, D_a is the Document set per author, D_a^h is the set of documents per author in the holdout set, and n is the number of times we wish to select each of the documents.

Since both datasets used a subset of 100 authors, the number of iterations is: $100 * (25/3) * 10 = 8000$. Eight thousand iterations are run for both the Chinese and the Arabic datasets on the Keselj algorithm. The number of iterations increases the likelihood of selecting each document 10 times, though there is no way to ensure that each document is randomly selected ten times. During each iteration, both the Original language and Translated text are run through the system, and performance metrics are calculated.

Arabic

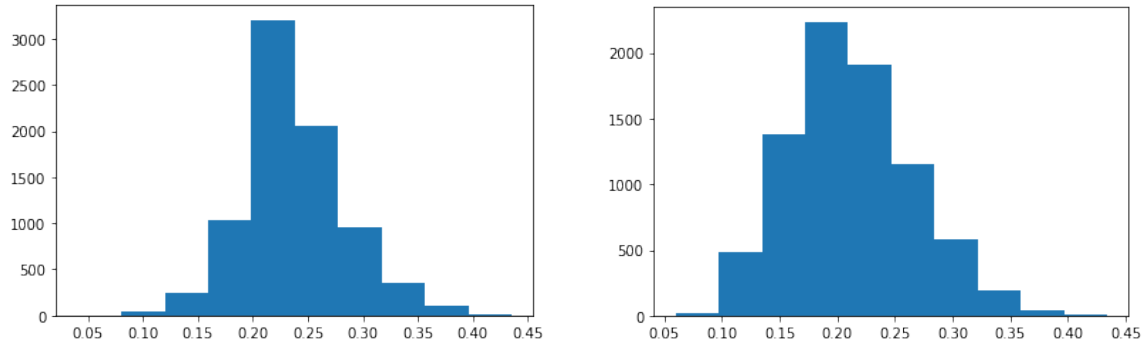
Running 8,000 iterations of the Arabic dataset against the Keselj algorithm provides a vector of performance metrics measuring the performance of both the translation and Arabic datasets. Performance metrics calculated during each iteration include the False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER). The vector of FAR and FRR values are used to calculate the mean FAR, mean FRR, the standard deviation of the FAR, and standard deviation of the FRR. These are then plotted to show the Mean Reciever Operator Curve (ROC), shown in Figure 26.

We see in this ROC that there is a noticeable difference between the two ROC curves. Since the FAR and the FRR are used in the ROC curve, better performance is obtained when the Area under the curve (AUC) is smaller - or the blue line (mean ROC) appears closer to the [0, 0] point on the charts. In this case, we can visually observe that the original text performs far better than the Translation data. This observation is confirmed by calculating the AUC for the mean ROCs: 0.1586 for the translated set, and 0.0890 for the original Arabic dataset. We see then that using the



(a) Translation **(b)** Arabic

Figure 26. The Mean ROC Curves with Standard Deviation for both the Translated (a) and Original (b) Arabic datasets.



(a) Translation **(b)** Arabic

Figure 27. The distribution of EER obtained from the Translated (a) and Original (b) Arabic datasets.

original text provides far better performance than using Translations of the text.

The EERs are also calculated for each iteration. The calculations provide a vector of EERs that can be used to obtain a distribution of the EERs for the system using the dataset. A histogram showing the distributions of both the translations and the Arabic dataset is shown in Figure 27. We see, based on this, that the distribution appears to be normal. This normal distribution is confirmed by a D'Agostino K^2 test for normality: the translation dataset has a p-value of 1.062×10^{-3}, and the Arabic dataset has a p-value of 3.06×10^{-13}. Both fall far below a p-value threshold of 0.05, indicating that the distributions are normal.

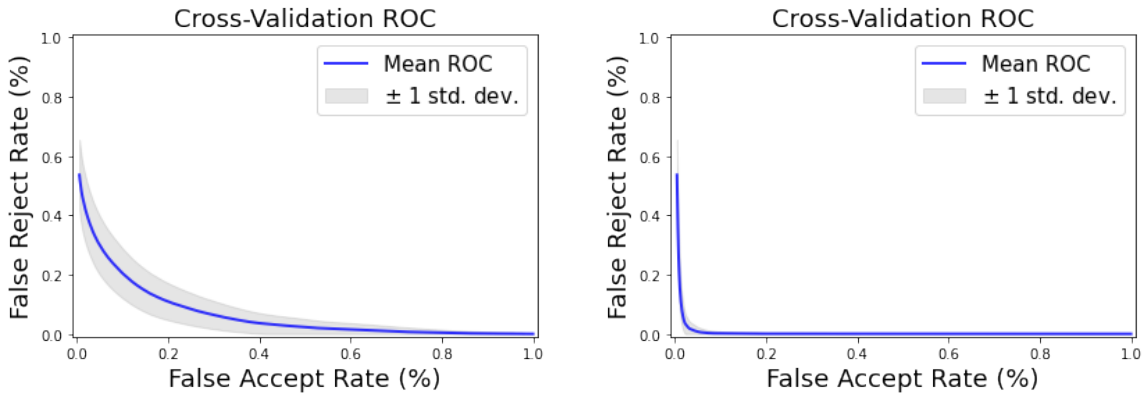
Knowing that the distributions are normal, we can calculate some basic statistics for each of the distributions. The Basic statistics for both distributions are shown in Table 27. We see an increase in the average EER between the translation and the Arabic datasets, while standard deviations and variance remain close between the two distributions. A Student's t-test is used to determine whether there is a statistically significant difference between the two distributions.

The t-test results in a t-value of 26.8131 with a p-value of 5.92×10^{-155}, resulting in a rejection of the null hypothesis. This rejection means that there is a statistically significant difference between the two distributions. Since the mean of the Translation dataset is larger than the Arabic dataset, this means that there is a statistically significant increase in the EERs when translations are used.

We can use the Cohen d measure to determine the effect size of the observed phenomenon. Cohen d measure results in a score of 0.424, a medium effect size between the distributions. Using Ruscio's Common Language conversion [194], there is a 61.8% chance that documents in the original language will provide a lower EER over the translated text. We also see statistically significant correlation coefficients. Pearson correlations result in a score of 0.6856 with a p-value close to 0 (floating-point rounding). A Spearman's ρ results in a score of 0.6726 with a p-value close to 0 (floating-point rounding). These p-values mean that there is a strong positive correlation between the two distributions and that the system monotonically increases. As a random variable selected in the Arabic dataset increases in error, the translation distribution on the same document will never decrease.

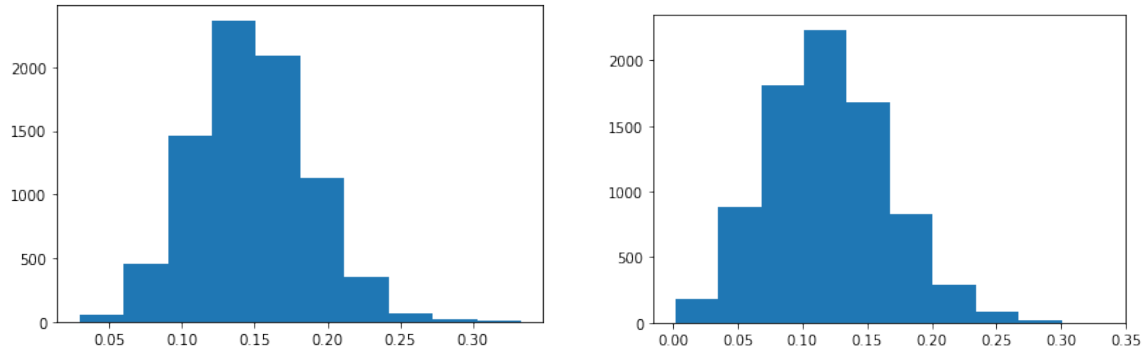
Chinese

The same experiments were run on the Chinese dataset, running 8,000 iterations against the Keselj algorithm. The same performance metrics were calculated during each iteration of the experiment. The mean ROC curves for the Chinese and



(a) Translation **(b)** Chinese

Figure 28. The Mean ROC Curves with Standard Deviation for both the Translated (a) and Original (b) Chinese datasets.



(a) Translation **(b)** Chinese

Figure 29. The distribution of EER obtained from the Translated (a) and Original (b) Chinese datasets.

Translated Chinese datasets are shown in Figure 28.

The differences in the ROC curves are more noticeably different between the two ROC curves. We again measure the FAR against the FRR, so better performance is obtained when the Area under the Curve (AUC) is smaller - the blue line (mean ROC) appears closer to the [0, 0] point in the charts. The performance is confirmed by measuring the AUC for the mean ROC for the translation, 0.0673, and the Original Chinese dataset, 0.0060. We see a significant performance increase when using Chinese logograms in the analysis. It is worth noting that the use of logograms appears to have an impact on the FAR and FRR. Further experimentation on Chinese char-

acters against the Keselj algorithm could help understand the impact of logograms on the system's overall performance.

The EERs are calculated for each iteration, providing a vector of EERs, allowing for the determination of the EER distribution. A histogram showing the distributions for both the translations and the Original Chinese dataset is shown in Figure 29. Both distributions appear to be normally distributed. This visual observation is confirmed by a D'Agostino K^2 test for normality: the translation dataset has a p-value of 3.73×10^{-15}, and the original Chinese dataset has a p-value of 2.08×10^{-24}. Both of these distributions are well below a threshold of 0.05, indicating that the distributions are normal.

Basic statistics are calculated for each of the distributions and are shown in Table 28. We see an increase in the average EER between the translation and the Chinese datasets. The variance and standard deviations remain close between the two distributions. A Student's t-test is used to determine whether there is a statistically significant difference between the two distributions.

The t-test results in a t-value of 26.3506, indicating a p-value of 7.83×10^{-150} and rejects the null hypothesis. The rejection of the null hypothesis means that there is a statistically significant difference between the two distributions. Since the mean of the Translation dataset is larger than the Chinese dataset, there is a statistically significant increase in the EERs when translations are used.

Cohen d is again used to measure the effect size of the observed phenomenon. The Cohen d measure results in a score of 0.4166, a medium effect size between the two distributions. Using Ruscio's Common Language conversion [194], there is a 61.4% chance that a document's original language will provide a lower EER over the translated text. Statistically significant correlation coefficients are also observed. A Pearson correlation coefficient results in a score of 0.6546 with a p-value close to 0 (floating-point rounding). A Spearman's ρ results in a score of 0.6481 with a p-

value close to 0 (floating-point rounding). These p-values mean a moderately strong, positive correlation between the two distributions exists, and that the system is monotonically increasing. As a random variable selected in the Chinese dataset increases in error, the translation distribution on the same document will never decrease.

From the analysis of both the Arabic and Chinese datasets, it is observed that the use of translations in this behavioral biometric statistically significantly increases the error rates. When using a behavioral biometric, the analyst should avoid using translations in preference for the original language of the documents.

Feature	Language	U-value	p-val.	Cohen D	Pearson r	Pearson p-val.	Spearman ρ	Spearman p-val.
Error O	Spanish	550616	0.0125	-0.1461	-0.0224	0.584	0.0546	0.1823
	French	97502	0.0162	0.4151	-0.0025	0.9792	0.0641	0.4998
	Italian	59862	0.0767	0.3567	-0.059	0.6325	0.0079	0.9493
	Portuguese	138741	0.2702	0.0044	0.0005	0.995	0.0289	0.7293
	Tigrinya	104222	0.2286	0.0642	-0.0448	0.6409	-0.0193	0.8404
	Swedish	108092	0.0899	0.0921	0.0296	0.7489	0.0618	0.5041
	Indonesian	76670	0.2403	0.1461	0.0796	0.477	-0.027	0.8097
	Dutch	42355	0.0505	0.3261	-0.1087	0.4526	-0.1712	0.2346
Error C	Spanish	557837	0.0372	0.0087	-0.0295	0.4721	-0.0276	0.5002
	French	107722	0.3129	-0.0018	-0.056	0.5559	0.0226	0.8121
	Italian	64446	0.322	-0.0368	-0.0358	0.7721	-0.0466	0.7061
	Portuguese	136854	0.1899	-0.1593	0.0326	0.6959	-0.0329	0.6935
	Tigrinya	107825	0.4381	-0.1073	-0.0569	0.5533	-0.0593	0.5367
	Swedish	113958	0.3378	-0.1375	-0.0952	0.303	-0.0754	0.4148
	Indonesian	75219	0.1629	0.0169	-0.1922	0.0837	-0.1821	0.1015
	Dutch	46168	0.2424	-0.2887	-0.0178	0.9024	-0.057	0.6942
Error E	Spanish	561005	0.0567	0.0601	0.0447	0.2751	0.0291	0.4769
	French	98705	0.0259	0.1539	-0.0575	0.5451	-0.0226	0.8119
	Italian	59582	0.0685	0.1731	0.2371	0.0515	0.2051	0.0934
	Portuguese	136103	0.1625	-0.0316	-0.0774	0.3533	0.0008	0.9926
	Tigrinya	90501	0.0014	0.2866	0.0647	0.5002	0.0604	0.529
	Swedish	100126	0.0047	-0.2981	0.0015	0.9873	-0.0214	0.8171
	Indonesian	78384	0.3528	0.0729	-0.0139	0.9013	-0.0838	0.454
	Dutch	48358	0.4371	0.1048	0.1476	0.3063	0.1048	0.4688
Error A	Spanish	568761	0.1372	-0.0232	-0.006	0.8828	0.0154	0.707
	French	100920	0.0562	0.3468	-0.0916	0.3348	-0.0713	0.453
	Italian	64237	0.3064	-0.0045	0.1833	0.1347	-0.2282	0.0612
	Portuguese	123581	0.003	0.314	-0.0514	0.5378	-0.0086	0.9182
	Tigrinya	98533	0.0473	0.312	0.0044	0.963	-0.0483	0.6149
	Swedish	114877	0.392	0.1317	0.007	0.9399	-0.0345	0.7092
	Indonesian	77516	0.2934	0.079	-0.1014	0.3645	-0.0835	0.4556
	Dutch	47730	0.377	0.0633	0.0245	0.8659	-0.0074	0.9591
Error N	Spanish	534660	0.0006	0.1668	0.0257	0.5306	-0.0217	0.5956
	French	104220	0.146	0.0214	-0.1368	0.1484	-0.1217	0.199
	Italian	64428	0.3207	-0.0718	-0.0029	0.9811	0.1633	0.1834
	Portuguese	134007	0.1003	-0.063	-0.1202	0.1483	-0.0305	0.7145
	Tigrinya	91306	0.0022	0.2276	-0.144	0.1315	-0.0227	0.8132
	Swedish	116098	0.4673	-0.1837	0.0618	0.5041	0.0419	0.6508
	Indonesian	77870	0.3171	0.0126	-0.1354	0.2253	-0.19	0.0873
	Dutch	45479	0.1925	-0.1223	0.0918	0.526	-0.0304	0.834

Table 24. Mann-Whitney U-Test results for each feature in each of the languages that contained at least 50 distinct authors having enough non-English text and meeting the minimum criteria of 1500 words.

Feature	Linguistic Family	U-value	p-val.	Cohen D	Pearson r	Pearson p-val.	Spearman ρ	Spearman p-val.
Error O	Indo-European	1.32e+06	0.4538	-0.0409	0.0425	0.1181	0.0497	0.0677
	Afro-Asiatic	104315	0.1883	0.0457	0.0356	0.7096	0.0438	0.6467
	Austronesian	76670	0.2403	0.1461	0.0796	0.477	-0.027	0.8097
	Uralic	67345	0.1089	0.3234	0.2068	0.0751	0.2201	0.0577
Error C	Indo-European	1.30e+06	0.1711	0.0258	0.0338	0.2139	0.0158	0.5605
	Afro-Asiatic	108977	0.4494	-0.1091	-0.029	0.7617	0.0014	0.9885
	Austronesian	75219	0.1629	0.0169	-0.1922	0.0837	-0.1821	0.1015
	Uralic	69854	0.2327	-0.1513	0.108	0.3562	-0.0043	0.9708
Error E	Indo-European	1.31e+06	0.3079	0.0252	-0.0005	0.9842	0.0302	0.2671
	Afro-Asiatic	90582	0.0009	0.2852	-0.0282	0.768	-0.0027	0.9772
	Austronesian	78384	0.3528	0.0729	-0.0139	0.9013	-0.0838	0.454
	Uralic	67784	0.1262	-0.1368	0.1629	0.1625	0.1471	0.208
Error A	Indo-European	1.28e+06	0.0307	0.0078	0.0236	0.3857	0.0279	0.3047
	Afro-Asiatic	100167	0.0597	0.3034	0.0884	0.3539	0.0825	0.3873
	Austronesian	77516	0.2934	0.079	-0.1014	0.3645	-0.0835	0.4556
	Uralic	62221	0.012	0.3423	-0.1182	0.3123	-0.0859	0.4636
Error N	Indo-European	1.24e+06	0.0006	0.1484	-0.0088	0.7474	-0.0265	0.3305
	Afro-Asiatic	92043	0.002	0.2238	-0.1682	0.0763	-0.0899	0.3456
	Austronesian	77870	0.3171	0.0126	-0.1354	0.2253	-0.19	0.0873
	Uralic	71125	0.3172	-0.1671	0.0493	0.6743	-0.0235	0.8412

Table 25. Mann-Whitney U-Test results for each feature in each of the Linguistic Families that contained at least 50 distinct authors having enough non-English text and meeting the minimum criteria of 1500 words.

Feature	EN median	FL median	KL divergence	U-value	p-value	H_0
Flesch Kincade Grade	9	16.8	-7858.93	1.16e+06	2.19e-234	Reject
Automated Readability Index	10.9	22.3	-9802.47	0.99e+06	2.47e-292	Reject
Gunning Fog	9.68	17.88	-10485.7	1.04e+06	8.47e-275	Reject
SMOG Index	9.9	12.7	-5593.67	1.13e+06	6.63e-242	Reject
Coleman Liau Index	7.54	9	-2197.25	1.30e+06	9.11e-195	Reject
Dale Chall Readability Score	6.3	7.5	-3058.69	1.15e+06	4.28e-239	Reject

Table 26. Tests on the readability measures used to calculate the estimated education level to understand the text. We also test whether the differences between the English dataset and the translation dataset are statistically significantly different.

Statistics	Translation	Arabic
Average	0.2382	0.2169
Variance	0.0024	0.0026
Standard Deviation	0.0494	0.0509
Minimum	0.0407	0.0598
Maximum	0.4355	0.4340

Table 27. Basic statistics calculated against the Equal Error Rate distributions for both the Translations and the Arabic BOLT datasets.

Statistics	Translation	Arabic
Average	0.1470	0.1282
Variance	0.0018	0.0022
Standard Deviation	0.0429	0.0471
Minimum	0.0295	0.0017
Maximum	0.3332	0.3335

Table 28. Basic statistics calculated against the Equal Error Rate distributions for both the Translations and the Chinese BOLT datasets.

CHAPTER V

CONCLUSIONS

While it may seem intuitive that one should not use translations for estimating traits using psycholinguistics, it has been the author's observation that many data scientists, social scientists, corporations, and other entities have often disregarded the potential effects of translation on their systems. This dissertation provides insight into the impact of translations on three different psychological and behavioral biometric measures, including Personality Trait estimation, Emotion Analysis, and Authorship Attribution. The experiments outlined in this dissertation also provide experimental designs that various scientists should use to test the impact of translations on their systems before relying on translated text to make judgments about psychological traits.

While this dissertation is by no means comprehensive for the entire field, it has shown that in at least three measures of psychological traits, translations statistically significantly increase the systems' error rates. Translations of text were never meant to maintain the emotional or behavioral state of the original authors. Instead, its purpose has been to convey the gist of a message to persons who cannot understand the original language. This dissertation does not intend to belittle Machine Translation research. Machine Translation still holds an essential place in protecting national security, improving business relations, and increasing globalization. This dissertation instead attempts to point out potential downfalls to other scientists who may rely

on translation systems to provide insights that Machine Translation systems are not currently built to handle.

1 Discussions and Speculations

The experiments conducted on Personality Trait estimation show an analysis of the impact of translations to determine if translations can be used in such systems. The results are sporadic, often depending on the specific language being translated. Therefore, it is impossible to say when the use of translations allows for personality traits to be appropriately estimated. Each of the traits rejects the null hypothesis when specific languages are used in the experiment; however, the analysis shows that one can more reliably say that Agreeableness and Neuroticism are the most likely traits to have increased error rates when using translations.

It does appear that the use of translations from other Indo-European languages to English provides the opportunity to use translations for estimating Openness, Conscientiousness, and Extraversion as the Null hypothesis cannot be rejected based on the p-values. However, looking at each language indicates that this may depend on the language that is being translated. For example, Spanish rejects the null hypothesis for Openness, Conscientiousness, and Neuroticism, while not rejecting Extraversion or Agreeableness. French rejects the Null hypothesis for Openness and Extraversion, while not rejecting the null hypothesis for Conscientiousness, Agreeableness, or Neuroticism. Portuguese only rejects the null hypothesis for Conscientiousness.

Based on this, it appears to be dependent on the language model that is used in translation. It can be speculated that this is a remnant of how well a language can be translated (or the BLEU scores) from one language to another. For example, if we assume a language that is a one-to-one translation that mimics English grammatical and lexical structures, which could achieve a BLEU score of 1, translations likely would not impact the resulting estimation of personality traits. However, such a

language does not currently exist, and BLEU scores rarely reach such scores. Such experiments may indicate the possibility of measuring the effect that a translation engine and translated language might have on the overall estimation of psychological traits, though further experimentation is necessary to prove such speculations.

The issues caused might also be caused by the effects of translations on the features that are being tested. The impact that translations have on the features tested in personality estimation is shown in Table A1. From this, we see that nearly all of the topics show a statistically significant difference between the translations and English language datasets. The fact that many of the features reject the null hypothesis indicates that translations have a detrimental effect on the topic distributions obtained from LDA. While the feature sets were not tested for each language, it can be speculated that each language has a unique impact on the topics discovered in each language as specific word choices from the translation engine might be preferred in their translations. Further experimentation around the Feature sets is necessary to discover how translations impact the system's topic model. The use of translations was also found to significantly impact the level of education required to comprehend the text, as indicated by various readability measures.

Emotion analysis experiments showed that in all cases for the four emotions tested, the emotions reject the null hypothesis with very low p-values. The low p-values indicate that the use of translations significantly increases, which can be observed between the histograms, which show a change in the distributions from normally distributed to non-normally distributed. We also find a lack of correlation between the translated and English language distributions. The lack of correlation indicates that translation appears the remove emotion words. In Machine Translation, the use of emotional adjectives likely adds little to the understanding of the message being conveyed and may get overlooked by the system.

Authorship attribution experiments resulted in the same effects. Since this was a

behavioral biometric, the resulting study explored different metrics, and required the exploration of the algorithms in both the original language and translated documents. We see that translation statistically significant effects all of the metrics used in our experiments, including False Accept Rate, False Reject Rate, and Equal Error Rates.

2 Impact

The exploration of the impact of translations on these three systems indicates that translations should be avoided or tested for statistical significance in all systems that attempt to estimate psychological traits and behavioral biometrics that rely on linguistic analysis. The impact of these findings has a minimal impact on the overall uses of psychological tasks that might be useful in adding context or determining potential malicious persons, as described in Chapter I. However, these findings show that each of these tasks should be done as close to the native language of the text as possible.

The performance of the Authorship attribution algorithms indicates that they are likely suitable for use in foreign languages. However, further experimentation should be done in understanding how effective the algorithms are when used on logogram based languages, such as hanzi (Chinese characters), instead of alphabetic languages (such as English and Arabic). The initial results in our exploration of the Keselj algorithm shows that such logogram based languages may be efficient.

However, research efforts should be increased to research and build natural language processing technologies that work in other languages. Much of the research efforts that have been found are primarily based on English language text. As globalization of economies increases and National Security threats expand worldwide for all nations, the importance of natural language processing technologies will be an essential aspect for identifying and contextualizing potential threats.

Identifying radical religiously motivated terrorists around the globe has given rise

to the need for Arabic language tools. Some tools were explored in Chapter IV, such as Stanford's CoreNLP, Columbia University's MADAMIRA, and University of Maryland's Arabic toolkit provides some attractive opportunities to explore Arabic for Emotion analysis. Initial exploration of Arabic emotion analysis showed that the tools have promise, but the preprocessing of Arabic using CoreNLP was not nearly as useful as the English language tools. Further exploration of MADAMIRA and the University of Maryland's Arabic toolkit may provide better results, but further exploration in Natural Language toolkits for twitter processing in various languages would prove beneficial. The use of native language toolkits would alleviate any need for translations to estimate psychological traits.

3 Future Research Efforts

This dissertation is meant to show that the effects of translations may be detrimental to the estimates that are being made by psychological tools. However, there is still much work that should go into understanding those effects to determine whether translations can work, when they fail, and how to conduct psychological trait estimates in native languages.

A significant difference is shown in the histograms for the Root Mean Squared Error for emotion analysis, but little was done to understand the underlying effect on the feature distribution. Further experimentation could provide insight into how translations affect the distributions of features between two sets of data. These effects could allow for a complete understanding of how translations impact the algorithms, allowing further development of the algorithms to consider translations, making them more useful.

Additionally, preprocessing methods, or exploration of other morphological and lexical toolkits (such as MADAMIRA and the University of Maryland Arabic Toolkit) might provide improved overall analytic capabilities when processing Native Language

text. While this would have no impact on the effect that translations have on the system, it would show the capabilities of the tools in Native Languages for additional comparisons.

Further experimentation on a linguistically homogeneous dataset needs to be conducted to determine whether the algorithm presented works in the native language. However, given the results of this experimentation, we strongly recommend building distinct models for individual languages to ensure that one can achieve the most accurate estimates possible in foreign languages. If translations must be used, scientists should conduct a full analysis of the impact their chosen translation system has on their estimates to determine which traits are more significantly affected.

A more comprehensive look at the languages that are being translated may provide insights into when translations fail, and whether traits can be more accurately detected in other languages. For example, rather than translating from Spanish to English, perhaps translations from Spanish to Portuguese may provide more accuracy. These types of studies, compared to the BLEU scores for translations from one language to another, could measure the potential impact that translations have on psychological estimates. However, this experimentation will require significant investments in building datasets across many different languages.

If there are specific languages that are less affected by translations, it would be worth exploring the translation of text to those languages and testing the systems using that language. Perhaps this could indicate that there is an intermediary or constructed language that could capture each language's nuances. By translating to this constructed language, perhaps using each of the systems on this constructed language could improve results across all languages.

Given the experimentation results in this dissertation, it is highly recommended to avoid the use of translations whenever the highest levels of accuracy are necessary, for instance, for National Security and Law Enforcement purposes. There is still

much potential experimentation needed to understand the impacts of translations further, but building distinct models for individual languages can ensure that analysts and scientists achieve the most accurate estimates possible in foreign languages. If translations must be used for any reason, scientists should conduct a full analysis of their chosen translation system's impact on their estimates to determine which traits are more significantly affected. This analysis will help ensure that analysts and scientists are better informed of the potential inaccuracies and change any resulting decisions from the data accordingly.

REFERENCES

[1] IBM, "Insider Threat Detection." [Online]. Available: https://www.ibm.com/security/security-intelligence/qradar/insider-threat

[2] ——, "IBM Watson: Personality Insights." [Online]. Available: https://www.ibm.com/watson/services/personality-insights/

[3] G. Skloot and D. D'Agostino, "Introducing Personality AI," Crystal Knows, Tech. Rep., 2020. [Online]. Available: https://www.crystalknows.com

[4] J. Feldman, "Artificial Intelligence in Cognitive Science," in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, pp. 792–796. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B0080430767016132

[5] K. D. Forbus, "AI and cognitive science: The past and next 30 years," *Topics in Cognitive Science*, vol. 2, no. 3, pp. 345–356, 2010.

[6] Raconteur, "A Day in Data," 2019. [Online]. Available: https://res.cloudinary.com/yumyoshojin/image/upload/v1/pdf/future-data-2019.pdf

[7] American Psychological Association, "2012 APA state licensing board list [Unpublished special analysis]," American Psychological Association, Washington, D.C., Tech. Rep., 2012.

[8] R. K. b. A. W. Khan, "Why Do We Need More Clinical Psychologists?" *The Malaysian Journal of Medical Sciences*, vol. 15, no. 2, pp. 1–2, 2008.

[9] A. W. Burgess, M. Mahoney, J. Visk, and L. Morgenbesser, "Cyber Child Sexual Exploitation," *Journal of Psychosocial Nursing and Mental Health Services*, vol. 46, no. 9, pp. 38–45, 9 2008. [Online]. Available: http://www.healio.com/doiresolver?doi=10.3928/02793695-20080901-01

[10] J. W. Patchin and S. Hinduja, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," *Youth Violence and Juvenile Justice*, vol. 4, no. 2, pp. 148–169, 2006.

[11] L. S.-L. Chen, H. H.-J. Tu, and E. S.-T. Wang, "Personality Traits and Life Satisfaction among Online Game Players," *CyberPsychology & Behavior*, vol. 11, no. 2, pp. 145–149, 4 2008. [Online]. Available: http://www.liebertpub.com/doi/10.1089/cpb.2007.0023

[12] K. S. YOUNG, "Internet Addiction: The Emergence of a New Clinical Disorder," *CyberPsychology & Behavior*, vol. 1, no. 3, pp. 237–244, 1 1998. [Online]. Available: http://www.liebertpub.com/doi/10.1089/cpb.1998.1.237

[13] A. Dhir, S. Chen, and M. Nieminen, "Predicting adolescent Internet addiction: The roles of demographics, technology accessibility, unwillingness to communicate and sought Internet gratifications," *Computers in Human Behavior*, vol. 51, pp. 24–33, 10 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0747563215003519

[14] C. Ko, J.-Y. Yen, C. Yen, C. Chen, C. Weng, and C. Chen, "The Association between Internet Addiction and Problematic Alcohol Use in Adolescents: The Problem Behavior Model," *CyberPsychology & Behavior*, vol. 11, no. 5, pp. 571–576, 10 2008. [Online]. Available: http://www.liebertpub.com/doi/10.1089/cpb.2007.0199

[15] S. Stieger, T. Eichinger, and B. Honeder, "Can Mate Choice Strategies Explain Sex Differences?" *Social Psychology*, vol. 40, no. 1, pp. 16–25, 1 2009. [Online]. Available: https://econtent.hogrefe.com/doi/10.1027/1864-9335.40.1.16

[16] E. von Zagorski, "Gender and modification of self-traits in online dating: The impact of anonymity, social desirability, and self-monitoring," Ph.D. dissertation, Walden University, 2013.

[17] J. T. Hancock, C. Toma, and N. Ellison, "The truth about lying in online dating profiles," *Conference on Human Factors in Computing Systems - Proceedings*, no. May 2014, pp. 449–452, 2014.

[18] D. Cappelli, A. Moore, R. Trzeciak, and T. J. Shimeall, "Common sense guide to prevention and detection of insider threats 3rd edition–version 3.1," Software Engineering Institute, Carnegie Mellon, Tech. Rep. January, 2009.

[19] US-CERT, "Combating the Insider Threat," *National Cybersecurity and Communications Integration Center*, no. May, pp. 61–64, 2014. [Online]. Available: https://www.us-cert.gov/sites/default/files/publications/CombatingtheInsiderThreat_0.pdf

[20] K. Le, M. Brent Donnellan, S. K. Spilman, O. P. Garcia, and R. Conger, "Workers behaving badly: Associations between adolescent reports of the Big Five and counterproductive work behaviors in adulthood," *Personality and Individual Differences*, vol. 61-62, pp. 7–12, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2013.12.016

[21] D. C. Blair, C. R. Barrett, C. W. Boustany Jr., S. Gorton, W. J. Lynn III, D. Wince-Smith, M. K. Young, and J. M. Huntsman Jr., "Update to the IP Commission Report on The Theft of American Intellectual Property: Reassessments of the challenge and United States Policy," U.S. Congress, Washington, D.C., Tech. Rep., 2017.

[22] B. Casey, "The Impact of Intellectual Property Theft on the Economy," US Senate, Washington, D.C., Tech. Rep. August, 2012.

[23] Cybersecurity Insiders, "Insider Threat: 2018 Report," Crowd Research Partners, Tech. Rep., 2018. [Online]. Available: https://cdn2.hubspot.net/hubfs/5260286/PDFs/Whitepapers/insider-threat-report-2018-wp.pdf

[24] A. Cummings, T. Lewellen, D. Mcintire, A. P. Moore, and R. Trzeciak, "Insider threat study: Illicit cyber activity involving fraud in the u. s. financial services sector," Software Engineering Institute, Carnegie Mellon, Tech. Rep. July, 2012. [Online]. Available: https://resources.sei.cmu.edu/asset_files/SpecialReport/2012_003_001_28137.pdf

[25] D. Cappelli, A. Moore, and R. Trzeciak, *The CERT Guide to Insider Threats: How to Prevent, Detect, and respond to Information Technology Crimes (Theft, Sabotage, Fraud).* Upper Saddle River, NJ: Addison-Wesley, 2012.

[26] A. P. Moore, D. M. Cappelli, T. C. Caron, E. Shaw, and R. F. Trzeciak, "Insider theft of intellectual property for business advantage: A preliminary model," *CEUR Workshop Proceedings*, vol. 469, pp. 1–21, 2009.

[27] J. Barling, K. E. Dupré, and E. K. Kelloway, "Predicting Workplace Aggression and Violence," *Annual Review of Psychology*, vol. 60, no. 1, pp. 671–692, 1 2009. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.psych.60.110707.163629

[28] M. Cullen and P. R. Sackett, "Peronality and counterproductive work behavior," in *Personality and Work*, M. Barrick and A. Ryan, Eds. San Franciscio, CA: Jossey-Bass, 2003, pp. 150–182.

[29] M. S. Hershcovis, N. Turner, J. Barling, K. A. Arnold, K. E. Dupré, M. Inness, M. M. LeBlanc, and N. Sivanathan, "Predicting workplace aggression: A meta-analysis." *Journal of Applied Psychology*, vol. 92, no. 1, pp. 228–238, 2007. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.92.1.228

[30] M. Mount, R. Ilies, and E. Johnson, "Relationship of Personality Traits and Counterproductive Work Behaviors: The Mediating Effects of Job Satisfaction," *Personnel Psychology*, vol. 59, no. 3, pp. 591–622, 9 2006. [Online]. Available: http://doi.wiley.com/10.1111/j.1744-6570.2006.00048.x

[31] J. F. Salgado, "The Big Five Personality Dimensions and Counterproductive Behaviors," *International Journal of Selection and Assessment*, vol. 10, no. 1&2, pp. 117–125, 3 2002. [Online]. Available: http://doi.wiley.com/10.1111/1468-2389.00198

[32] J. Anglim, F. Lievens, L. Everton, S. L. Grant, and A. Marty, "HEXACO personality predicts counterproductive work behavior and organizational

citizenship behavior in low-stakes and job applicant contexts," *Journal of Research in Personality*, vol. 77, pp. 11–20, 2018. [Online]. Available: https://doi.org/10.1016/j.jrp.2018.09.003

[33] P. R. Sackett and C. J. Devore, "Counterproductive Behaviors at Work," in *Handbook of Industrial, Work and Organizational Psychology: Personnel Psychology handbook of industrial, work and organizational psychology: Personnel psychology.* 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2001, pp. 145–164. [Online]. Available: http://sk.sagepub.com/reference/hdbk_orgpsych1/n9.xml

[34] P. E. Spector and S. Fox, "The Stressor-Emotion Model of Counterproductive Work Behavior." in *Counterproductive work behavior: Investigations of actors and targets.*, S. Fox and P. E. Spector, Eds. Washington: American Psychological Association, 2005, pp. 151–174. [Online]. Available: http://content.apa.org/books/10893-007

[35] I. N. A. M. F. Kozako, S. Z. Safin, and A. R. A. Rahim, "The Relationship of Big Five Personality Traits on Counterproductive Work Behaviour among Hotel Employees: An Exploratory Study," *Procedia Economics and Finance*, vol. 7, no. Icebr, pp. 181–187, 2013. [Online]. Available: http://dx.doi.org/10.1016/S2212-5671(13)00233-5

[36] C. M. Berry, D. S. Ones, and P. R. Sackett, "Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis." *Journal of Applied Psychology*, vol. 92, no. 2, pp. 410–424, 2007. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0021-9010.92.2.410

[37] L. M. R. Bolton, L. K. Becker, and L. K. Barber, "Big Five trait predictors of differential counterproductive work behavior dimensions," *Personality and Individual Differences*, vol. 49, no. 5, pp. 537–541, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2010.03.047

[38] J. M. Jensen and P. C. Patel, "Predicting counterproductive work behavior from the interaction of personality traits," *Personality and Individual Differences*, vol. 51, no. 4, pp. 466–471, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2011.04.016

[39] S. Basu, Y. H. Victoria Chua, M. Wah Lee, W. G. Lim, T. Maszczyk, Z. Guo, and J. Dauwels, "Towards a data-driven behavioral approach to prediction of insider-threat," *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 4994–5001, 2019.

[40] M. Maasberg, J. Warren, and N. L. Beebe, "The dark side of the insider: Detecting the insider threat through examination of dark triad personality traits," *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2015-March, pp. 3518–3526, 2015.

[41] S. M. Ho, M. Kaarst-Brown, and I. Benbasat, "Trustworthiness attribution: Inquiry into insider threat detection," *Journal of the Association for Information Science and Technology*, vol. 69, no. 2, pp. 271–280, 2018.

[42] F. L. Greitzer, J. D. Lee, J. Purl, and A. K. Zaidi, "Design and Implementation of a Comprehensive Insider Threat Ontology," *Procedia Computer Science*, vol. 153, pp. 361–369, 2019. [Online]. Available: https://doi.org/10.1016/j.procs.2019.05.090

[43] G. Yang, L. Cai, A. Yu, J. Ma, D. Meng, and Y. Wu, "Potential Malicious Insiders Detection Based on a Comprehensive Security Psychological Model," *Proceedings - IEEE 4th International Conference on Big Data Computing Service and Applications, BigDataService 2018*, no. 2017, pp. 9–16, 2018.

[44] National Intelligence Council (U.S.), "Global trends : paradox of progress," U.S. Office of the Director of National Intelligence, Tech. Rep., 2017. [Online]. Available: https://www.dni.gov/index.php/global-trends-home

[45] M. Crenshaw, "The psychology of terrorism: An agenda for the 21st century," *Political Psychology*, vol. 21, no. 2, pp. 405–420, 2000.

[46] A. Abbasi and H. Chen, "Affect Intensity Analysis of Dark Web Forums," in *2007 IEEE Intelligence and Security Informatics*. IEEE, 5 2007, pp. 282–288. [Online]. Available: http://ieeexplore.ieee.org/document/4258712/

[47] M. Petrovskiy and M. Chikunov, "Online extremism discovering through social network structure analysis," *2019 IEEE 2nd International Conference on Information and Computer Technologies, ICICT 2019*, pp. 243–249, 2019.

[48] I. B. Arpinar, U. Kursuncu, and D. Achilov, "Social media analytics to identify and counter islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online," *Proceedings - 2016 International Conference on Collaboration Technologies and Systems, CTS 2016*, pp. 611–612, 2016.

[49] K. S. Douglas, C. D. Webster, S. D. Hart, and H. Belfrage, *HCR-20V3: Assessing risk for violence - User guide*. Burnaby, Canaga: Mental Health, Law, and Policy Institute, Simon Fraser University, 2013.

[50] R. Lara-Cabrera, A. Gonzalez-Pardo, M. Barhamgi, and D. Camacho, "Extracting radicalisation behavioural patterns from social network data," *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, vol. 2017-Augus, pp. 6–10, 2017.

[51] R. Lara-Cabrera, A. Gonzalez Pardo, K. Benouaret, N. Faci, D. Benslimane, and D. Camacho, "Measuring the Radicalisation Risk in Social Networks," *IEEE Access*, vol. 5, pp. 10 892–10 900, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7935371/

[52] R. Lara-Cabrera, A. Gonzalez-Pardo, and D. Camacho, "Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter," *Future Generation Computer Systems*, vol. 93, pp. 971–978, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2017.10.046

[53] J. Torregrosa, I. Gilpérez-López, R. Lara-Cabrera, D. Garriga, and D. Camacho, "Can an automatic tool assess risk of radicalization online? A case study on Facebook," *Proceedings - 2017 European Intelligence and Security Informatics Conference, EISIC 2017*, vol. 2017-Janua, p. 165, 2017.

[54] I. Gilpérez-López, J. Torregrosa, M. Barhamgi, and D. Camacho, "An initial study on radicalization risk factors: Towards an assessment software tool," *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, vol. 2017-Augus, pp. 11–16, 2017.

[55] M. Obaidi, R. Bergh, N. Akrami, and J. F. Dovidio, "The Personality of Extremists: Examining Violent and Non-Violent Defense of Muslims," *PsyArxiv*, 2020. [Online]. Available: https://psyarxiv.com/kry38/

[56] J. Ginges, S. Atran, S. Sachdeva, and D. Medin, "Psychology out of the laboratory: The challenge of violent extremism." *American Psychologist*, vol. 66, no. 6, pp. 507–519, 2011. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0024715

[57] N. J. Smelser, *Theory of collective behavior*. New Orleans, LA: Quid Pro, 2011.

[58] A. Lankford, "Précis of The Myth of Martyrdom: What Really Drives Suicide Bombers, Rampage Shooters, and Other Self-Destructive Killers," *Behavioral and Brain Sciences*, vol. 37, no. 4, pp. 351–362, 8 2014. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0140525X13001581/type/journal_article

[59] J. Victoroff, "The Mind of the Terrorist," *Journal of Conflict Resolution*, vol. 49, no. 1, pp. 3–42, 2 2005. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0022002704272040

[60] M. D. Silber and A. Bhatt, "Radicalization in the west: The homegrown threat," The City of New York Police Department, NYPD Intelligence Division, New York, New York, USA, Tech. Rep., 2007. [Online]. Available: http://prtl-prd-web.nyc.gov/html/nypd/downloads/pdf/public_information/NYPD_Report-Radicalization_in_the_West.pdf

[61] M. Sageman, *Leaderless Jihad: Terror Networks in the Twenty-First Century*. Philadelphia, PA: University of Pennsylvania Press, 2008.

[62] G. Morf, *Le Terrorisme Quebecois*. Montreal, Canada: Editions de l'Homme, 1970.

[63] M. Crenshaw, "The Causes of Terrorism," *Comparative Politics*, vol. 13, no. 4, p. 379, 1981. [Online]. Available: https://www.jstor.org/stable/421717

[64] L. Leone, A. Chirumbolo, and M. Desimoni, "The impact of the HEXACO personality model in predicting socio-political attitudes: The moderating role of interest in politics," *Personality and Individual Differences*, vol. 52, no. 3, pp. 416–421, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2011.10.049

[65] M. Alizadeh, I. Weber, C. Cioffi-Revilla, S. Fortunato, and M. Macy, "Psychological and Personality Profiles of Political Extremists," *(Preprint)*, 4 2017. [Online]. Available: http://arxiv.org/abs/1704.00119

[66] E. Bell, M. A. Woodley, J. A. Schermer, and P. A. Vernon, "Politics and the General Factor of Personality," *Personality and Individual Differences*, vol. 53, no. 5, pp. 546–551, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2012.04.027

[67] S. Trip, M. I. Marian, A. Halmajan, M. I. Drugas, C. H. Bora, and G. Roseanu, "Irrational beliefs and personality traits as psychological mechanisms underlying the adolescents' extremist mind-set," *Frontiers in Psychology*, vol. 10, no. MAY, pp. 1–12, 2019.

[68] H. Chabrol, J. Bronchain, C. I. Morgades Bamba, and P. Raynal, "The Dark Tetrad and radicalization: personality profiles in young women," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 12, no. 2, pp. 157–168, 4 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/19434472.2019.1646301

[69] M. Dewing, "Social Media: An Introduction," Library of Parliament, Canada, Ottowa, Canada, Tech. Rep., 20120. [Online]. Available: https://bdp.parl.ca/staticfiles/PublicWebsite/Home/ResearchPublications/InBriefs/PDF/2010-03-e.pdf

[70] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[71] K. K. Kim, A. R. Lee, and U. K. Lee, "Impact of anonymity on roles of personal and group identities in online communities," *Information and Management*, vol. 56, no. 1, pp. 109–121, 2019. [Online]. Available: https://doi.org/10.1016/j.im.2018.07.005

[72] P. B. Lowry, J. Zhang, C. Wang, and M. Siponen, "Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model," *Information Systems Research*, vol. 27, no. 4, pp. 962–986, 2016.

[73] C.S.-W., "What doxxing is, and why it matters," 2014. [Online]. Available: https://www.economist.com/the-economist-explains/2014/03/10/what-doxxing-is-and-why-it-matters

[74] Federal Bureau of Investivation, "Don't Make the Call: The New Phenomenon of 'Swatting'," 2008. [Online]. Available: https://archives.fbi.gov/archives/news/stories/2008/february/swatting020408

[75] M. J. Moore, T. Nakano, A. Enomoto, and T. Suda, "Anonymity and roles associated with aggressive posts in an online forum," *Computers in Human Behavior*, vol. 28, no. 3, pp. 861–867, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2011.12.005

[76] D. Bradbury, "Unveiling the dark web," *Network Security*, vol. 2014, no. 4, pp. 14–17, 2014. [Online]. Available: http://dx.doi.org/10.1016/S1353-4858(14)70042-X

[77] J. S. Donath, "Identity Deception in the Virtual Community," *Communities in Cyberspace*, no. August 1996, pp. 29–59, 1999.

[78] N. Ellison, R. Heino, and J. Gibbs, "Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment," *Journal of Computer-Mediated Communication*, vol. 11, no. 2, pp. 415–441, 2006.

[79] P. D. Ekstrom and C. M. Federico, "Personality and political preferences over time: Evidence from a multiwave longitudinal study," *Journal of Personality*, vol. 87, no. 2, pp. 398–412, 2019.

[80] N. Satherley, C. G. Sibley, and D. Osborne, "Identity, ideology, and personality: Examining moderators of affective polarization in New Zealand," *Journal of Research in Personality*, vol. 87, p. 103961, 2020. [Online]. Available: https://doi.org/10.1016/j.jrp.2020.103961

[81] F. Mosteller and D. Wallace, "Notes on an authorship problem," in *Harvard Symposium on Digital Computers and their Applications*, 1962, pp. 163–197.

[82] ——, "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963.

[83] ——, *Inference and disputed authorship: The Federalist*. Center for the Study of Language and Inf., 1964.

[84] N. A. Novino, K. A. Sohn, and T. S. Chung, "A graph model based author attribution technique for single-class e-mail classification," *2015 IEEE/ACIS 14th International Conference on Computer and Information Science, ICIS 2015 - Proceedings*, pp. 191–196, 2015.

[85] L. Fridman, S. Weber, R. Greenstadt, and M. Kam, "Active Authentication on Mobile Devices via Stylometry, Application Usage, Web Browsing, and GPS Location," *IEEE Systems Journal*, vol. 11, no. 2, pp. 513–521, 2017.

[86] J. Herz and A. Bellaachia, "The Authorship of Audacity: Data Mining and Stylometric Analysis of Barack Obama Speeches," *Proceedings of the International Conference on Data Mining (DMIN)*, 2014.

[87] A. M. Kuruvilla and S. Varghese, "A Detection System to Counter Identity Deception in Social Media Applications," in *2015 International Conference on Circuit, Power, and Computer Technologies (ICCPCT)*. IEEE, 2015.

[88] R. Ragel, P. Herath, and U. Senanayake, "Authorship detection of SMS messages using unigrams," *2013 IEEE 8th International Conference on Industrial and Information Systems, ICIIS 2013 - Conference Proceedings*, pp. 387–392, 2013.

[89] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying Stylometry Techniques and Applications," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–36, 11 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3161158.3132039

[90] C. L. Huth, "The insider threat and employee privacy: An overview of recent case law," *Computer Law and Security Review*, vol. 29, no. 4, pp. 368–381, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.clsr.2013.05.014

[91] Legal Information Institute, "Electronic Surveillance," 2017. [Online]. Available: https://www.law.cornell.edu/wex/electronic_surveillance

[92] "Executive Order 12333," 1981. [Online]. Available: https://www.archives.gov/federal-register/codification/executive-order/12333.html

[93] H. Chen, "Sentiment and affect analysis of Dark Web forums: Measuring radicalization on the internet," in *2008 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 6 2008, pp. 104–109. [Online]. Available: http://ieeexplore.ieee.org/document/4565038/

[94] G. Grefenstette, Y. Qu, D. A. Evans, and J. G. Shanahan, "Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes," *AAAI Spring Symposium - Technical Report*, vol. SS-04-07, pp. 63–70, 2005.

[95] C. Ma, H. Prendinger, and M. Ishizuka, "Emotion estimation and reasoning based on affective textual interaction," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3784 LNCS, pp. 622–628, 2005.

[96] J. Donath, K. Karahalios, and F. Viegas, "Visualizing conversation," *Proceedings of the Hawaii International Conference on System Sciences*, p. 74, 1999.

[97] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 4, pp. 483–496, 2001.

[98] Z. J. Chuang and C.-h. Wu, "Multi-modal emotion recognition from speech and text," *Journal of Computational Linguistics and Chinese*, vol. 9, no. 2, pp. 45–62, 2004. [Online]. Available: http://www.aclweb.org/anthology/O/O04/O04-3004.pdf

[99] G. Mishne and M. De Rijke, "Capturing global mood levels using blog posts," *AAAI Spring Symposium - Technical Report*, vol. SS-06-03, no. August, pp. 145–152, 2006.

[100] G. Mishne, "Experiments with mood classification in blog posts," *Proceedings of ACM SIGIR 2005 workshop on stylistic ...*, 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.2693&rep=rep1&type=pdf

[101] K. S. Sabra, R. N. Zantout, M. A. El Abed, and L. Hamandi, "Sentiment Analysis: Arabic sentiment lexicons," *2017 Sensors Networks Smart and Emerging Technologies, SENSET 2017*, vol. 2017-Janua, pp. 1–4, 2017.

[102] A. Abbasi, Hsinchun Chen, S. Thoms, and Tianjun Fu, "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168–1180, 9 2008. [Online]. Available: http://ieeexplore.ieee.org/document/4479460/

[103] W. JAMES, "II.—WHAT IS AN EMOTION ?" *Mind*, vol. os-IX, no. 34, pp. 188–205, 1884. [Online]. Available: https://academic.oup.com/mind/article-lookup/doi/10.1093/mind/os-IX.34.188

[104] W. B. Cannon, "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory," *The American Journal of Psychology*, vol. 100, no. 3/4, p. 567, 1987. [Online]. Available: https://www.jstor.org/stable/1422695?origin=crossref

[105] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state." *Psychological Review*, vol. 69, no. 5, pp. 379–399, 1962. [Online]. Available: http://content.apa.org/journals/rev/69/5/379

[106] M. B. Arnold, *Emotion and personality*. New York, New York, USA: Columbia University Press, 1960.

[107] N. H. Frijda, *The emotions*. Cambridge: Cambridge University Press, 1986.

[108] R. S. Lazarus, *Psychological Stress and the Coping Process*. New York, New York, USA: McGraw-Hill, 1966.

[109] I. J. Roseman and C. A. Smith, "Appraisal Theory: Overview, Assumptions, Varieties, controversies," in *Appraisal processes in emotion: Theory, methods, research*, K. R. Scherer, A. Schorr, and T. Johnstone, Eds. London: London University Press, 2001, pp. 3–19.

[110] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*, R. Plutchik and H. Kellerman, Eds. New York, New York, USA: Academic Press, 1980, pp. 3–33.

[111] ——, "Emotions: a general psychoevolutionary theory," in *Approaches to Emotions*, K. R. Scherer and P. Ekman, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1984, pp. 197–219.

[112] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3, pp. 169–200, 1992. [Online]. Available: http://www.paulekman.com/wp-content/uploads/2009/02/Universality-Of-Emotional-Expression-A-personal-History.pdfhttp://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.99.3.550

[113] ——, "Basic Emotions," in *Handbook of Cognition and Emotions*, T. Dalgleish and M. J. Power, Eds. Chichester, UK: John Wiley & Sons, Ltd, 1999, ch. 3, pp. 45–60. [Online]. Available: https://www.paulekman.com/wp-content/uploads/2013/07/Basic-Emotions.pdfhttp://doi.wiley.com/10.1002/0470013494

[114] ——, "Universals and Cultural Differences in Facial Expressions of Emotion," in *Nebraska Symposium on Motivation*, J. Cole, Ed. Lincoln University of Nebraska Press, 1972, vol. 19, pp. 207–282. [Online]. Available: papers3://publication/uuid/FDC5E29A-0E28-4DDF-B1A4-F53FEE0B4F70

[115] ——, "Are There Basic Emotions?" pp. 550–553, 1992.

[116] Machine Elf 1735, "File:Plutchik-wheel.svg," 2011. [Online]. Available: https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg

[117] ChaoticBrain, "File:Plutchik Dyads.svg," 2019. [Online]. Available: https://commons.wikimedia.org/wiki/File:Plutchik_Dyads.svg

[118] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," *ACL 2007 - SemEval 2007 - Proceedings of the 4th International Workshop on Semantic Evaluations*, no. June, pp. 70–74, 2007.

[119] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 1–17, 2018.

[120] The MITRE Corporation, "Technical Report and Annotation Instructions: Language Processing for Virtual Communications: Egyptian Arabic," The MITRE Corporation, Tech. Rep., 2017.

[121] G. J. Boyle, "Myers-Briggs Type Indicator (MBTI): Some Psychometric Limitations," *Australian Psychologist*, vol. 30, no. 1, pp. 71–74, 3 1995. [Online]. Available: http://doi.wiley.com/10.1111/j.1742-9544.1995.tb01750.x

[122] S. R. WALLACE, W. V. CLARKE, and R. J. DRY, "The Activity Vector Analysis as a Selector of Life Insurance Salesmen," *Personnel Psychology*, vol. 9, no. 3, pp. 337–345, 9 1956. [Online]. Available: http://doi.wiley.com/10.1111/j.1744-6570.1956.tb01072.x

[123] W. M. Marston, *Emotions of Normal People*. London: Kegan Paul, Trench, Trubner & Co, 1928. [Online]. Available: https://archive.org/details/emotionsofnormal032195mbp/page/n7/mode/2up

[124] W. J. Camara, J. S. Nathan, and A. E. Puente, "Psychological test usage: Implications in professional psychology." *Professional Psychology: Research and Practice*, vol. 31, no. 2, pp. 141–154, 4 2000. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0735-7028.31.2.141

[125] B. C. Schiele, A. B. Baker, and S. R. Hathaway, "The Minnesota Multiphasic Peronality Inventory," *Journal-lancet*, no. 63, pp. 292–297, 1943.

[126] Y. S. Ben-Porath and D. L. Davis, *Case Studies for Interpreting the MMPI-A*. Minneapolis, MN: University of Minnesota Press, 1996.

[127] Y. S. Ben-Porath, *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press, 2012.

[128] J. N. Butcher and C. L. Williams, "Personality Assessment with the MMPI-2: Historical Roots, International Adaptations, and Current Challenges," *Applied Psychology: Health and Well-Being*, vol. 1, no. 1, pp. 105–135, 3 2009. [Online]. Available: http://doi.wiley.com/10.1111/j.1758-0854.2008.01007.x

[129] R. B. Cattell, *Use of Factor Analysis in Behavioral and Life Sciences*. New York, New York, USA: Plenum, 1978.

[130] E. C. Tupes and R. E. Christal, "Recurrent Personality Factors Based on Trait Ratings," *USAF ASD Technical Report*, vol. 60, no. 61-97, pp. 225–251, 1961. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-6494.1992.tb00973.x

[131] L. R. Goldberg, "From Ace to Zombie: Same explorations in the language of personality," in *Advances in Personality Assessment*, J. N. Butcher, Ed. Hillsdale, NJ: Erlbaum, 1982, pp. 201–234.

[132] R. R. McCrae and P. T. Costa, "Validation of the five-factor model of personality across instruments and observers." *Journal of Personality and Social Psychology*, vol. 52, no. 1, pp. 81–90, 1987. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.52.1.81

[133] B. S. Connelly, D. S. Ones, and O. S. Chernyshenko, "Introducing the Special Section on Openness to Experience: Review of Openness Taxonomies, Measurement, and Nomological Net," *Journal of Personality Assessment*, vol. 96, no. 1, pp. 1–16, 1 2014. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00223891.2013.830620

[134] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.

[135] R. Hogan, *Hogan Personality Inventory Manual*. Minneapolis, MN: National Computer Systems, 1986.

[136] D. J. Ozer and V. Benet-Martínez, "Personality and the Prediction of Consequential Outcomes," *Annual Review of Psychology*, vol. 57, no. 1, pp. 401–421, 1 2006. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.psych.57.102904.190127

[137] D. Watson and L. A. Clark, "Extraversion and Its Positive Emotional Core," in *Handbook of Personality Psychology*. Elsevier, 1997, pp. 767–793. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780121346454500305

[138] S. Rothmann and E. P. Coetzer, "The big five personality dimensions and job performance," *SA Journal of Industrial Psychology*, vol. 29, no. 1, 10 2003. [Online]. Available: http://sajip.co.za/index.php/sajip/article/view/88

[139] W. G. Graziano and N. Eisenberg, "Agreeableness: A Dimension of Personality," in *Handbook of Personality Psychology*. Elsevier, 1997, pp. 795–824. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780121346454500317

[140] A. B. Zonderman, S. V. Stone, and P. T. Costa, "Age and neuroticism as risk factors for the incidence of diagnoses of psychotic and neurotic disorders." in *Annual Convention of the American Psychological Association*, New Orleans, LA, 1989.

[141] M. C. Ashton, K. Lee, M. Perugini, P. Szarota, R. E. de Vries, L. Di Blas, K. Boies, and B. De Raad, "A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions From Psycholexical Studies in Seven Languages." *Journal of Personality and Social Psychology*, vol. 86, no. 2, pp. 356–366, 2004. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.86.2.356

[142] M. C. Ashton and K. Lee, "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure," *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 150–166, 2007.

[143] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical

considerations, and practical guidelines," *American Psychologist*, vol. 70, no. 6, pp. 543–556, 2015.

[144] N. Shuyo, "Language Detection Library for Java," 2010. [Online]. Available: https://github.com/shuyo/language-detection

[145] S. Day, J. Brown, Z. Thomas, I. Gregory, L. Bass, and G. Dozier, "Adversarial authorship, AuthorWebs, and entropy-based evolutionary clustering," *2016 25th International Conference on Computer Communications and Networks, ICCCN 2016*, no. August, 2016.

[146] C. Faust, G. Dozier, J. Xu, and M. C. King, "Adversarial authorship, interactive evolutionary hill-climbing, and author CAAT-III," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 11 2017, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/document/8285355/

[147] J. Gaston, M. Narayanan, G. Dozier, D. L. Cothran, C. Arms-Chavez, M. Rossi, M. C. King, and J. Xu, "Authorship Attribution via Evolutionary Hybridization of Sentiment Analysis, LIWC, and Topic Modeling Features," *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, no. September 2019, pp. 933–940, 2019.

[148] B. Onyshkevych, "Broad Operational Language Translation (BOLT) (Archived)." [Online]. Available: https://www.darpa.mil/program/broad-operational-language-translation

[149] J. Tracey, H. Lee, S. Strassel, and S. Ismael, "BOLT Arabic Discussion Forums LDC2018T10," Philadelphia, 2018. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2018T10

[150] J. Tracey, H. Lee, S. Strassel, and S. Chen, "BOLT Chinese Discussion Forums LDC2016T05," Philadelphia, 2016. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016T05

[151] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[152] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[153] L. Bloomfield, *Language*. London: George Allen & Unwin, 1935.

[154] J. Coleman, *Life of Slang*. Oxford: Oxford University Press, 2012.

[155] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[156] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 1094–1101, 2014.

[157] D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter, "Standard Arabic Morphological Analyzer (SAMA) Version 3.1," 2009. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010L01

[158] P. Rodrigues, V. Novak, C. Anton Rytting, J. Yelle, and J. Boutz, "Arabic data science toolkit: An API for Arabic language feature extraction," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 1239–1245, 2019.

[159] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[160] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 4, pp. 1036–1040, 2015.

[161] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical Predictors of Personality Type," in *2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[162] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[163] M. Wilson, "MRC psycholinguistic database: Machine-usable dictionary, version 2.00," *Behavior Research Methods, Instruments, & Computers*, vol. 20, no. 1, pp. 6–10, 1988.

[164] J. Oberlander and S. Nowson, "Whose thumb is it anyway? Classifying author personaltiy from weblog text," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, no. July, 2006, pp. 627–634.

[165] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large scale personality classification of personal weblogs," *ICWSM 2007 - International Conference on Weblogs and Social Media*, 2007.

[166] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*. New York, New York, USA: ACM Press, 2011, p. 253. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1979742.1979614

[167] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, pp. 149–156, 2011.

[168] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," in *IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing*, 2011, pp. 180–185.

[169] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," *AAAI Workshop - Technical Report*, vol. WS-13-01, pp. 2–5, 2013.

[170] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Modeling and User-Adapted Interaction*, vol. 26, no. 2-3, pp. 109–142, 2016.

[171] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. P. Seligman, "Automatic personality assessment through social media language." *Journal of Personality and Social Psychology*, vol. 108, no. 6, pp. 934–952, 6 2015. [Online]. Available: http://dx.doi.org/10.1037/pspp0000020%5Cnhttp://dx.doi.org/10.1037/pspp0000020.supphttp://doi.apa.org/getdoi.cfm?doi=10.1037/pspp0000020

[172] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, and L. H. Ungar, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS ONE*, vol. 8, no. 9, 2013.

[173] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[174] K. Sundararajan, T. Neal, Y. Yan, A. Fatima, Y. Xiang, K. Reese, and D. Woodard, "Performance and Feature Analysis for Large-Scale Authorship Attribution," *Journal for IC Research & Development (submitted)*, 2017.

[175] M. Potthast, S. Braun, T. Buz, F. Duffhauss, F. Friederich, J. M. Gulzow, J. Kohler, W. Lotzsch, F. Muller, M. E. Muller, R. Passmann, B. Reinke, L. Rettenmeier, T. Romestsch, T. Sommer, M. Trager, S. Wilhelm, B. Stein, E. Stamatatos, and M. Hagen, "Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello, Eds. Cham: Springer International Publishing, 2016, vol. 9626, pp. 393–407. [Online]. Available: http://link.springer.com/10.1007/978-3-319-30671-1

[176] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-Gram-based Author Profiles for Authorship Attribution," *Pacific Association for Computational Linguistics*, 2003.

[177] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 3 2011. [Online]. Available: http://link.springer.com/10.1007/s10579-009-9111-2https://link.aps.org/doi/10.1103/PhysRevLett.88.048702

[178] E. Stamatatos, "Author Identification Using Imbalanced and Limited Training Texts," in *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*. IEEE, 9 2007, pp. 237–241. [Online]. Available: http://ieeexplore.ieee.org/document/4312893/

[179] W. J. Teahan and D. J. Harper, "Using Compression-Based Language Models for Text Categorization," in *Language Modeling for Information Retrieval*. Dordrecht: Springer Netherlands, 2003, pp. 141–165. [Online]. Available: http://link.springer.com/10.1007/978-94-017-0171-6_7

[180] W. R. Bennett, *Scientific and engineering problem-solving with the computer.* Englewood Cliffs, New Jersey: Prentice-Hall, 1976.

[181] D. Benedetto, E. Caglioti, and V. Loreto, "Language Trees and Zipping," *Physical Review Letters*, vol. 88, no. 4, p. 048702, 1 2002. [Online]. Available: http://arxiv.org/abs/cond-mat/0108530http://dx.doi.org/10.1103/PhysRevLett.88.048702https://link.aps.org/doi/10.1103/PhysRevLett.88.048702

[182] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring Differentiability: Unmasking Pseudonymous Authors," *Journal of Machine Learning Research*, vol. 8, no. Jun, pp. 1261–1276, 2007. [Online]. Available: http://www.jmlr.org/papers/v8/koppel07a.html

[183] K. Sundararajan and D. L. Woodard, "What constitutes " style " in authorship attribution?" *27th International Conference on Computational Linguistics*, pp. 2814–2822, 2018.

[184] T. Neal, K. Sundararajan, and D. Woodard, "Exploiting Linguistic Style as a Cognitive Biometric for Continuous Verification," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2 2018, pp. 270–276. [Online]. Available: https://ieeexplore.ieee.org/document/8411232/

[185] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, 9 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[186] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating English to Arabic Machine Translation Using BLEU," in *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, 2013. [Online]. Available: www.ijacsa.thesai.org

[187] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040.pdf

[188] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989.

[189] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, New York, USA: Springer, 2011.

[190] P. Zhang, "Model Selection via Multifold Cross Validation," *Annals of Statistics*, vol. 21, no. 1, pp. 299–313, 1993. [Online]. Available: https://projecteuclid.org/download/pdf_1/euclid.aos/1176349027

[191] Q. S. Xu and Y. Z. Liang, "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[192] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 3 1947. [Online]. Available: http://projecteuclid.org/euclid.aoms/1177730491

[193] M. P. Fay and M. A. Proschan, "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules," *Statistics Surveys*, vol. 4, pp. 1–39, 2010. [Online]. Available: http://projecteuclid.org/euclid.ssu/1266847666

[194] J. Ruscio, "A probability-based measure of effect size: Robustness to base rates and other factors." *Psychological Methods*, vol. 13, no. 1, pp. 19–30, 2008. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.13.1.19

APPENDIX A: FEATURE SIGNIFICANCE TABLE

Table A1. Personality features were tested to determine if translations had a significant impact on the features used to estimate personality traits in the algorithm. In total, 141 of the features Accepted the null hypothesis (they were not statistically significantly different). The remaining 1,870 rejected the null hypothesis indicating that translations had an impact on those 1,870 features.

Feature	EN median	FL median	u-value	p-value	H_0
topic0	3.4723e-05	2.9037e-05	2.32522e+06	6.8712e-12	Reject
topic1	0.00017084	0.00014516	2.33841e+06	5.0089e-11	Reject
topic2	9.9062e-05	5.3927e-05	1.60226e+06	3.5321e-116	Reject
topic3	0.0001228	4.5225e-05	1.2822e+06	1.867e-197	Reject
topic4	0.0001521	8.7758e-05	1.83851e+06	9.701e-70	Reject
topic5	0.00017083	0.0001161	1.98804e+06	1.9628e-46	Reject
topic6	7.955e-05	2.3991e-05	1.58556e+06	4.1755e-120	Reject
topic7	0.00011667	3.7927e-05	1.33048e+06	1.7698e-184	Reject
topic8	0.00013431	9.4224e-05	2.20309e+06	1.3158e-20	Reject
topic9	0.00010734	5.0461e-05	1.88832e+06	9.4281e-61	Reject
topic10	4.2269e-05	2.583e-05	2.18252e+06	3.2475e-23	Reject
topic11	0.00013196	7.8082e-05	1.83702e+06	7.785e-70	Reject
topic12	2.2291e-05	9.005e-06	1.76914e+06	5.6476e-82	Reject
topic13	0.00012994	7.9005e-05	1.7571e+06	6.882e-84	Reject
topic14	0.0001007	0.00011062	2.58663e+06	0.18914	Accept
topic15	0.00015581	0.00011087	2.02953e+06	1.8441e-40	Reject
topic16	0.00011785	5.5671e-05	1.32806e+06	9.0062e-185	Reject
topic17	0.00010016	9.7456e-05	2.52681e+06	0.013353	Reject
topic18	0.00017933	0.00012495	1.90933e+06	3.6783e-58	Reject
topic19	0.000132	9.4173e-05	2.03634e+06	4.2564e-40	Reject
topic20	9.3253e-05	7.0657e-05	2.17258e+06	1.1269e-24	Reject
topic21	0.00011291	7.4583e-05	2.12176e+06	1.1088e-29	Reject
topic22	6.1965e-05	8.568e-06	993255	3.6174e-291	Reject
topic23	0.00013168	9.1414e-05	1.89857e+06	1.4562e-59	Reject
topic24	0.00011575	0.00010644	2.37961e+06	1.2539e-08	Reject
topic25	2.5295e-05	1.1833e-05	1.56563e+06	5.5042e-123	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic26	0.00013506	0.00012425	2.35307e+06	3.541e-10	Reject
topic27	1.5173e-05	4.3376e-05	1.81909e+06	1.5092e-71	Reject
topic28	8.5455e-05	7.5134e-05	2.42072e+06	2.8534e-06	Reject
topic29	1.1959e-05	2.3282e-05	2.06166e+06	7.6006e-36	Reject
topic30	0.00011504	4.8327e-05	1.51046e+06	6.5558e-137	Reject
topic31	6.3176e-05	5.085e-05	2.0597e+06	2.8118e-37	Reject
topic32	4.9245e-05	3.4181e-05	1.83492e+06	1.6214e-70	Reject
topic33	7.0588e-05	7.8216e-05	2.55991e+06	0.076794	Accept
topic34	0.00010055	7.2569e-05	2.19083e+06	1.3607e-22	Reject
topic35	0.00012157	8.316e-05	2.05699e+06	2.3063e-37	Reject
topic36	0.00014085	0.0001232	2.18818e+06	4.7848e-23	Reject
topic37	8.2977e-05	3.4252e-05	1.81734e+06	1.9978e-73	Reject
topic38	3.3733e-05	3.7172e-05	2.52716e+06	0.017495	Reject
topic39	0.00013867	9.0814e-05	2.03443e+06	1.0297e-39	Reject
topic40	0.00014826	0.00011978	2.18262e+06	3.4598e-23	Reject
topic41	1.6333e-05	4.9401e-05	1.59318e+06	2.0475e-107	Reject
topic42	0.0001557	0.00015573	2.58001e+06	0.14617	Accept
topic43	4.459e-05	3.5758e-05	2.29424e+06	3.8387e-14	Reject
topic44	0.00013897	0.00011702	2.21992e+06	4.1647e-20	Reject
topic45	4.3344e-05	3.0869e-05	2.03071e+06	5.9201e-41	Reject
topic46	0.00012187	7.7275e-05	2.03189e+06	1.1343e-40	Reject
topic47	4.4802e-05	7.9851e-06	1.76358e+06	7.6292e-83	Reject
topic48	0.00011235	9.5149e-05	2.36648e+06	6.1234e-09	Reject
topic49	5.7574e-05	9.7752e-05	2.09357e+06	4.2161e-33	Reject
topic50	0.00013808	0.00010388	2.16531e+06	2.6756e-25	Reject
topic51	9.4129e-05	2.1356e-05	956151	4.1023e-300	Reject
topic52	1.0375e-05	1.8286e-05	2.10604e+06	3.4693e-16	Reject
topic53	0.00013377	9.2553e-05	1.81816e+06	2.8782e-73	Reject
topic54	6.7346e-05	7.7496e-05	2.50224e+06	0.0026373	Reject
topic55	7.3768e-05	5.2348e-05	2.04387e+06	2.9123e-39	Reject
topic56	0.00014608	0.00010516	2.08850e+06	1.4565e-33	Reject
topic57	8.9475e-05	6.467e-05	2.15276e+06	2.9216e-26	Reject
topic58	2.475e-05	6.9883e-05	1.63202e+06	3.8613e-103	Reject
topic59	0.00012192	0.00014649	2.40530e+06	2.867e-07	Reject
topic60	0.00016811	0.0001895	2.36916e+06	3.205e-09	Reject
topic61	0.00017849	0.00011179	2.00585e+06	3.2849e-43	Reject
topic62	0.00016293	0.00018767	2.42398e+06	3.2355e-06	Reject
topic63	0.0001333	0.00010114	2.15325e+06	1.5668e-26	Reject
topic64	2.9494e-05	1.5315e-05	1.72692e+06	5.2988e-90	Reject
topic65	0.00010102	7.4709e-05	1.95592e+06	2.8376e-51	Reject
topic66	0.00015426	0.00012229	2.32757e+06	2.2627e-11	Reject
topic67	6.2202e-05	2.922e-05	2.05226e+06	5.791e-38	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic68	0.00018674	1.4795e-05	1.58884e+06	5.2741e-119	Reject
topic69	0.00021944	0.00021067	2.48408e+06	0.00058938	Reject
topic70	0.00013265	9.9467e-05	1.89826e+06	6.6957e-60	Reject
topic71	0.00010101	0.0001273	2.2164e+06	2.015e-20	Reject
topic72	4.4463e-05	1.5545e-05	1.22424e+06	7.2164e-215	Reject
topic73	0.00010664	9.1728e-05	2.3687e+06	3.9964e-09	Reject
topic74	6.0936e-05	6.2628e-05	2.54912e+06	0.040566	Reject
topic75	0.00014489	9.5295e-05	1.93678e+06	5.7943e-54	Reject
topic76	1.2605e-06	1.3695e-06	2.51975e+06	0.010852	Reject
topic77	0.00013411	8.4171e-05	1.69384e+06	2.0769e-96	Reject
topic78	5.5942e-05	4.0437e-05	1.8527e+06	1.7165e-67	Reject
topic79	0.00012624	8.5989e-05	2.00275e+06	1.4471e-42	Reject
topic80	0.00010953	6.6959e-05	1.80619e+06	4.3672e-75	Reject
topic81	0.00011271	0.00010578	2.51095e+06	0.0043776	Reject
topic82	0.0001564	0.00018839	2.27217e+06	9.8093e-16	Reject
topic83	2.9445e-05	1.5755e-05	2.0572e+06	1.0255e-36	Reject
topic84	0.0001301	0.00016004	2.21566e+06	1.3861e-20	Reject
topic85	0.00013555	0.00012589	2.45132e+06	3.5308e-05	Reject
topic86	0.00012728	9.2444e-05	1.98380e+06	9.7434e-47	Reject
topic87	6.8162e-05	6.4495e-05	2.50706e+06	0.0039483	Reject
topic88	0.00012525	7.2791e-05	1.71739e+06	2.1551e-91	Reject
topic89	0.0001062	1.1345e-05	775855	0	Reject
topic90	0.00015547	0.00019079	2.23373e+06	5.4945e-19	Reject
topic91	9.3534e-05	0.00012449	2.17169e+06	1.1653e-24	Reject
topic92	6.8368e-05	1.0705e-05	999699	7.4865e-285	Reject
topic93	3.9405e-05	2.5394e-05	2.13532e+06	2.0168e-28	Reject
topic94	7.7104e-05	4.7369e-05	1.55485e+06	9.0476e-127	Reject
topic95	9.192e-05	6.3074e-05	1.89564e+06	2.5747e-60	Reject
topic96	3.3086e-05	1.9049e-05	1.97527e+06	4.3754e-48	Reject
topic97	0.00014788	0.00012186	2.12434e+06	5.8477e-29	Reject
topic98	8.815e-05	7.9517e-05	2.45237e+06	4.8076e-05	Reject
topic99	0.00014017	0.00011367	2.05833e+06	1.9033e-37	Reject
topic100	5.9221e-05	4.0836e-05	2.15389e+06	2.2833e-26	Reject
topic101	0.00012433	8.9841e-05	2.15953e+06	3.639e-25	Reject
topic102	8.3935e-05	0.00032635	1.4122e+06	9.1558e-159	Reject
topic103	0.00014681	0.00010939	2.07448e+06	5.3729e-35	Reject
topic104	0.00013607	0.00017841	2.21048e+06	1.3658e-20	Reject
topic105	0.00010995	0.00010048	2.53939e+06	0.029698	Reject
topic106	1.8722e-05	1.4427e-05	2.19514e+06	2.1575e-22	Reject
topic107	8.7836e-05	6.7303e-05	2.19059e+06	1.0038e-22	Reject
topic108	7.323e-05	5.4748e-05	2.06124e+06	5.8151e-37	Reject
topic109	3.5925e-08	2.4595e-06	369056	6.079e-52	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic110	8.2209e-05	4.8885e-05	1.93505e+06	3.1803e-54	Reject
topic111	0.00012561	0.00011831	2.47754e+06	0.00038398	Reject
topic112	0.00016978	0.00015631	2.41262e+06	6.5975e-07	Reject
topic113	0.00012804	8.8228e-05	1.98597e+06	2.7465e-43	Reject
topic114	8.6921e-05	9.3108e-05	2.51975e+06	0.0071665	Reject
topic115	9.7732e-05	7.8165e-05	2.15093e+06	8.9865e-27	Reject
topic116	5.8407e-05	9.3373e-07	973495	1.3641e-289	Reject
topic117	0.00016956	0.00012287	2.11433e+06	1.3124e-30	Reject
topic118	7.8983e-05	5.6494e-05	2.01615e+06	1.7899e-42	Reject
topic119	5.5432e-05	3.9737e-05	2.22031e+06	6.9729e-20	Reject
topic120	0.00013976	0.00010139	1.8068e+06	4.1321e-75	Reject
topic121	0.0001491	0.00013732	2.43704e+06	8.833e-06	Reject
topic122	9.2863e-05	9.0642e-05	2.59167e+06	0.22898	Accept
topic123	0.00012471	7.6553e-05	1.97697e+06	1.967e-47	Reject
topic124	0.00019261	0.00015137	2.0645e+06	1.8905e-36	Reject
topic125	0.00014716	8.7657e-05	1.99081e+06	8.6068e-46	Reject
topic126	0.00013678	0.00014562	2.47163e+06	0.00021462	Reject
topic127	0.0001176	9.6031e-05	2.07462e+06	4.2243e-35	Reject
topic128	0.00014393	7.7164e-05	1.61514e+06	4.6664e-112	Reject
topic129	0.00036299	0.000224	1.43736e+06	6.3433e-156	Reject
topic130	0.00011232	0.00010191	2.55056e+06	0.043439	Reject
topic131	0.00012016	0.00010835	2.39326e+06	7.8809e-08	Reject
topic132	9.8222e-05	9.0431e-05	2.58864e+06	0.18759	Accept
topic133	5.1108e-05	7.0088e-06	979371	4.6639e-294	Reject
topic134	9.3321e-05	4.3711e-05	1.81434e+06	8.5985e-74	Reject
topic135	0.00014815	7.1495e-05	1.48171e+06	2.464e-142	Reject
topic136	0.00013264	9.1084e-05	2.15628e+06	5.1427e-26	Reject
topic137	5.8038e-05	2.7227e-05	2.02289e+06	7.5832e-42	Reject
topic138	0.00014145	0.00015831	2.39452e+06	8.0174e-08	Reject
topic139	0.00012936	7.7971e-05	1.63853e+06	4.2351e-108	Reject
topic140	8.3502e-05	5.3442e-05	1.95426e+06	5.6289e-51	Reject
topic141	4.3341e-05	6.7753e-05	1.81861e+06	2.3764e-73	Reject
topic142	0.00019794	0.0001697	2.30739e+06	3.4225e-13	Reject
topic143	0.00010873	7.158e-05	1.85360e+06	3.5046e-67	Reject
topic144	6.9961e-05	6.0761e-05	2.4693e+06	0.00019419	Reject
topic145	0.0002174	0.0003016	1.82776e+06	1.319e-71	Reject
topic146	0.00011261	6.7127e-05	1.97198e+06	8.0967e-49	Reject
topic147	6.3286e-05	1.6514e-06	894187	0	Reject
topic148	6.3383e-05	3.9049e-06	710068	0	Reject
topic149	7.4537e-05	3.7064e-05	1.65284e+06	1.5866e-104	Reject
topic150	9.7895e-05	6.7104e-05	2.2026e+06	1.3802e-21	Reject
topic151	3.0901e-05	6.3344e-06	1.22388e+06	2.2098e-212	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic152	4.8606e-05	1.9832e-05	1.61111e+06	3.2025e-114	Reject
topic153	0.00013576	8.7081e-05	2.08448e+06	1.8883e-33	Reject
topic154	0.00012415	0.00013075	2.53427e+06	0.020213	Reject
topic155	6.0564e-05	4.5116e-05	2.0047e+06	3.8663e-44	Reject
topic156	0.00014031	9.3187e-05	1.91171e+06	1.2151e-57	Reject
topic157	0.00015173	0.00012024	2.05281e+06	3.9115e-38	Reject
topic158	0.0001093	9.0851e-05	2.32681e+06	7.4101e-12	Reject
topic159	0.00016015	0.00010257	1.73203e+06	5.0335e-89	Reject
topic160	0.00012132	9.1075e-05	2.18317e+06	4.7344e-23	Reject
topic161	0.00012841	0.00011539	2.29881e+06	8.2834e-14	Reject
topic162	0.00013564	0.00012954	2.55785e+06	0.054847	Accept
topic163	0.00015866	7.4065e-05	1.4013e+06	1.5327e-164	Reject
topic164	0.00010746	9.6977e-05	2.40953e+06	5.2299e-07	Reject
topic165	7.5267e-05	4.7586e-05	2.17202e+06	1.2575e-24	Reject
topic166	4.9863e-05	5.2534e-05	2.62426e+06	0.47374	Accept
topic167	0.00010155	0.00011697	2.43567e+06	8.6208e-06	Reject
topic168	3.4169e-05	2.378e-05	2.08210e+06	3.3807e-34	Reject
topic169	0.00013573	7.4175e-05	1.70667e+06	1.1261e-93	Reject
topic170	7.3288e-05	7.3473e-05	2.56872e+06	0.09159	Accept
topic171	0.00012121	9.3157e-05	2.16849e+06	5.5895e-25	Reject
topic172	0.00011874	5.9432e-05	1.68156e+06	1.985e-99	Reject
topic173	0.00013834	0.00010469	1.98853e+06	2.2947e-46	Reject
topic174	0.00014038	0.00010804	2.25556e+06	8.4451e-17	Reject
topic175	0.00010318	6.2311e-05	1.98115e+06	2.1421e-47	Reject
topic176	0.0001212	9.1654e-05	1.99237e+06	5.8556e-46	Reject
topic177	0.00013347	0.00010511	2.18758e+06	6.6637e-23	Reject
topic178	7.9712e-05	6.5014e-05	2.26631e+06	2.7887e-16	Reject
topic179	9.4473e-05	0.0001215	2.22404e+06	7.803e-20	Reject
topic180	9.1151e-05	9.0206e-05	2.55134e+06	0.040516	Reject
topic181	7.3412e-05	6.0206e-05	2.22936e+06	3.4482e-19	Reject
topic182	0.00014185	9.2458e-05	2.05873e+06	2.133e-37	Reject
topic183	7.0501e-05	5.4531e-05	2.18662e+06	3.3147e-23	Reject
topic184	5.1584e-05	2.8584e-05	1.57893e+06	1.9643e-121	Reject
topic185	0.00010149	9.7866e-05	2.58226e+06	0.15745	Accept
topic186	0.00014194	0.00012134	2.23841e+06	1.7126e-18	Reject
topic187	0.0001294	8.8477e-05	2.1148e+06	8.7385e-31	Reject
topic188	4.3839e-05	2.9864e-05	1.93863e+06	1.0642e-53	Reject
topic189	6.8592e-05	4.882e-05	2.11768e+06	3.0263e-30	Reject
topic190	0.00012619	8.1922e-05	2.07048e+06	7.8264e-36	Reject
topic191	0.0001157	4.1367e-05	1.50548e+06	6.7731e-138	Reject
topic192	0.00021556	0.00019889	2.36898e+06	3.1316e-09	Reject
topic193	0.00012856	0.00010749	2.24107e+06	3.551e-18	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic194	0.00018155	0.00010738	1.88637e+06	8.6666e-61	Reject
topic195	6.4504e-05	3.7002e-05	1.8878e+06	3.1391e-60	Reject
topic196	6.0633e-05	2.491e-05	1.50313e+06	1.7414e-138	Reject
topic197	0.00012831	0.00011628	2.51409e+06	0.0094807	Reject
topic198	0.00010227	4.9035e-05	1.66054e+06	3.7972e-99	Reject
topic199	5.4469e-05	1.1358e-05	776602	0	Reject
topic200	9.0705e-05	1.0595e-05	1.02389e+06	1.7273e-278	Reject
topic201	0.00013342	0.00011365	2.22911e+06	2.1778e-19	Reject
topic202	3.6609e-05	4.1113e-05	2.51389e+06	0.0061778	Reject
topic203	3.7197e-05	3.5956e-05	2.57216e+06	0.12483	Accept
topic204	6.8096e-05	4.653e-05	2.12541e+06	1.6979e-29	Reject
topic205	0.00010018	5.6224e-05	1.81604e+06	1.7133e-73	Reject
topic206	9.5243e-05	5.272e-05	1.46872e+06	9.1255e-148	Reject
topic207	3.9606e-05	2.7502e-05	2.0861e+06	7.4236e-34	Reject
topic208	6.0614e-05	5.7341e-05	2.49745e+06	0.0016095	Reject
topic209	0.00013285	0.0001569	2.49855e+06	0.0018912	Reject
topic210	0.00014747	0.00014155	2.57226e+06	0.10527	Accept
topic211	8.9838e-05	8.1421e-05	2.50921e+06	0.010479	Reject
topic212	4.8421e-05	3.1802e-05	2.08159e+06	2.1725e-34	Reject
topic213	9.2774e-06	5.5262e-06	2.19377e+06	1.6315e-22	Reject
topic214	8.8465e-05	0.00020847	1.89671e+06	4.2122e-52	Reject
topic215	0.00015971	7.3597e-05	1.30106e+06	2.1466e-192	Reject
topic216	3.96e-05	3.4675e-05	2.44212e+06	2.769e-05	Reject
topic217	0.00011511	7.5568e-05	1.53966e+06	1.6962e-130	Reject
topic218	2.613e-05	4.922e-05	1.87735e+06	2.5491e-58	Reject
topic219	0.00017897	0.00018457	2.57512e+06	0.11224	Accept
topic220	0.00015962	0.00010637	1.95667e+06	9.6948e-51	Reject
topic221	4.0753e-05	2.2389e-05	1.59861e+06	1.9364e-116	Reject
topic222	0.00011339	9.0132e-05	2.14659e+06	3.0985e-27	Reject
topic223	7.284e-05	5.9766e-05	2.27532e+06	1.4225e-15	Reject
topic224	3.8741e-05	8.5645e-07	959916	3.5521e-302	Reject
topic225	0.00010842	0.00010369	2.62252e+06	0.44772	Accept
topic226	0.00013986	9.5426e-05	1.88178e+06	2.1307e-62	Reject
topic227	0.00011157	8.211e-05	2.29907e+06	8.653e-14	Reject
topic228	0.00010947	8.6152e-05	2.28471e+06	3.8661e-14	Reject
topic229	4.7196e-05	4.2719e-06	1.31501e+06	2.4541e-186	Reject
topic230	0.00014233	0.00012505	2.31768e+06	2.1215e-12	Reject
topic231	0.00012838	6.9432e-05	1.60573e+06	7.3572e-115	Reject
topic232	9.5982e-05	2.9754e-05	1.6984e+06	7.8778e-96	Reject
topic233	0.00017673	0.0001145	1.84072e+06	3.2229e-69	Reject
topic234	0.00012173	0.00012849	2.51423e+06	0.0054485	Reject
topic235	4.5425e-05	3.0777e-05	2.17443e+06	2.1743e-23	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic236	1.0911e-05	2.3793e-05	1.82685e+06	6.5999e-65	Reject
topic237	0.00013158	9.7533e-05	2.00317e+06	1.7624e-44	Reject
topic238	0.000198	0.00015126	2.06207e+06	5.5225e-37	Reject
topic239	7.6242e-05	6.5852e-05	2.39643e+06	1.1533e-07	Reject
topic240	9.5048e-05	9.9487e-05	2.52221e+06	0.008331	Reject
topic241	5.6724e-05	5.1384e-05	2.41981e+06	1.4586e-06	Reject
topic242	8.9502e-05	4.7419e-05	1.60095e+06	1.8137e-116	Reject
topic243	8.8398e-05	3.9025e-05	1.21527e+06	8.071e-218	Reject
topic244	0.00012174	0.00011233	2.44047e+06	1.9195e-05	Reject
topic245	0.00011092	7.4119e-05	2.01747e+06	1.4602e-42	Reject
topic246	0.0001021	0.00011233	2.40452e+06	2.6164e-07	Reject
topic247	0.00013101	7.7636e-05	1.72341e+06	1.0528e-90	Reject
topic248	0.00011277	0.00011312	2.61067e+06	0.35522	Accept
topic249	9.9454e-05	7.128e-05	2.04418e+06	1.0258e-38	Reject
topic250	0.00011881	9.2757e-05	2.24893e+06	1.0716e-17	Reject
topic251	8.5233e-05	7.2467e-05	2.13036e+06	5.88e-29	Reject
topic252	4.1287e-05	3.0559e-05	2.17247e+06	2.2456e-24	Reject
topic253	0.00013634	7.2357e-05	1.7404e+06	2.0794e-87	Reject
topic254	7.026e-05	3.43e-05	1.66414e+06	2.5376e-102	Reject
topic255	5.2416e-05	3.4648e-05	1.8551e+06	8.9447e-67	Reject
topic256	0.00018353	0.0001504	2.12935e+06	3.5278e-29	Reject
topic257	9.486e-05	8.3015e-05	2.30097e+06	1.4234e-13	Reject
topic258	0.00012739	0.00014934	2.39961e+06	2.1874e-07	Reject
topic259	5.9801e-05	3.7294e-05	1.66187e+06	3.6804e-103	Reject
topic260	8.3552e-05	8.3558e-05	2.57446e+06	0.12497	Accept
topic261	0.00020352	0.00017582	2.23969e+06	2.2e-18	Reject
topic262	0.00011035	7.2684e-05	1.62405e+06	2.1612e-111	Reject
topic263	0.00012562	8.1926e-05	1.98129e+06	7.7032e-47	Reject
topic264	0.00014818	7.4241e-05	1.62209e+06	1.1987e-111	Reject
topic265	2.5377e-05	1.3968e-05	1.86688e+06	4.3803e-64	Reject
topic266	2.5566e-05	2.4408e-05	2.46418e+06	0.00016681	Reject
topic267	9.8021e-06	1.4152e-06	1.5246e+06	4.0392e-133	Reject
topic268	0.00010433	7.0144e-05	2.08457e+06	4.8937e-34	Reject
topic269	0.00013114	7.8728e-05	1.67089e+06	4.3782e-101	Reject
topic270	0.00012256	0.00011279	2.41672e+06	1.1772e-06	Reject
topic271	0.00014389	7.5037e-05	2.00678e+06	7.1677e-44	Reject
topic272	5.1701e-06	1.1671e-06	1.87711e+06	1.6304e-61	Reject
topic273	0.0001332	0.00010873	2.37867e+06	1.9367e-08	Reject
topic274	0.00015183	0.00013584	2.43253e+06	6.9036e-06	Reject
topic275	0.00011692	9.0914e-05	2.16365e+06	1.8171e-25	Reject
topic276	0.00017296	0.00011773	1.78384e+06	2.9976e-79	Reject
topic277	0.0001602	0.00024516	1.81913e+06	4.2619e-73	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic278	0.0001475	9.114e-05	1.72708e+06	7.9131e-90	Reject
topic279	7.699e-05	5.0709e-05	2.22474e+06	2.0474e-19	Reject
topic280	4.5189e-05	2.9624e-05	2.13872e+06	1.1995e-25	Reject
topic281	8.8321e-05	5.2586e-05	2.1122e+06	9.6468e-31	Reject
topic282	0.00020098	0.00019042	2.45612e+06	5.5055e-05	Reject
topic283	9.1762e-05	9.5635e-05	2.59107e+06	0.21005	Accept
topic284	0.00010219	8.5671e-05	2.38866e+06	7.4919e-08	Reject
topic285	0.00020664	0.00034847	1.64262e+06	1.4078e-107	Reject
topic286	3.9795e-05	3.8791e-05	2.4935e+06	0.0019773	Reject
topic287	6.948e-05	5.1316e-05	1.98779e+06	9.9118e-47	Reject
topic288	8.7883e-05	5.8525e-05	1.37573e+06	1.3329e-171	Reject
topic289	0.00014128	0.00017361	2.26684e+06	3.673e-16	Reject
topic290	0.00014036	0.00011911	2.45136e+06	5.9395e-05	Reject
topic291	7.3326e-05	4.9962e-05	1.871e+06	5.3264e-64	Reject
topic292	0.00014705	0.00012607	2.23393e+06	5.7189e-19	Reject
topic293	0.00011795	0.00010541	2.41376e+06	9.6168e-07	Reject
topic294	0.00011245	7.7858e-05	1.70386e+06	9.8595e-95	Reject
topic295	7.9142e-05	6.6949e-05	2.39363e+06	8.2402e-08	Reject
topic296	0.00012942	0.00011668	2.42439e+06	3.8042e-06	Reject
topic297	6.4771e-05	3.8418e-05	1.73654e+06	3.6262e-88	Reject
topic298	3.4849e-05	1.6738e-05	1.85739e+06	4.1652e-66	Reject
topic299	7.7021e-05	7.8144e-05	2.61301e+06	0.37571	Accept
topic300	0.00017635	0.0001297	2.4127e+06	0.00085957	Reject
topic301	0.00016564	0.00012536	1.97279e+06	1.0582e-48	Reject
topic302	0.00011998	0.00015827	2.21255e+06	9.0367e-21	Reject
topic303	0.00014663	0.00015697	2.48636e+06	0.00070382	Reject
topic304	0.00012488	9.143e-05	2.16948e+06	7.0114e-25	Reject
topic305	0.00010374	0.00011318	2.51675e+06	0.0068285	Reject
topic306	9.2333e-05	8.9867e-05	2.5522e+06	0.044423	Reject
topic307	7.3953e-05	7.2055e-05	2.51852e+06	0.0070899	Reject
topic308	0.00011071	0.00010799	2.59273e+06	0.22089	Accept
topic309	9.9786e-05	0.00010654	2.52597e+06	0.014612	Reject
topic310	0.00014106	0.00011244	2.33105e+06	3.7529e-11	Reject
topic311	0.00014794	0.00014771	2.55278e+06	0.045853	Reject
topic312	8.5507e-05	2.2795e-05	973390	1.9594e-298	Reject
topic313	2.4752e-05	1.7764e-05	2.17467e+06	2.303e-24	Reject
topic314	0.00011944	5.8465e-05	1.58679e+06	6.8055e-119	Reject
topic315	5.156e-05	2.4496e-05	1.55255e+06	1.3551e-125	Reject
topic316	2.3654e-05	8.1678e-06	1.69521e+06	8.9359e-96	Reject
topic317	2.7732e-08	8.3569e-08	1.0416e+06	6.388e-43	Reject
topic318	0.00015376	0.00017872	2.31534e+06	1.2324e-12	Reject
topic319	0.00014556	8.2452e-05	1.79743e+06	5.9583e-77	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic320	0.00011198	9.0273e-05	2.25635e+06	6.5577e-17	Reject
topic321	0.00017205	0.00020118	2.39483e+06	1.076e-07	Reject
topic322	6.6753e-05	3.8273e-05	1.53191e+06	2.4995e-132	Reject
topic323	0.00015989	0.00011927	2.11342e+06	1.0389e-30	Reject
topic324	0.00011575	7.1518e-05	2.07974e+06	8.4377e-34	Reject
topic325	4.2565e-05	3.9351e-06	1.0561e+06	3.0117e-263	Reject
topic326	9.9814e-05	6.237e-05	2.0645e+06	1.8895e-36	Reject
topic327	0.0001193	3.0114e-05	1.21646e+06	6.1641e-216	Reject
topic328	0.00016103	0.00015671	2.57726e+06	0.13781	Accept
topic329	8.2718e-05	5.193e-05	2.27495e+06	1.9248e-15	Reject
topic330	9.0184e-05	3.1622e-05	1.45267e+06	1.3489e-151	Reject
topic331	0.00018446	0.00018262	2.56155e+06	0.067987	Accept
topic332	0.00012222	8.3534e-05	1.67371e+06	6.9992e-101	Reject
topic333	0.00014553	0.00013995	2.55883e+06	0.066626	Accept
topic334	0.00010207	7.4378e-05	1.93022e+06	1.1279e-54	Reject
topic335	0.00012163	8.7037e-05	1.92371e+06	2.2779e-55	Reject
topic336	9.2281e-05	2.9861e-05	1.28974e+06	1.7706e-192	Reject
topic337	0.00013097	8.5104e-05	2.04206e+06	5.1152e-39	Reject
topic338	4.7971e-05	1.6567e-05	1.8851e+06	2.6955e-61	Reject
topic339	8.3342e-05	8.1816e-05	2.5445e+06	0.0289	Reject
topic340	7.2582e-05	5.0618e-05	2.08759e+06	5.7277e-33	Reject
topic341	8.5107e-05	7.6869e-05	2.39216e+06	6.8974e-08	Reject
topic342	1.9642e-05	1.3789e-05	1.89586e+06	3.8257e-60	Reject
topic343	0.00014339	9.4621e-05	2.0025e+06	1.1635e-43	Reject
topic344	8.6949e-05	9.2586e-05	2.52197e+06	0.0087599	Reject
topic345	9.0413e-05	2.7531e-05	1.355e+06	1.2645e-176	Reject
topic346	8.1356e-05	6.1632e-05	1.83628e+06	2.7819e-70	Reject
topic347	0.00016311	7.4148e-05	1.51944e+06	8.9552e-133	Reject
topic348	0.00010441	6.6214e-05	1.94775e+06	3.4485e-52	Reject
topic349	0.0001258	7.153e-05	1.82312e+06	3.0863e-72	Reject
topic350	8.7593e-05	6.2101e-05	2.01374e+06	4.6567e-43	Reject
topic351	0.00011332	7.7796e-05	1.9655e+06	1.7757e-49	Reject
topic352	0.00020395	0.00018718	2.3902e+06	9.1246e-08	Reject
topic353	7.4849e-05	4.5193e-05	2.09641e+06	2.0324e-32	Reject
topic354	0.00012021	0.00012049	2.59256e+06	0.21201	Accept
topic355	0.00011997	6.6609e-05	1.60286e+06	2.6291e-115	Reject
topic356	6.9722e-05	5.8065e-05	2.32674e+06	1.6418e-11	Reject
topic357	0.00011431	0.00012887	2.43885e+06	1.0588e-05	Reject
topic358	0.00021259	0.00015786	1.81639e+06	9.664e-74	Reject
topic359	0.00010313	3.4122e-05	1.62072e+06	6.2919e-112	Reject
topic360	0.00016387	0.00013154	2.18565e+06	2.6709e-23	Reject
topic361	0.00012224	8.2631e-05	2.1454e+06	3.8348e-27	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic362	1.1386e-05	1.0268e-05	2.48564e+06	0.00066553	Reject
topic363	7.5499e-05	6.7753e-05	2.37527e+06	8.2792e-09	Reject
topic364	3.8688e-05	6.064e-05	2.11767e+06	6.0797e-29	Reject
topic365	9.6917e-05	6.9086e-05	2.06351e+06	1.1038e-36	Reject
topic366	2.929e-05	2.0754e-05	2.14329e+06	3.7116e-27	Reject
topic367	0.00010264	5.516e-05	1.82967e+06	4.2274e-71	Reject
topic368	4.4843e-05	1.3145e-05	1.46608e+06	4.8107e-148	Reject
topic369	9.9088e-05	5.6429e-05	1.95713e+06	7.6931e-51	Reject
topic370	0.00012629	0.00012322	2.59155e+06	0.19814	Accept
topic371	0.00010333	0.00011238	2.47055e+06	0.00021566	Reject
topic372	6.4555e-05	4.1585e-05	1.769e+06	5.3099e-82	Reject
topic373	0.00017048	0.00011976	2.16897e+06	2.0228e-24	Reject
topic374	0.00012822	0.00011256	2.30638e+06	3.4632e-13	Reject
topic375	0.00013535	0.00010017	1.9481e+06	2.0093e-52	Reject
topic376	0.00011148	7.783e-05	1.78816e+06	8.5185e-79	Reject
topic377	0.00012713	8.496e-05	2.06798e+06	1.1708e-35	Reject
topic378	0.00010562	5.7975e-05	1.91698e+06	2.0014e-56	Reject
topic379	0.00014129	0.00010649	2.06811e+06	7.0974e-36	Reject
topic380	6.2499e-05	3.3378e-05	1.81627e+06	1.2961e-73	Reject
topic381	0.00011165	5.3839e-05	1.55523e+06	6.6278e-126	Reject
topic382	0.00013523	0.0001946	1.97068e+06	5.2879e-49	Reject
topic383	3.2517e-06	9.0716e-06	84829	2.7468e-28	Reject
topic384	0.00015148	0.00011227	2.12991e+06	3.9752e-28	Reject
topic385	8.759e-05	3.1436e-05	1.19166e+06	1.1995e-224	Reject
topic386	0.00013747	8.0058e-05	1.89133e+06	3.0303e-60	Reject
topic387	0.00017728	0.00023118	2.1358e+06	1.7561e-28	Reject
topic388	7.8978e-05	6.6503e-05	2.36021e+06	1.7262e-09	Reject
topic389	7.1892e-05	3.0042e-05	1.1386e+06	4.5584e-242	Reject
topic390	7.7281e-05	5.1667e-05	1.93263e+06	1.8639e-54	Reject
topic391	0.00014469	9.7105e-05	2.01942e+06	4.7093e-42	Reject
topic392	0.00013901	0.00015955	2.35864e+06	1.1985e-09	Reject
topic393	0.00012587	9.5491e-05	2.19973e+06	1.3581e-21	Reject
topic394	3.3598e-05	6.9453e-05	1.83228e+06	3.2286e-65	Reject
topic395	0.00010412	7.2103e-05	2.12435e+06	2.1315e-29	Reject
topic396	7.6117e-05	9.5607e-05	2.43439e+06	7.5758e-06	Reject
topic397	0.00012741	7.132e-05	1.85329e+06	2.9485e-65	Reject
topic398	0.00011216	8.5407e-05	1.99522e+06	1.4447e-45	Reject
topic399	5.6838e-05	3.3047e-05	1.68706e+06	2.6326e-98	Reject
topic400	0.0001308	0.00010443	2.14798e+06	7.1263e-27	Reject
topic401	0.00013071	8.6943e-05	2.04264e+06	8.7554e-39	Reject
topic402	0.00012052	0.00011187	2.5242e+06	0.010072	Reject
topic403	5.2545e-05	5.3123e-05	2.56353e+06	0.073673	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic404	9.0033e-05	8.9209e-05	2.62472e+06	0.46627	Accept
topic405	0.00010055	3.979e-05	1.83091e+06	4.0806e-70	Reject
topic406	9.7999e-05	6.5547e-05	2.12674e+06	3.89e-29	Reject
topic407	7.0035e-05	6.5935e-05	2.51953e+06	0.007592	Reject
topic408	5.3263e-05	5.9762e-05	2.5029e+06	0.0032067	Reject
topic409	0.00013512	0.00011309	2.19828e+06	3.4652e-22	Reject
topic410	0.00014888	0.00012551	2.26126e+06	1.3172e-16	Reject
topic411	0.00011939	0.00013951	2.40407e+06	2.8286e-07	Reject
topic412	0.00016077	0.00013274	2.4006e+06	1.8891e-07	Reject
topic413	8.8121e-05	9.8216e-05	2.48975e+06	0.001076	Reject
topic414	9.3237e-05	6.1287e-05	1.99109e+06	9.685e-46	Reject
topic415	0.00013625	8.2025e-05	1.6387e+06	1.6529e-107	Reject
topic416	9.3967e-05	5.5222e-05	1.92805e+06	6.9829e-55	Reject
topic417	0.00018302	0.0002023	2.42675e+06	6.748e-06	Reject
topic418	8.9194e-05	9.8551e-05	2.48655e+06	0.00084542	Reject
topic419	0.00014145	0.00010013	2.1329e+06	1.1065e-28	Reject
topic420	0.00015164	0.00010656	1.96575e+06	1.2848e-48	Reject
topic421	0.00015062	0.00011223	2.38081e+06	2.5695e-08	Reject
topic422	0.00012136	8.3868e-05	1.94768e+06	2.356e-52	Reject
topic423	1.7567e-05	1.3396e-05	2.39826e+06	2.3743e-07	Reject
topic424	2.9043e-05	5.9085e-06	1.47361e+06	1.5635e-145	Reject
topic425	4.9174e-05	1.3322e-05	1.49456e+06	6.154e-140	Reject
topic426	8.5887e-05	7.9133e-05	2.49612e+06	0.0018755	Reject
topic427	8.6583e-05	4.2775e-05	1.3336e+06	8.1169e-184	Reject
topic428	0.00012442	8.8794e-05	1.94667e+06	1.722e-52	Reject
topic429	0.00014977	0.00015426	2.56563e+06	0.076907	Accept
topic430	0.00010926	0.00011759	2.54222e+06	0.027309	Reject
topic431	0.00013541	0.00013707	2.60467e+06	0.31692	Accept
topic432	8.8671e-05	0.00013338	1.99812e+06	3.6067e-45	Reject
topic433	0.00010623	7.4701e-05	1.95088e+06	9.9838e-52	Reject
topic434	8.1318e-05	5.5902e-06	753790	0	Reject
topic435	0.00015097	0.00012394	2.23224e+06	5.0539e-19	Reject
topic436	8.8553e-05	6.2008e-05	2.22025e+06	3.5876e-20	Reject
topic437	0.00014714	0.00010074	1.86556e+06	1.3896e-64	Reject
topic438	0.00010152	9.1629e-05	2.4269e+06	3.5044e-06	Reject
topic439	0.00011003	5.1409e-05	1.69746e+06	1.1164e-95	Reject
topic440	0.00012921	9.163e-05	2.14135e+06	8.8597e-28	Reject
topic441	0.00019171	0.0002079	2.53395e+06	0.017583	Reject
topic442	0.00013881	0.00013985	2.61189e+06	0.36623	Accept
topic443	9.3249e-05	5.844e-05	1.92653e+06	5.9119e-55	Reject
topic444	9.849e-05	6.4548e-05	1.95223e+06	3.0216e-51	Reject
topic445	0.00013002	0.0001679	2.08666e+06	5.1017e-34	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic446	0.00016145	0.00017927	2.44964e+06	3.3539e-05	Reject
topic447	0.00011246	0.00011032	2.59863e+06	0.24495	Accept
topic448	2.6405e-05	5.4087e-05	1.77831e+06	2.5173e-55	Reject
topic449	4.7243e-05	1.4523e-05	1.28623e+06	1.0747e-196	Reject
topic450	0.00010657	8.0983e-05	2.06405e+06	1.2539e-36	Reject
topic451	5.8551e-05	1.5933e-05	1.76114e+06	1.8796e-81	Reject
topic452	7.1444e-05	4.8285e-05	2.10018e+06	2.5698e-30	Reject
topic453	0.00014039	0.00011314	2.16296e+06	1.9751e-25	Reject
topic454	0.00016736	0.000158	2.48134e+06	0.00047461	Reject
topic455	7.7031e-05	0.00026477	1.41412e+06	4.1185e-144	Reject
topic456	0.0001055	2.5112e-05	1.07724e+06	5.6396e-259	Reject
topic457	8.6419e-05	3.5984e-05	1.27548e+06	7.9596e-200	Reject
topic458	0.00010251	9.2886e-05	2.46494e+06	0.00023574	Reject
topic459	0.00015062	8.4468e-05	1.89437e+06	2.2187e-60	Reject
topic460	6.9982e-05	6.0334e-05	2.45505e+06	8.2159e-05	Reject
topic461	0.00010568	4.5311e-05	1.76042e+06	1.3745e-83	Reject
topic462	0.00011429	9.1465e-05	2.02196e+06	4.2363e-42	Reject
topic463	0.00025627	0.00023667	2.40228e+06	2.2988e-07	Reject
topic464	6.4159e-05	2.8434e-05	1.19593e+06	3.2853e-224	Reject
topic465	2.9812e-05	2.6347e-05	2.28701e+06	1.342e-14	Reject
topic466	7.6624e-05	1.9105e-05	1.28629e+06	1.0648e-196	Reject
topic467	9.6082e-05	2.6683e-05	1.2634e+06	8.3055e-204	Reject
topic468	8.7756e-05	4.1479e-05	1.88123e+06	2.4658e-62	Reject
topic469	6.7515e-05	5.3091e-05	2.31829e+06	2.3358e-12	Reject
topic470	0.00013004	8.9264e-05	1.72985e+06	1.3199e-89	Reject
topic471	0.00012298	6.3417e-05	1.5556e+06	1.3508e-126	Reject
topic472	0.00015051	0.00013843	2.36696e+06	2.7378e-09	Reject
topic473	4.5262e-05	1.8307e-05	1.79555e+06	1.8659e-77	Reject
topic474	0.0001323	7.5423e-05	1.77806e+06	3.8311e-80	Reject
topic475	0.00013057	0.00013861	2.50788e+06	0.0038843	Reject
topic476	0.00010156	8.845e-05	2.41019e+06	6.4618e-07	Reject
topic477	0.00015755	0.00013848	2.26448e+06	2.4283e-16	Reject
topic478	8.2728e-05	9.106e-05	2.54007e+06	0.027429	Reject
topic479	2.3888e-05	2.6781e-05	2.45642e+06	6.217e-05	Reject
topic480	0.00015931	0.00016882	2.54535e+06	0.030181	Reject
topic481	8.9217e-05	5.0091e-05	1.90278e+06	4.8852e-59	Reject
topic482	0.00011918	6.9483e-05	2.02956e+06	1.8596e-40	Reject
topic483	0.00012108	0.00010247	2.31736e+06	1.6964e-12	Reject
topic484	9.1311e-05	4.1951e-05	2.08503e+06	1.7959e-31	Reject
topic485	3.9423e-05	1.1719e-05	1.52004e+06	3.6831e-135	Reject
topic486	0.0001013	7.1546e-05	1.9966e+06	2.9634e-45	Reject
topic487	6.7568e-05	3.0406e-05	1.72644e+06	1.0021e-87	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic488	9.4283e-05	0.00011347	2.31674e+06	1.5372e-12	Reject
topic489	0.00013982	0.00010221	2.05679e+06	1.1524e-36	Reject
topic490	3.5265e-06	1.9862e-06	1.85276e+06	9.8686e-59	Reject
topic491	6.4021e-05	6.7368e-05	2.56958e+06	0.090415	Accept
topic492	8.0017e-05	3.378e-05	1.58760e+06	4.3223e-119	Reject
topic493	0.00011899	5.7351e-05	1.84886e+06	2.8139e-64	Reject
topic494	0.00012311	0.00011022	2.45172e+06	4.487e-05	Reject
topic495	8.3791e-05	2.4202e-06	872329	0	Reject
topic496	0.00011211	6.6619e-05	2.07025e+06	5.5284e-36	Reject
topic497	0.00018398	0.00013585	1.88336e+06	1.9559e-62	Reject
topic498	5.6925e-05	6.7026e-05	2.388e+06	3.6097e-08	Reject
topic499	3.0212e-05	3.169e-07	1.9078e+06	1.409e-18	Reject
topic500	0.00013979	0.00010927	2.07894e+06	6.1744e-35	Reject
topic501	0.00012572	7.2447e-05	1.50259e+06	3.5572e-139	Reject
topic502	4.2346e-05	2.9749e-05	1.64861e+06	4.0611e-106	Reject
topic503	3.7455e-05	0.00010038	1.71239e+06	2.3125e-92	Reject
topic504	0.0001179	0.00012482	2.55434e+06	0.0493	Reject
topic505	0.00013493	8.0734e-05	1.76733e+06	8.5992e-81	Reject
topic506	7.0473e-05	4.7718e-05	2.1625e+06	3.5862e-25	Reject
topic507	9.3095e-05	6.1488e-05	2.25598e+06	5.0036e-17	Reject
topic508	0.00012328	0.00014815	2.23587e+06	1.0372e-18	Reject
topic509	0.00010689	6.1641e-05	1.89344e+06	9.1995e-59	Reject
topic510	0.00011392	0.00010118	2.50531e+06	0.0035136	Reject
topic511	0.00011183	7.9983e-05	1.81993e+06	4.0434e-73	Reject
topic512	0.00011387	7.4844e-05	1.8271e+06	1.5179e-71	Reject
topic513	1.8566e-06	9.1089e-06	1.9704e+06	1.106e-39	Reject
topic514	0.00014905	0.0001271	2.17128e+06	1.6752e-24	Reject
topic515	0.00010776	6.7159e-05	1.85533e+06	1.9504e-66	Reject
topic516	9.6078e-05	4.8246e-05	1.98843e+06	4.1539e-46	Reject
topic517	0.0001197	7.14e-05	1.67778e+06	7.6199e-100	Reject
topic518	0.00013585	9.1634e-05	1.65894e+06	1.4206e-103	Reject
topic519	7.2749e-05	5.424e-05	2.346e+06	2.8456e-10	Reject
topic520	6.716e-05	5.2316e-05	2.22775e+06	1.9959e-18	Reject
topic521	7.8941e-05	3.4116e-05	1.25375e+06	7.3094e-207	Reject
topic522	0.00013269	8.7374e-05	1.98928e+06	3.9829e-46	Reject
topic523	2.1721e-05	2.7273e-05	2.50802e+06	0.0052034	Reject
topic524	8.5683e-05	2.0479e-05	917671	0	Reject
topic525	0.00015583	0.00011676	2.11982e+06	1.4159e-29	Reject
topic526	0.00014136	0.0001097	1.98178e+06	1.9672e-47	Reject
topic527	6.3248e-05	3.451e-05	1.9642e+06	8.3546e-50	Reject
topic528	0.00011568	7.5093e-05	1.83231e+06	5.0932e-70	Reject
topic529	0.00010411	6.3003e-05	2.00624e+06	6.8105e-43	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic530	0.00011384	8.2878e-05	2.19821e+06	6.6204e-22	Reject
topic531	3.0013e-05	1.3139e-05	1.49222e+06	6.2163e-142	Reject
topic532	0.00010046	8.8338e-05	2.55595e+06	0.050294	Accept
topic533	0.00015902	0.00014451	2.37843e+06	1.0773e-08	Reject
topic534	9.8619e-05	6.1844e-05	1.86459e+06	3.4042e-65	Reject
topic535	0.00014912	9.2492e-05	2.04923e+06	1.0014e-37	Reject
topic536	0.00012147	0.00013078	2.46281e+06	0.00011092	Reject
topic537	0.00012255	9.5324e-05	2.24139e+06	3.7088e-18	Reject
topic538	6.8442e-05	4.7957e-05	2.16624e+06	1.6751e-24	Reject
topic539	0.00013597	9.4329e-05	2.36212e+06	1.4427e-09	Reject
topic540	0.00012099	0.00012531	2.55934e+06	0.068112	Accept
topic541	0.00013243	9.1487e-05	2.0166e+06	1.1179e-42	Reject
topic542	7.3195e-05	6.1675e-05	2.37699e+06	8.956e-09	Reject
topic543	0.00018998	0.0002096	2.37598e+06	7.8569e-09	Reject
topic544	0.00013988	7.4477e-05	1.75166e+06	3.0479e-85	Reject
topic545	5.803e-05	3.3752e-05	2.02952e+06	9.8677e-41	Reject
topic546	8.5974e-05	5.8733e-05	1.61507e+06	1.5394e-113	Reject
topic547	0.0001463	0.00011071	2.30506e+06	3.2836e-13	Reject
topic548	0.00013646	0.00012	2.17578e+06	2.3415e-24	Reject
topic549	9.0075e-05	8.1278e-05	2.42727e+06	3.2368e-06	Reject
topic550	0.00011528	6.5838e-05	1.92625e+06	7.2798e-55	Reject
topic551	0.00011142	0.00011016	2.53944e+06	0.035298	Reject
topic552	0.00012453	0.00013359	2.5649e+06	0.20269	Accept
topic553	0.00012978	0.00011071	2.24622e+06	7.8275e-18	Reject
topic554	0.00010564	5.9787e-05	1.78928e+06	2.0036e-78	Reject
topic555	0.00013448	0.00012758	2.48864e+06	0.00091446	Reject
topic556	4.8508e-05	4.6853e-07	742866	0	Reject
topic557	0.00011997	9.2367e-05	2.10919e+06	2.6698e-31	Reject
topic558	4.579e-05	4.4013e-05	2.53291e+06	0.015561	Reject
topic559	4.8428e-05	3.4545e-05	2.12528e+06	4.4306e-28	Reject
topic560	5.4707e-05	4.2481e-05	2.30103e+06	1.4386e-13	Reject
topic561	3.7014e-05	0.00011693	1.61314e+06	9.4317e-106	Reject
topic562	0.00013917	6.5197e-05	1.43455e+06	8.0796e-155	Reject
topic563	0.00013448	0.00010043	1.84628e+06	1.4169e-68	Reject
topic564	8.7262e-05	6.2819e-05	2.35879e+06	1.0662e-09	Reject
topic565	0.00022166	0.00024356	2.46102e+06	9.4741e-05	Reject
topic566	0.00013862	9.4987e-05	1.98844e+06	1.6745e-46	Reject
topic567	1.8502e-07	2.1673e-07	2.39127e+06	0.00017918	Reject
topic568	0.00013751	6.0578e-05	1.42362e+06	1.7497e-159	Reject
topic569	9.8167e-05	9.9942e-05	2.62322e+06	0.4434	Accept
topic570	0.00013954	0.00016222	2.35853e+06	7.5824e-10	Reject
topic571	4.954e-05	2.483e-05	1.84187e+06	1.565e-68	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic572	2.1092e-05	9.2274e-06	1.68513e+06	1.6125e-98	Reject
topic573	0.000131	0.00010727	2.13305e+06	8.8933e-29	Reject
topic574	0.00013575	0.00010384	2.16593e+06	3.9347e-25	Reject
topic575	0.00013598	6.7242e-05	1.65463e+06	7.5828e-105	Reject
topic576	2.3581e-05	7.3624e-06	1.56786e+06	1.0483e-121	Reject
topic577	0.00017258	0.00019603	2.42632e+06	2.9286e-06	Reject
topic578	5.946e-05	5.2792e-05	2.41364e+06	7.3915e-07	Reject
topic579	0.00016555	0.00016161	2.56364e+06	0.074336	Accept
topic580	0.00013316	0.00010413	2.08794e+06	9.4875e-34	Reject
topic581	0.00013039	5.0199e-05	1.06118e+06	4.2322e-266	Reject
topic582	7.0865e-06	7.5194e-06	2.5695e+06	0.15083	Accept
topic583	8.4373e-05	2.4162e-05	1.07281e+06	1.7278e-263	Reject
topic584	7.0808e-05	6.0416e-05	2.21161e+06	5.9505e-21	Reject
topic585	6.531e-05	7.2387e-05	2.44222e+06	1.6486e-05	Reject
topic586	9.82e-05	0.00011892	2.27333e+06	9.9692e-16	Reject
topic587	0.00014262	0.00010749	2.19594e+06	4.1328e-22	Reject
topic588	7.639e-05	3.173e-05	1.50423e+06	1.3332e-137	Reject
topic589	3.9799e-05	1.8925e-05	1.77864e+06	3.3427e-80	Reject
topic590	0.00013294	0.00013571	2.56254e+06	0.070951	Accept
topic591	0.00011572	0.00012886	2.54288e+06	0.037821	Reject
topic592	5.2591e-05	3.0174e-05	1.81427e+06	4.088e-74	Reject
topic593	4.6787e-05	3.3983e-05	1.63463e+06	4.1863e-109	Reject
topic594	4.5591e-05	1.8806e-05	1.54324e+06	1.0224e-128	Reject
topic595	0.00019833	0.00017939	2.308e+06	5.2963e-13	Reject
topic596	3.1828e-05	1.8224e-05	2.01513e+06	5.2604e-43	Reject
topic597	0.00010324	6.9691e-05	1.98071e+06	2.5526e-47	Reject
topic598	0.00012926	0.0001386	2.52256e+06	0.0090777	Reject
topic599	0.00012516	0.00015981	2.21172e+06	6.087e-21	Reject
topic600	0.00012259	5.5832e-05	1.64043e+06	9.0712e-107	Reject
topic601	4.4753e-05	3.1062e-05	2.09615e+06	1.4465e-32	Reject
topic602	7.1526e-05	5.5091e-05	2.03257e+06	1.0308e-40	Reject
topic603	0.00010822	6.4819e-05	1.86505e+06	5.7884e-65	Reject
topic604	0.00010539	9.8079e-05	2.52274e+06	0.0092298	Reject
topic605	0.0001299	0.00013542	2.58198e+06	0.14417	Accept
topic606	4.4407e-05	2.4197e-05	2.2609e+06	1.2523e-16	Reject
topic607	0.00012675	0.00012296	2.55624e+06	0.050977	Accept
topic608	7.0559e-05	9.7305e-05	2.27827e+06	2.4053e-15	Reject
topic609	4.1323e-05	2.3423e-05	1.87183e+06	5.2772e-64	Reject
topic610	9.9941e-05	6.8781e-05	2.04888e+06	1.6789e-38	Reject
topic611	5.6211e-05	4.2946e-07	820053	0	Reject
topic612	0.00017918	9.6603e-05	1.58806e+06	3.6725e-119	Reject
topic613	7.8513e-05	4.6181e-05	1.75771e+06	8.9682e-84	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic614	6.7295e-05	3.6763e-05	1.74159e+06	5.4361e-87	Reject
topic615	7.4699e-05	3.1677e-05	1.5067e+06	5.6025e-138	Reject
topic616	0.00019657	0.00018601	2.49432e+06	0.001385	Reject
topic617	8.5067e-08	8.0299e-08	2.13989e+06	0.024365	Reject
topic618	4.6597e-05	1.2679e-05	1.17544e+06	1.3445e-230	Reject
topic619	9.8512e-06	7.7095e-06	2.41708e+06	2.4888e-06	Reject
topic620	0.00012181	5.4699e-05	1.54576e+06	1.6895e-128	Reject
topic621	8.4809e-05	3.878e-05	2.1538e+06	2.2832e-26	Reject
topic622	0.00014406	0.00012951	2.459e+06	9.6144e-05	Reject
topic623	0.00012529	0.00010002	2.11922e+06	3.5352e-30	Reject
topic624	1.2197e-05	2.4365e-05	2.06517e+06	4.3415e-30	Reject
topic625	0.00011624	8.4464e-05	2.1594e+06	5.261e-26	Reject
topic626	0.00015959	0.00011056	2.07283e+06	1.0148e-34	Reject
topic627	9.647e-05	7.0272e-05	2.17688e+06	3.0021e-24	Reject
topic628	4.2137e-05	4.2021e-05	2.59289e+06	0.21349	Accept
topic629	4.1918e-05	3.4557e-05	2.40768e+06	4.8118e-07	Reject
topic630	0.00010419	9.5957e-05	2.48949e+06	0.0012432	Reject
topic631	0.00013817	8.5745e-05	1.81264e+06	2.7209e-73	Reject
topic632	0.0001319	0.00012365	2.6231e+06	0.49356	Accept
topic633	5.5737e-05	4.1725e-05	2.05452e+06	8.5256e-38	Reject
topic634	6.091e-05	5.6872e-05	2.48454e+06	0.00066799	Reject
topic635	0.00010244	5.0576e-05	1.7286e+06	3.4845e-89	Reject
topic636	0.00015256	0.00024484	1.74738e+06	2.1201e-86	Reject
topic637	5.8948e-05	2.6712e-05	1.52805e+06	7.1771e-133	Reject
topic638	0.00012912	0.00015096	2.2634e+06	2.9539e-16	Reject
topic639	0.00020363	0.00011617	2.1693e+06	6.7159e-24	Reject
topic640	9.8545e-05	6.3547e-05	1.94986e+06	9.5245e-52	Reject
topic641	0.00018582	0.00015897	2.28372e+06	6.2782e-15	Reject
topic642	0.00013114	8.1567e-05	1.87291e+06	5.5685e-64	Reject
topic643	0.00010747	7.4387e-05	2.20051e+06	2.1727e-21	Reject
topic644	0.00013355	0.00011479	2.04364e+06	2.721e-39	Reject
topic645	3.0016e-05	2.1617e-05	2.1105e+06	8.3002e-31	Reject
topic646	0.00012054	0.00012294	2.60947e+06	0.3651	Accept
topic647	4.4318e-05	2.3331e-05	1.92972e+06	3.399e-54	Reject
topic648	0.00010174	0.00014092	2.13037e+06	5.8933e-29	Reject
topic649	4.8324e-05	5.0336e-05	2.58174e+06	0.14901	Accept
topic650	0.00010449	9.7777e-05	2.49302e+06	0.0016194	Reject
topic651	0.00010051	6.3849e-05	1.82587e+06	8.9754e-72	Reject
topic652	1.029e-05	3.0683e-05	1.69098e+06	3.414e-79	Reject
topic653	1.1165e-05	1.8374e-05	2.21486e+06	9.3077e-20	Reject
topic654	9.1006e-05	6.1886e-05	2.05237e+06	3.4423e-38	Reject
topic655	0.00014061	0.00011083	2.05916e+06	2.4169e-37	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic656	5.6344e-05	5.2989e-05	2.61264e+06	0.3717	Accept
topic657	0.00012264	6.8488e-05	1.57107e+06	1.8642e-122	Reject
topic658	8.3359e-05	0.0001118	2.19257e+06	1.2552e-22	Reject
topic659	8.4936e-05	7.687e-05	2.37795e+06	1.1691e-08	Reject
topic660	4.1458e-05	1.3341e-05	1.13562e+06	1.2616e-242	Reject
topic661	5.8267e-05	5.079e-05	2.38767e+06	5.2389e-08	Reject
topic662	0.00014862	0.00017325	2.42847e+06	4.0875e-06	Reject
topic663	0.00011974	5.3404e-05	1.51956e+06	2.6344e-134	Reject
topic664	0.0001189	0.00010559	2.41063e+06	5.9838e-07	Reject
topic665	0.00015434	0.00011028	2.09968e+06	4.7103e-32	Reject
topic666	0.00013664	0.00011925	2.17429e+06	1.6662e-24	Reject
topic667	6.3759e-05	9.6473e-05	2.16863e+06	4.534e-25	Reject
topic668	0.00011286	8.0223e-05	1.93084e+06	5.1784e-54	Reject
topic669	3.6789e-05	3.7261e-05	2.61156e+06	0.35272	Accept
topic670	4.8288e-05	1.1531e-05	950899	1.221e-305	Reject
topic671	2.8889e-05	2.466e-05	2.58420e+06	0.16869	Accept
topic672	0.00012567	9.1878e-05	1.99224e+06	1.0496e-45	Reject
topic673	8.4792e-05	8.8186e-05	2.60205e+06	0.2873	Accept
topic674	3.8599e-05	2.4129e-05	1.59463e+06	4.5636e-118	Reject
topic675	0.0001226	8.6333e-05	1.93324e+06	3.2156e-54	Reject
topic676	0.00013557	0.00014492	2.58936e+06	0.2621	Accept
topic677	0.00013452	0.0001174	2.34097e+06	1.6232e-10	Reject
topic678	6.8857e-05	3.961e-05	1.83718e+06	5.7303e-70	Reject
topic679	5.9031e-05	3.9549e-05	1.69337e+06	5.0542e-97	Reject
topic680	0.00011868	6.2601e-05	1.42628e+06	2.1321e-158	Reject
topic681	7.8035e-06	3.7795e-07	2.06103e+06	5.4067e-20	Reject
topic682	0.00015909	0.00013492	2.24704e+06	1.3564e-17	Reject
topic683	0.00012688	9.7721e-05	2.29134e+06	6.8433e-14	Reject
topic684	0.00011501	0.00012811	2.42722e+06	3.6226e-06	Reject
topic685	0.00010707	3.5639e-05	1.68560e+06	2.2217e-97	Reject
topic686	1.5022e-05	2.7578e-05	1.97208e+06	2.0659e-48	Reject
topic687	9.847e-05	7.4221e-05	1.90067e+06	1.6132e-59	Reject
topic688	0.000127	5.6788e-05	1.70035e+06	1.6298e-69	Reject
topic689	0.00011078	6.2111e-05	1.74726e+06	4.4429e-86	Reject
topic690	0.00011579	9.7307e-05	2.30594e+06	2.7007e-13	Reject
topic691	7.7907e-05	9.8673e-05	2.28824e+06	1.3774e-14	Reject
topic692	0.00010958	8.9654e-05	2.28532e+06	1.0016e-14	Reject
topic693	2.6291e-05	1.4486e-05	2.00763e+06	7.0576e-44	Reject
topic694	8.907e-05	8.1376e-05	2.34316e+06	8.5541e-11	Reject
topic695	1.1251e-05	5.1224e-06	2.02373e+06	9.5238e-40	Reject
topic696	5.7032e-05	4.9359e-05	2.36994e+06	4.0661e-09	Reject
topic697	0.00016609	0.0001281	1.90178e+06	1.7048e-59	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic698	0.00011947	5.2011e-05	1.2966e+06	1.847e-193	Reject
topic699	7.4711e-05	3.6444e-05	2.16033e+06	1.4272e-24	Reject
topic700	6.5192e-05	4.5381e-05	2.09444e+06	5.4587e-33	Reject
topic701	4.1971e-05	2.5969e-05	1.64102e+06	1.441e-107	Reject
topic702	9.0406e-05	3.3012e-05	1.52684e+06	2.45e-133	Reject
topic703	0.00010865	0.00011072	2.56461e+06	0.081282	Accept
topic704	8.9491e-05	6.7939e-05	2.27003e+06	9.6767e-16	Reject
topic705	0.00013349	0.00011402	2.42465e+06	4.8925e-06	Reject
topic706	5.7095e-05	3.3787e-05	1.80241e+06	3.1948e-76	Reject
topic707	0.00011652	3.6838e-05	1.69711e+06	1.6004e-94	Reject
topic708	0.00013355	8.1853e-05	1.92394e+06	2.548e-55	Reject
topic709	0.00013165	0.00018988	1.869e+06	8.8828e-65	Reject
topic710	0.00014064	0.00010688	1.94583e+06	2.4961e-52	Reject
topic711	0.00011778	9.2553e-05	2.12271e+06	1.0875e-29	Reject
topic712	0.00015307	9.6461e-05	1.88936e+06	1.4231e-60	Reject
topic713	9.6113e-05	7.6523e-05	2.41368e+06	1.5903e-05	Reject
topic714	0.00019847	0.00019903	2.56043e+06	0.067857	Accept
topic715	0.00014264	0.00010063	2.06714e+06	4.0728e-36	Reject
topic716	8.8873e-05	8.0222e-05	2.45395e+06	6.1688e-05	Reject
topic717	0.00022123	0.00017517	1.91015e+06	3.4946e-58	Reject
topic718	8.8405e-05	4.6059e-05	1.41414e+06	9.3247e-162	Reject
topic719	3.8286e-05	2.5677e-05	2.06741e+06	4.2822e-36	Reject
topic720	8.9804e-05	5.1576e-05	1.46341e+06	6.3618e-148	Reject
topic721	0.00011993	9.4582e-05	2.1613e+06	1.0483e-25	Reject
topic722	0.00012858	8.9867e-05	1.97678e+06	5.3415e-48	Reject
topic723	0.00018479	0.00019916	2.48184e+06	0.00049419	Reject
topic724	6.9616e-05	8.2506e-06	971226	4.6142e-298	Reject
topic725	0.00011672	8.5865e-05	2.03452e+06	3.2454e-40	Reject
topic726	0.00012661	0.00010774	2.13759e+06	2.7355e-28	Reject
topic727	9.0029e-05	4.6059e-05	1.40525e+06	2.8792e-163	Reject
topic728	0.00013305	0.0001019	2.17142e+06	1.0954e-24	Reject
topic729	0.00017089	0.00013537	2.27331e+06	1.2055e-15	Reject
topic730	0.00010274	6.7266e-05	1.88754e+06	1.8031e-61	Reject
topic731	0.00014009	0.00011724	2.34269e+06	9.3832e-11	Reject
topic732	0.00012604	0.00013309	2.51944e+06	0.007028	Reject
topic733	0.00013009	0.00011436	2.40886e+06	4.8471e-07	Reject
topic734	9.7007e-05	3.8698e-05	1.34264e+06	4.4332e-181	Reject
topic735	5.0789e-05	3.4788e-05	2.15774e+06	1.8845e-25	Reject
topic736	9.4611e-05	6.2569e-05	2.22784e+06	2.5421e-19	Reject
topic737	0.0001122	8.0861e-05	2.13918e+06	6.7073e-28	Reject
topic738	0.00015159	0.00011638	2.13199e+06	1.4389e-28	Reject
topic739	0.00012461	8.6054e-05	2.04019e+06	4.1801e-39	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic740	0.00012209	4.9325e-05	1.65404e+06	1.1619e-102	Reject
topic741	4.3071e-05	8.2321e-06	976033	6.0022e-294	Reject
topic742	0.00010366	6.312e-05	2.0093e+06	3.8072e-43	Reject
topic743	0.00013925	9.9173e-05	2.06681e+06	1.8689e-35	Reject
topic744	1.9621e-05	2.2911e-05	2.41834e+06	1.7944e-06	Reject
topic745	0.00010325	8.4023e-05	2.3932e+06	7.7279e-08	Reject
topic746	0.00018037	0.000191	2.49785e+06	0.0016563	Reject
topic747	0.00016417	0.00016074	2.61477e+06	0.41901	Accept
topic748	0.00013253	0.00011342	2.23413e+06	5.9558e-19	Reject
topic749	2.7091e-07	2.8766e-07	2.51157e+06	0.076391	Accept
topic750	0.00011258	7.9348e-05	1.82424e+06	4.8324e-72	Reject
topic751	0.00013913	7.9256e-05	1.57009e+06	1.2653e-123	Reject
topic752	9.6819e-05	9.6168e-05	2.5747e+06	0.11047	Accept
topic753	0.0001253	7.8742e-05	2.07274e+06	2.4512e-35	Reject
topic754	6.2987e-06	5.8262e-06	2.36292e+06	0.0081378	Reject
topic755	9.7051e-05	6.2277e-05	2.01142e+06	4.1943e-43	Reject
topic756	4.6907e-05	3.5452e-05	2.19235e+06	8.9816e-22	Reject
topic757	2.186e-05	7.6512e-05	1.7471e+06	3.5436e-79	Reject
topic758	0.00010542	8.0767e-05	1.98277e+06	1.9766e-47	Reject
topic759	3.0785e-05	2.1423e-05	2.22305e+06	3.307e-18	Reject
topic760	0.0001181	0.00012007	2.61445e+06	0.37694	Accept
topic761	9.4506e-05	5.2857e-05	1.87041e+06	4.801e-63	Reject
topic762	3.5079e-05	2.6288e-05	2.19097e+06	7.0135e-23	Reject
topic763	0.00015539	0.00012781	2.09224e+06	2.3038e-33	Reject
topic764	0.00010969	7.9294e-05	2.13276e+06	1.7838e-28	Reject
topic765	5.1866e-05	1.8473e-05	1.5804e+06	8.3512e-119	Reject
topic766	0.00010711	8.0862e-05	2.22253e+06	1.3357e-19	Reject
topic767	4.9473e-05	1.4138e-05	1.47043e+06	9.9035e-147	Reject
topic768	0.00010985	5.6445e-05	1.77481e+06	5.9928e-80	Reject
topic769	6.5113e-05	4.3841e-05	2.26588e+06	3.1358e-16	Reject
topic770	5.7068e-05	2.3623e-05	1.46001e+06	6.1309e-150	Reject
topic771	7.2901e-05	6.2964e-05	2.45772e+06	0.00015381	Reject
topic772	0.00015018	0.00011304	1.87174e+06	5.1011e-64	Reject
topic773	3.7752e-05	2.8132e-05	2.13203e+06	6.8971e-29	Reject
topic774	0.00013634	7.8346e-05	2.13944e+06	5.4268e-28	Reject
topic775	0.00020919	0.00018138	2.39634e+06	1.3035e-07	Reject
topic776	6.382e-05	1.0588e-05	736682	0	Reject
topic777	8.4125e-05	6.222e-05	2.23152e+06	1.7481e-18	Reject
topic778	0.00013479	0.00014692	2.46933e+06	0.00019468	Reject
topic779	0.00011702	8.6542e-05	2.16182e+06	1.8847e-25	Reject
topic780	2.7022e-05	1.8075e-05	2.23927e+06	3.0211e-18	Reject
topic781	8.5775e-05	8.3876e-05	2.61243e+06	0.36994	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic782	0.00011334	0.00010697	2.40876e+06	4.26e-07	Reject
topic783	0.00011604	8.0182e-05	1.59887e+06	4.0252e-117	Reject
topic784	4.1026e-05	0.00010459	1.68159e+06	6.1915e-99	Reject
topic785	0.00014975	0.0001726	2.39414e+06	8.7629e-08	Reject
topic786	8.5192e-05	3.3504e-05	1.31843e+06	2.9784e-187	Reject
topic787	0.00010106	0.00011344	2.49209e+06	0.0012887	Reject
topic788	0.00012059	7.7762e-05	2.065e+06	2.2337e-36	Reject
topic789	2.8263e-05	3.3405e-05	2.30835e+06	4.7727e-13	Reject
topic790	0.00012878	0.00011324	2.43826e+06	1.117e-05	Reject
topic791	0.00014117	0.00011773	2.37327e+06	7.3763e-09	Reject
topic792	9.2667e-05	9.2359e-05	2.56044e+06	0.071704	Accept
topic793	0.00010352	6.1295e-05	1.74239e+06	5.1993e-87	Reject
topic794	0.00017466	0.00012727	2.14445e+06	3.1162e-27	Reject
topic795	8.6358e-05	5.8565e-05	1.7292e+06	2.201e-89	Reject
topic796	3.7104e-05	2.0734e-05	1.57335e+06	7.0577e-123	Reject
topic797	5.5321e-05	3.7977e-05	1.98234e+06	3.232e-47	Reject
topic798	0.00013332	0.00012061	2.45456e+06	6.5181e-05	Reject
topic799	0.00014325	7.9582e-05	1.32023e+06	2.582e-186	Reject
topic800	8.9747e-05	0.00010825	2.39279e+06	6.4999e-08	Reject
topic801	0.00012782	9.4354e-05	2.03372e+06	1.4552e-40	Reject
topic802	8.7691e-05	5.8406e-05	2.06755e+06	4.5715e-36	Reject
topic803	0.00013175	8.4087e-05	1.88319e+06	1.4216e-61	Reject
topic804	9.0919e-05	8.571e-05	2.53774e+06	0.022974	Reject
topic805	1.9518e-06	1.7213e-06	2.32094e+06	4.1363e-12	Reject
topic806	3.8778e-05	5.75e-05	2.16928e+06	3.3858e-23	Reject
topic807	0.00013909	5.9611e-05	1.64906e+06	9.3695e-105	Reject
topic808	3.014e-05	5.055e-06	1.59515e+06	1.3718e-117	Reject
topic809	0.00025779	0.0001946	1.92273e+06	4.18e-56	Reject
topic810	2.9309e-05	2.1308e-05	1.97359e+06	1.8371e-48	Reject
topic811	0.0001291	8.997e-05	2.31877e+06	7.691e-12	Reject
topic812	1.4848e-05	8.7744e-06	1.98499e+06	7.5817e-47	Reject
topic813	8.207e-05	6.4385e-05	2.31064e+06	1.1503e-12	Reject
topic814	6.6977e-05	4.2883e-05	1.79496e+06	6.4613e-77	Reject
topic815	8.5991e-05	7.8935e-05	2.49298e+06	0.001366	Reject
topic816	0.00012538	5.2853e-05	1.45101e+06	3.1867e-151	Reject
topic817	4.3771e-05	3.7546e-06	771152	0	Reject
topic818	0.00015245	0.00012635	2.23423e+06	2.5017e-18	Reject
topic819	8.6059e-05	7.2078e-05	2.25038e+06	1.736e-17	Reject
topic820	3.2549e-05	2.4337e-05	2.54514e+06	0.35103	Accept
topic821	0.00012039	8.484e-05	2.04723e+06	1.0397e-38	Reject
topic822	9.5323e-05	4.572e-05	1.49156e+06	7.117e-142	Reject
topic823	0.00011984	7.459e-05	2.09726e+06	1.1573e-32	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic824	1.8905e-05	5.3878e-06	1.85347e+06	1.0237e-63	Reject
topic825	3.4613e-05	2.2603e-05	1.87915e+06	7.9875e-63	Reject
topic826	0.00012764	0.00014358	2.40742e+06	3.6569e-07	Reject
topic827	9.3963e-05	9.3347e-05	2.62128e+06	0.43679	Accept
topic828	0.00016805	7.9894e-05	1.25045e+06	5.4163e-207	Reject
topic829	0.00022532	0.00023519	2.62293e+06	0.46102	Accept
topic830	2.7144e-05	3.1065e-05	2.55866e+06	0.072851	Accept
topic831	0.00013741	0.00011653	2.29661e+06	9.7368e-14	Reject
topic832	0.00011727	0.00011091	2.46246e+06	9.7292e-05	Reject
topic833	0.00013644	0.00010665	2.16541e+06	2.3302e-24	Reject
topic834	8.5989e-05	4.1378e-05	1.62378e+06	1.4916e-110	Reject
topic835	0.00014459	0.00011103	2.30198e+06	9.3299e-13	Reject
topic836	0.00012337	7.7163e-05	2.00887e+06	3.423e-43	Reject
topic837	8.6315e-05	7.926e-05	2.52445e+06	0.010948	Reject
topic838	4.3985e-05	2.2144e-05	1.84114e+06	1.9653e-67	Reject
topic839	6.2371e-05	6.122e-05	2.56765e+06	0.087378	Accept
topic840	8.5215e-05	8.9983e-05	2.62218e+06	0.45441	Accept
topic841	0.00012405	0.00013666	2.3948e+06	8.2861e-08	Reject
topic842	9.211e-05	5.7793e-06	856086	0	Reject
topic843	0.00010084	8.566e-05	2.33857e+06	7.0944e-11	Reject
topic844	4.3772e-05	3.2229e-06	1.10723e+06	2.0323e-251	Reject
topic845	0.00011066	1.7387e-05	1.36558e+06	1.1546e-169	Reject
topic846	9.4706e-05	7.7666e-05	2.40921e+06	6.4935e-07	Reject
topic847	9.8539e-05	8.1815e-05	2.31749e+06	2.0586e-12	Reject
topic848	8.8077e-05	8.3778e-05	2.54142e+06	0.026063	Reject
topic849	0.00013481	8.2936e-05	1.78577e+06	6.5078e-79	Reject
topic850	0.00016699	0.00013022	2.11706e+06	2.643e-30	Reject
topic851	0.00034899	0.00032694	2.44201e+06	1.4464e-05	Reject
topic852	0.00017401	0.00014471	2.10477e+06	6.4377e-32	Reject
topic853	5.3872e-05	4.1372e-05	2.32126e+06	7.1193e-12	Reject
topic854	9.3926e-05	0.00014416	2.01139e+06	3.9363e-43	Reject
topic855	0.00011203	0.00010537	2.50713e+06	0.0031658	Reject
topic856	8.9677e-05	0.0001583	1.59389e+06	4.8287e-118	Reject
topic857	0.00018066	0.00013464	2.10693e+06	6.4819e-31	Reject
topic858	0.0001019	8.3643e-05	2.35505e+06	4.673e-10	Reject
topic859	5.5359e-05	2.5931e-05	1.25126e+06	2.1963e-207	Reject
topic860	7.4346e-05	6.5924e-05	2.36765e+06	2.6216e-09	Reject
topic861	7.3655e-05	6.8062e-05	2.58924e+06	0.2286	Accept
topic862	0.00014604	0.00015237	2.60592e+06	0.34463	Accept
topic863	0.00014926	7.8738e-05	1.80276e+06	2.2049e-75	Reject
topic864	7.392e-05	3.9885e-05	1.7051e+06	2.6162e-94	Reject
topic865	0.00014404	0.00015843	2.57778e+06	0.14535	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic866	5.064e-05	3.6991e-05	2.10211e+06	3.3065e-31	Reject
topic867	8.6464e-05	9.3815e-05	2.53021e+06	0.014268	Reject
topic868	0.00012593	0.0001771	2.00168e+06	1.5042e-44	Reject
topic869	0.00010908	6.5976e-05	1.57395e+06	2.3451e-122	Reject
topic870	3.1748e-05	1.3541e-05	1.57567e+06	2.0843e-121	Reject
topic871	9.2494e-05	3.1863e-05	1.55294e+06	3.1874e-126	Reject
topic872	0.00014929	0.00010576	2.24887e+06	2.8339e-17	Reject
topic873	0.00010961	6.9184e-05	1.56863e+06	5.8647e-124	Reject
topic874	0.00012758	0.00012816	2.59301e+06	0.2073	Accept
topic875	9.5538e-05	7.9902e-05	2.25177e+06	1.843e-17	Reject
topic876	3.4725e-05	1.0657e-05	1.85091e+06	1.4576e-66	Reject
topic877	0.00014499	8.7521e-05	1.5268e+06	9.6939e-134	Reject
topic878	0.00010314	0.00013055	2.30463e+06	2.1772e-13	Reject
topic879	0.00011051	4.9135e-05	1.80278e+06	1.1188e-75	Reject
topic880	0.00011771	6.6523e-05	1.74778e+06	8.2642e-86	Reject
topic881	9.6696e-05	4.9952e-05	1.57726e+06	1.9758e-121	Reject
topic882	9.6561e-05	7.2011e-05	2.15638e+06	2.5728e-26	Reject
topic883	8.3012e-05	5.1878e-05	2.14868e+06	4.0811e-27	Reject
topic884	0.00010147	9.9875e-05	2.57411e+06	0.11767	Accept
topic885	0.00013481	0.0001041	2.11892e+06	3.278e-30	Reject
topic886	7.2113e-05	5.5666e-05	2.44345e+06	2.0553e-05	Reject
topic887	0.00010485	5.2061e-05	1.73894e+06	1.6198e-87	Reject
topic888	0.00011987	0.00011326	2.5301e+06	0.017993	Reject
topic889	0.0001039	8.4424e-05	2.22645e+06	1.5767e-19	Reject
topic890	0.00014568	9.3062e-05	1.95766e+06	5.0845e-51	Reject
topic891	0.00010576	2.3821e-05	861706	0	Reject
topic892	9.6344e-05	9.021e-05	2.53157e+06	0.014429	Reject
topic893	0.00012212	7.9557e-05	2.09661e+06	1.2478e-32	Reject
topic894	8.4085e-05	3.2214e-05	1.36258e+06	3.7023e-175	Reject
topic895	5.278e-05	1.6599e-06	944410	2.0212e-304	Reject
topic896	0.00012379	0.00010742	2.17024e+06	6.5713e-25	Reject
topic897	5.5779e-05	1.0723e-05	726075	0	Reject
topic898	4.5574e-05	2.4961e-05	1.68407e+06	2.1576e-98	Reject
topic899	0.00014419	0.00011975	2.25482e+06	8.6765e-17	Reject
topic900	0.000115	7.503e-05	1.99789e+06	6.2572e-45	Reject
topic901	0.00017695	0.0001297	2.09991e+06	2.3382e-32	Reject
topic902	1.6419e-08	5.1788e-08	751476	9.2238e-41	Reject
topic903	7.9375e-05	0.00011637	2.3572e+06	7.2545e-10	Reject
topic904	8.5474e-05	0.0001206	2.17234e+06	1.682e-24	Reject
topic905	6.0106e-05	4.8556e-05	2.43864e+06	1.1496e-05	Reject
topic906	5.0711e-05	4.4882e-05	2.21506e+06	1.2248e-20	Reject
topic907	0.00014594	0.00010762	2.21018e+06	8.4095e-21	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic908	9.7886e-05	8.6827e-05	2.33887e+06	4.5612e-11	Reject
topic909	0.00013234	0.00011493	2.35914e+06	8.2486e-10	Reject
topic910	0.00015628	8.3759e-05	1.62236e+06	1.1813e-110	Reject
topic911	0.00018449	0.00014546	2.23135e+06	7.9868e-19	Reject
topic912	6.3691e-05	3.5561e-05	1.64219e+06	1.1413e-107	Reject
topic913	0.00011918	8.3527e-05	2.0581e+06	2.3751e-37	Reject
topic914	0.00015734	0.00021655	1.9986e+06	3.0795e-45	Reject
topic915	0.00015793	0.00012844	2.14842e+06	6.319e-27	Reject
topic916	0.00011042	0.00013722	2.43248e+06	6.9348e-06	Reject
topic917	9.2462e-05	8.3303e-05	2.46229e+06	0.00010507	Reject
topic918	0.00024038	0.00019848	2.37285e+06	6.0394e-09	Reject
topic919	8.9418e-05	5.6337e-05	2.20208e+06	1.2111e-21	Reject
topic920	3.9814e-05	1.9758e-05	1.75127e+06	2.5727e-85	Reject
topic921	0.00013138	0.00010409	2.01619e+06	7.2878e-43	Reject
topic922	3.8766e-06	1.897e-06	1.82034e+06	2.1267e-72	Reject
topic923	0.00015279	9.7404e-05	1.68427e+06	1.0361e-98	Reject
topic924	5.5978e-05	5.793e-05	2.51675e+06	0.0073829	Reject
topic925	6.2823e-05	1.8099e-05	1.09020e+06	2.7649e-255	Reject
topic926	2.3346e-05	8.7752e-06	1.30649e+06	7.6916e-191	Reject
topic927	0.00010244	7.5071e-05	2.16412e+06	1.5896e-25	Reject
topic928	6.428e-05	3.9556e-05	2.04022e+06	1.6744e-38	Reject
topic929	0.00014862	0.00018775	2.22859e+06	1.9619e-19	Reject
topic930	9.3202e-05	9.368e-05	2.62551e+06	0.48486	Accept
topic931	7.9694e-05	0.00013106	2.01978e+06	2.4851e-38	Reject
topic932	0.00013463	0.00011371	2.3611e+06	1.2568e-09	Reject
topic933	0.00012088	8.5728e-05	2.26686e+06	6.5219e-16	Reject
topic934	2.3735e-05	5.3043e-06	1.45924e+06	1.1845e-147	Reject
topic935	7.6102e-05	3.239e-05	1.55554e+06	1.9141e-125	Reject
topic936	2.9554e-05	5.8718e-06	1.46238e+06	6.2033e-149	Reject
topic937	5.565e-05	6.2413e-05	2.49367e+06	0.0013296	Reject
topic938	0.00010106	8.5665e-05	2.36468e+06	2.0163e-09	Reject
topic939	0.00012487	0.00010689	2.1264e+06	2.1803e-29	Reject
topic940	8.2818e-05	6.8192e-05	2.08394e+06	3.132e-34	Reject
topic941	0.00012963	7.3352e-05	1.76958e+06	6.7888e-82	Reject
topic942	5.8171e-05	3.3153e-05	1.43432e+06	9.6791e-156	Reject
topic943	9.7063e-05	0.00013505	2.06072e+06	3.7623e-37	Reject
topic944	0.00011725	9.7856e-05	2.34344e+06	1.2268e-10	Reject
topic945	0.0001444	0.00014982	2.52101e+06	0.0088667	Reject
topic946	8.9332e-05	1.5428e-05	1.75382e+06	3.5503e-83	Reject
topic947	0.00014985	0.00010655	2.17462e+06	4.6462e-24	Reject
topic948	0.00016217	7.845e-05	1.3821e+06	2.8384e-169	Reject
topic949	7.0134e-05	5.9628e-05	2.38896e+06	5.3448e-08	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic950	0.00014039	0.00016402	2.47713e+06	0.0005732	Reject
topic951	0.00014591	0.00012051	2.36632e+06	3.4236e-09	Reject
topic952	0.00015459	0.00012135	2.19881e+06	4.8756e-22	Reject
topic953	0.00013247	8.7396e-05	1.92241e+06	7.3501e-56	Reject
topic954	0.0001254	9.6437e-05	1.93111e+06	8.1075e-55	Reject
topic955	6.068e-05	0.00022	1.50344e+06	7.4875e-138	Reject
topic956	3.0931e-06	3.5184e-06	2.52606e+06	0.173	Accept
topic957	9.2948e-05	2.645e-05	1.28869e+06	3.553e-196	Reject
topic958	6.5963e-05	2.1834e-05	1.3048e+06	2.6547e-187	Reject
topic959	5.9296e-05	4.67e-06	1.25692e+06	1.7507e-205	Reject
topic960	8.8096e-05	6.5639e-05	2.21573e+06	2.1434e-20	Reject
topic961	8.4845e-05	4.8856e-05	1.79386e+06	1.9814e-77	Reject
topic962	0.00017542	0.00013934	2.10429e+06	7.4153e-32	Reject
topic963	8.7191e-05	3.0703e-05	1.69434e+06	5.9448e-96	Reject
topic964	9.6682e-05	0.00010393	2.62756e+06	0.49242	Accept
topic965	0.00016808	0.00012706	2.00223e+06	1.2753e-44	Reject
topic966	0.00014762	0.00015087	2.60875e+06	0.35825	Accept
topic967	0.00012846	9.7926e-05	1.94579e+06	9.1565e-53	Reject
topic968	0.00017425	0.00013465	1.90574e+06	1.3781e-58	Reject
topic969	5.8414e-05	6.3192e-05	2.5958e+06	0.22555	Accept
topic970	1.6969e-05	8.837e-06	1.87137e+06	4.2956e-64	Reject
topic971	0.00011589	8.7952e-05	2.07307e+06	1.2149e-35	Reject
topic972	8.1979e-05	3.0745e-05	1.3818e+06	9.4181e-170	Reject
topic973	6.5944e-05	5.7888e-05	2.57062e+06	0.11191	Accept
topic974	0.00012669	0.00011956	2.50254e+06	0.0026921	Reject
topic975	0.00011503	7.2661e-05	1.78195e+06	2.8118e-79	Reject
topic976	0.00014386	0.00015898	2.40838e+06	5.2707e-07	Reject
topic977	0.00019771	0.00021117	2.42764e+06	3.7864e-06	Reject
topic978	0.00015503	0.00013007	2.40834e+06	7.4961e-07	Reject
topic979	5.1381e-05	4.3577e-05	2.35504e+06	1.522e-09	Reject
topic980	8.9447e-05	7.2524e-05	2.21987e+06	3.3135e-20	Reject
topic981	0.00010435	7.7576e-05	2.1292e+06	1.1756e-28	Reject
topic982	0.00014845	0.00018383	2.31449e+06	1.81e-12	Reject
topic983	1.8321e-05	1.2334e-05	1.9328e+06	1.05e-53	Reject
topic984	3.6295e-05	1.0967e-05	1.19737e+06	5.7532e-222	Reject
topic985	0.00014739	0.0001184	2.14151e+06	9.012e-28	Reject
topic986	0.00010116	0.00010891	2.4378e+06	1.0565e-05	Reject
topic987	5.1522e-05	3.768e-05	2.35869e+06	2.9305e-09	Reject
topic988	8.8653e-05	5.7835e-05	2.00197e+06	2.1205e-44	Reject
topic989	2.0456e-06	2.4377e-06	1.89862e+06	5.8413e-05	Reject
topic990	0.00012506	9.0681e-05	1.81044e+06	2.571e-74	Reject
topic991	2.2736e-05	6.6192e-05	1.63686e+06	1.1507e-102	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic992	0.00013231	8.6243e-05	1.76465e+06	8.4943e-83	Reject
topic993	7.0397e-05	6.7027e-05	2.4641e+06	0.00011244	Reject
topic994	3.8778e-05	3.8028e-05	2.40238e+06	4.3412e-07	Reject
topic995	9.371e-05	6.6518e-05	2.04181e+06	3.8387e-39	Reject
topic996	7.47e-05	4.7662e-05	1.68581e+06	2.2165e-98	Reject
topic997	0.00010873	0.00010403	2.54153e+06	0.024796	Reject
topic998	9.6703e-05	7.2613e-05	2.42179e+06	2.5943e-06	Reject
topic999	0.00014031	9.7485e-05	2.0458e+06	6.8494e-39	Reject
topic1000	8.9869e-05	6.0533e-05	1.7856e+06	2.9061e-79	Reject
topic1001	7.2252e-05	4.9425e-05	2.20292e+06	4.2656e-21	Reject
topic1002	7.8228e-05	5.0258e-05	2.14113e+06	8.4003e-28	Reject
topic1003	6.2746e-05	6.2007e-05	2.60443e+06	0.29642	Accept
topic1004	0.00012961	0.00012753	2.58928e+06	0.18432	Accept
topic1005	2.536e-05	9.1785e-05	1.48234e+06	3.9707e-131	Reject
topic1006	0.00010829	6.5286e-05	1.97161e+06	5.2043e-49	Reject
topic1007	0.00011571	8.4737e-05	1.81597e+06	8.1477e-74	Reject
topic1008	7.9436e-05	5.261e-05	2.02689e+06	4.4947e-41	Reject
topic1009	0.00021377	0.00011987	1.69425e+06	5.932e-96	Reject
topic1010	0.00013883	0.00014944	2.49395e+06	0.0017189	Reject
topic1011	7.4235e-05	8.059e-05	2.54424e+06	0.037977	Reject
topic1012	0.00016132	7.3572e-05	1.28113e+06	9.1601e-198	Reject
topic1013	1.0724e-05	1.2195e-05	2.43889e+06	0.00043143	Reject
topic1014	1.7978e-05	1.1076e-05	2.00897e+06	3.4371e-43	Reject
topic1015	0.00012526	0.00010579	2.16578e+06	3.7218e-25	Reject
topic1016	0.00023608	0.00023792	2.61888e+06	0.40545	Accept
topic1017	0.00013063	9.0865e-05	2.08592e+06	7.068e-34	Reject
topic1018	4.8105e-05	0.00011162	1.7005e+06	3.2164e-93	Reject
topic1019	0.00012555	0.00013151	2.5604e+06	0.071271	Accept
topic1020	7.4382e-05	6.606e-05	2.25204e+06	2.3816e-17	Reject
topic1021	9.4378e-07	7.4445e-07	2.56956e+06	0.31462	Accept
topic1022	4.4864e-05	2.8742e-05	2.02431e+06	6.8336e-41	Reject
topic1023	6.3067e-05	5.6537e-05	2.42095e+06	4.1705e-06	Reject
topic1024	0.00013302	8.9065e-05	1.89888e+06	1.1869e-59	Reject
topic1025	0.00010329	0.00013948	2.42905e+06	5.4813e-06	Reject
topic1026	0.0001067	6.4256e-05	1.71129e+06	4.4632e-93	Reject
topic1027	0.00018028	6.5163e-05	1.15938e+06	5.4493e-233	Reject
topic1028	0.00014534	0.00010545	1.8562e+06	9.2333e-67	Reject
topic1029	7.2104e-05	4.74e-05	2.01697e+06	1.253e-42	Reject
topic1030	4.4953e-05	3.4848e-05	2.13212e+06	1.9219e-28	Reject
topic1031	6.0307e-05	2.2914e-05	1.53498e+06	3.0093e-130	Reject
topic1032	0.00012486	0.00010325	2.11244e+06	6.1987e-31	Reject
topic1033	2.6251e-05	4.1184e-05	2.03264e+06	1.0528e-40	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1034	0.00012452	0.00012067	2.55878e+06	0.06024	Accept
topic1035	0.0001038	8.3953e-05	2.23133e+06	4.2134e-19	Reject
topic1036	7.3886e-05	6.1133e-05	2.26648e+06	3.4374e-16	Reject
topic1037	0.00011996	0.00018701	1.88235e+06	1.3423e-62	Reject
topic1038	0.00011435	5.9308e-05	2.06434e+06	1.5667e-35	Reject
topic1039	4.9834e-05	3.5788e-05	1.77405e+06	3.1156e-81	Reject
topic1040	1.9088e-05	2.4637e-05	2.34161e+06	1.8342e-09	Reject
topic1041	0.00015567	0.00017669	2.42386e+06	2.2557e-06	Reject
topic1042	4.2346e-05	2.1786e-05	1.67359e+06	2.8052e-99	Reject
topic1043	2.1784e-05	2.6878e-06	1.3767e+06	6.9768e-169	Reject
topic1044	0.00015139	0.00013418	2.2442e+06	4.3067e-18	Reject
topic1045	0.00010754	7.4124e-05	1.87385e+06	7.9169e-64	Reject
topic1046	5.1692e-05	4.0639e-05	2.3678e+06	3.5939e-09	Reject
topic1047	4.8593e-05	2.0703e-05	1.65026e+06	3.1746e-101	Reject
topic1048	0.00011843	0.00010284	2.18855e+06	5.1903e-23	Reject
topic1049	0.00022741	0.0001743	2.03211e+06	8.9988e-41	Reject
topic1050	0.00013479	0.00012167	2.42857e+06	3.7076e-06	Reject
topic1051	7.0686e-05	4.1345e-05	1.55453e+06	7.9769e-127	Reject
topic1052	8.1145e-05	6.7297e-05	2.24078e+06	3.2941e-18	Reject
topic1053	0.00013821	8.6333e-05	1.71952e+06	1.8901e-91	Reject
topic1054	0.0001176	7.8429e-05	1.94981e+06	1.2616e-51	Reject
topic1055	9.6976e-05	3.4297e-05	2.13065e+06	7.9256e-28	Reject
topic1056	1.1031e-05	1.101e-05	2.61155e+06	0.37259	Accept
topic1057	6.3858e-05	4.7611e-05	2.24065e+06	7.0876e-18	Reject
topic1058	5.0562e-05	3.526e-05	2.25638e+06	6.4834e-17	Reject
topic1059	5.2899e-05	6.437e-05	2.38517e+06	2.9168e-08	Reject
topic1060	8.1391e-05	0.00012724	2.18604e+06	2.9122e-23	Reject
topic1061	0.00011994	7.2093e-05	1.81279e+06	4.5951e-74	Reject
topic1062	3.768e-05	2.3797e-05	1.71642e+06	2.0561e-92	Reject
topic1063	0.00013405	4.9982e-05	1.13057e+06	6.2087e-243	Reject
topic1064	0.00014177	0.00011478	2.20165e+06	2.6089e-21	Reject
topic1065	7.9857e-08	8.199e-08	273624	0.2761	Accept
topic1066	0.00010657	0.00010437	2.58371e+06	0.15945	Accept
topic1067	0.00022354	0.00016232	2.00813e+06	8.033e-44	Reject
topic1068	0.00013554	8.8757e-05	1.95976e+06	1.9657e-50	Reject
topic1069	6.0736e-05	6.7845e-06	959818	7.3593e-301	Reject
topic1070	8.9241e-05	0.00011286	2.12513e+06	1.2205e-29	Reject
topic1071	0.00014653	0.00013041	2.4249e+06	8.589e-06	Reject
topic1072	6.9675e-05	6.8858e-05	2.57784e+06	0.14719	Accept
topic1073	1.0817e-05	3.8515e-05	1.63018e+06	5.1731e-106	Reject
topic1074	0.00011005	8.1183e-05	2.27239e+06	1.4541e-15	Reject
topic1075	0.00010699	0.00010568	2.57361e+06	0.12653	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1076	8.9391e-05	7.0407e-05	2.12999e+06	4.1461e-29	Reject
topic1077	0.00018393	0.00018925	2.52526e+06	0.011412	Reject
topic1078	9.6036e-05	7.0841e-05	2.06498e+06	8.6923e-36	Reject
topic1079	5.9931e-05	4.498e-05	2.17954e+06	6.9438e-24	Reject
topic1080	4.354e-05	3.6826e-05	2.32521e+06	1.1182e-11	Reject
topic1081	2.187e-05	1.6266e-05	2.15544e+06	3.3691e-26	Reject
topic1082	4.844e-05	3.1305e-05	2.16828e+06	6.7477e-24	Reject
topic1083	1.8229e-05	1.562e-05	2.41399e+06	2.8947e-06	Reject
topic1084	6.0236e-05	5.5578e-05	2.48034e+06	0.00048031	Reject
topic1085	0.00010126	5.1549e-05	1.49252e+06	1.172e-141	Reject
topic1086	0.00010178	7.5503e-05	2.23173e+06	5.5363e-19	Reject
topic1087	0.0001213	0.00012699	2.48464e+06	0.00061554	Reject
topic1088	9.9408e-05	5.1546e-05	1.86674e+06	7.6833e-65	Reject
topic1089	0.00013185	0.00013264	2.61915e+06	0.45933	Accept
topic1090	9.4214e-05	6.4128e-05	1.72953e+06	3.6759e-89	Reject
topic1091	8.2797e-05	6.7448e-05	2.31137e+06	7.6539e-13	Reject
topic1092	0.00012877	6.0955e-05	1.51655e+06	4.7984e-135	Reject
topic1093	5.1474e-05	2.2235e-05	1.83232e+06	7.1302e-70	Reject
topic1094	0.00014889	0.00014857	2.52601e+06	0.011202	Reject
topic1095	1.8481e-06	1.8411e-06	2.56635e+06	0.10398	Accept
topic1096	5.2749e-05	3.8955e-05	2.27243e+06	1.228e-15	Reject
topic1097	0.00010878	7.8405e-05	2.14126e+06	1.8043e-27	Reject
topic1098	0.00016074	0.00012167	2.22591e+06	1.7197e-19	Reject
topic1099	0.00020419	0.00023486	2.35691e+06	6.055e-10	Reject
topic1100	0.0002106	0.00019449	2.47683e+06	0.0003302	Reject
topic1101	6.8308e-05	7.2827e-05	2.49596e+06	0.0017071	Reject
topic1102	0.00010245	5.8598e-05	1.64663e+06	2.3517e-106	Reject
topic1103	0.0001466	0.00015365	2.57384e+06	0.1165	Accept
topic1104	0.00012551	7.7066e-05	1.78614e+06	7.8802e-79	Reject
topic1105	0.00013911	0.00010807	2.34975e+06	4.8231e-10	Reject
topic1106	0.00011119	4.0773e-05	1.297e+06	8.8764e-194	Reject
topic1107	0.00011547	0.00010508	2.37938e+06	1.403e-08	Reject
topic1108	2.727e-05	3.4693e-05	2.48234e+06	0.0006628	Reject
topic1109	0.00010615	7.2915e-05	2.31927e+06	3.2325e-12	Reject
topic1110	0.00013009	6.626e-05	1.68869e+06	2.9466e-97	Reject
topic1111	0.00013196	9.5635e-05	2.09015e+06	2.2136e-33	Reject
topic1112	4.3569e-05	1.1337e-05	1.33324e+06	5.9229e-180	Reject
topic1113	0.00010992	7.6758e-05	2.40032e+06	1.6016e-07	Reject
topic1114	0.00015456	0.00013978	2.41817e+06	1.7068e-06	Reject
topic1115	0.00012224	8.0727e-05	2.05031e+06	5.7406e-38	Reject
topic1116	9.295e-05	5.0123e-05	1.86677e+06	5.4469e-65	Reject
topic1117	0.00013472	0.00010248	1.96372e+06	5.3033e-50	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1118	0.00010165	4.214e-05	1.45436e+06	2.2763e-151	Reject
topic1119	0.00018955	9.5194e-05	1.87084e+06	4.8633e-64	Reject
topic1120	0.00010521	7.3331e-05	2.03249e+06	1.008e-40	Reject
topic1121	0.00018849	0.00019513	2.52979e+06	0.013938	Reject
topic1122	0.00012397	7.72e-05	1.60438e+06	1.6197e-115	Reject
topic1123	8.126e-05	4.8496e-05	1.70887e+06	9.489e-94	Reject
topic1124	6.1639e-05	8.6174e-05	2.12503e+06	2.0019e-29	Reject
topic1125	3.9096e-05	2.4592e-05	1.82378e+06	2.7682e-72	Reject
topic1126	0.00012638	7.9559e-05	1.8145e+06	1.3834e-73	Reject
topic1127	0.00013093	0.00012586	2.62426e+06	0.47288	Accept
topic1128	0.00013535	0.00014234	2.50587e+06	0.0031227	Reject
topic1129	0.00012264	0.00010397	2.41605e+06	1.2246e-06	Reject
topic1130	0.00013327	9.3538e-05	2.39113e+06	6.8865e-08	Reject
topic1131	4.4951e-05	2.3567e-05	1.6509e+06	8.1129e-106	Reject
topic1132	0.0001124	9.0508e-05	2.22562e+06	3.8248e-19	Reject
topic1133	0.00013976	0.00013628	2.51336e+06	0.0047821	Reject
topic1134	0.00018093	0.00017208	2.48764e+06	0.00092496	Reject
topic1135	0.00011308	3.4342e-05	1.22614e+06	1.0383e-214	Reject
topic1136	0.00012414	0.00010322	2.22861e+06	4.4535e-19	Reject
topic1137	3.733e-05	1.8716e-05	2.13674e+06	4.7609e-28	Reject
topic1138	6.0055e-05	5.7543e-05	2.55633e+06	0.059987	Accept
topic1139	0.00012733	0.00015603	2.29021e+06	3.3113e-14	Reject
topic1140	4.8278e-05	3.2009e-05	2.23057e+06	8.4589e-19	Reject
topic1141	9.2788e-05	8.2613e-05	2.32214e+06	3.5969e-12	Reject
topic1142	0.00012796	8.9281e-05	1.96127e+06	2.3486e-50	Reject
topic1143	6.3785e-05	5.604e-05	2.4196e+06	2.5583e-06	Reject
topic1144	4.0814e-05	5.3715e-06	927296	0	Reject
topic1145	0.0002344	0.0002964	2.14715e+06	2.8209e-27	Reject
topic1146	5.3844e-05	3.6765e-05	2.10021e+06	2.53e-32	Reject
topic1147	0.00013368	0.00010905	2.40554e+06	3.7716e-07	Reject
topic1148	8.9447e-05	6.5657e-05	2.16985e+06	1.2088e-24	Reject
topic1149	5.8268e-05	3.7656e-05	2.22207e+06	7.7124e-19	Reject
topic1150	2.6605e-05	1.9475e-05	2.2035e+06	4.2104e-20	Reject
topic1151	5.8984e-05	7.0141e-05	2.49007e+06	0.0010937	Reject
topic1152	8.2236e-05	6.24e-05	2.25213e+06	2.9145e-17	Reject
topic1153	9.5502e-05	7.0931e-05	2.23418e+06	6.0154e-19	Reject
topic1154	6.3162e-05	3.6707e-05	1.64298e+06	2.4709e-107	Reject
topic1155	0.00014087	6.5491e-05	1.53132e+06	1.8049e-132	Reject
topic1156	0.00017473	0.00013144	1.98795e+06	3.5668e-46	Reject
topic1157	0.00013833	0.00010397	2.37221e+06	4.7983e-09	Reject
topic1158	0.00013415	8.15e-05	1.73534e+06	7.022e-88	Reject
topic1159	0.00013377	0.00010997	2.19411e+06	2.2109e-22	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1160	1.6235e-05	2.8302e-05	2.13004e+06	3.0447e-28	Reject
topic1161	0.00011773	0.00011063	2.55893e+06	0.063591	Accept
topic1162	0.00015318	9.9337e-05	2.39773e+06	4.7411e-07	Reject
topic1163	7.0983e-05	3.3705e-05	2.05675e+06	2.1529e-37	Reject
topic1164	3.1302e-05	1.93e-05	2.1077e+06	3.102e-30	Reject
topic1165	0.00017197	0.00013427	2.18185e+06	2.2637e-23	Reject
topic1166	0.00013367	7.2732e-05	1.76448e+06	2.3641e-82	Reject
topic1167	0.00011425	8.601e-05	2.11338e+06	1.0036e-30	Reject
topic1168	0.00014332	0.00010628	1.95208e+06	1.4922e-51	Reject
topic1169	7.7149e-05	5.4838e-05	1.73812e+06	5.0853e-88	Reject
topic1170	0.00010282	7.1313e-05	1.73477e+06	1.1915e-88	Reject
topic1171	8.8862e-05	4.8517e-05	1.76013e+06	2.5559e-83	Reject
topic1172	6.1947e-05	6.2482e-05	2.62829e+06	0.49892	Accept
topic1173	5.2012e-05	4.0139e-05	2.24246e+06	8.4777e-18	Reject
topic1174	0.00011383	7.705e-05	2.23306e+06	5.8348e-19	Reject
topic1175	0.00010267	9.5054e-05	2.40242e+06	2.0491e-07	Reject
topic1176	2.1041e-05	8.8816e-06	1.63549e+06	3.265e-108	Reject
topic1177	0.00013212	8.3717e-05	1.78006e+06	2.8088e-80	Reject
topic1178	5.6461e-06	9.0363e-06	1.83815e+06	1.7852e-18	Reject
topic1179	0.00012691	9.0277e-05	2.12494e+06	1.1621e-29	Reject
topic1180	4.9383e-05	3.4347e-05	2.26804e+06	6.7542e-16	Reject
topic1181	3.9027e-05	1.9068e-05	1.81038e+06	1.783e-74	Reject
topic1182	0.00013865	0.00011861	2.35584e+06	5.2204e-10	Reject
topic1183	6.8139e-05	5.4423e-05	2.18556e+06	5.1611e-23	Reject
topic1184	0.00013044	8.4063e-05	1.79095e+06	8.3381e-78	Reject
topic1185	0.00013371	7.3705e-05	1.86673e+06	1.0597e-64	Reject
topic1186	0.00011567	8.384e-05	2.12483e+06	1.4661e-29	Reject
topic1187	2.3888e-05	1.0561e-05	1.82979e+06	2.593e-67	Reject
topic1188	9.8451e-05	9.2144e-05	2.5532e+06	0.046761	Reject
topic1189	0.00013535	0.00012579	2.4906e+06	0.0014669	Reject
topic1190	6.5438e-05	1.1747e-05	646060	0	Reject
topic1191	0.0001141	0.00011579	2.61709e+06	0.39008	Accept
topic1192	5.8537e-05	3.0179e-05	1.72134e+06	1.92e-91	Reject
topic1193	7.3478e-05	3.7295e-05	1.64592e+06	1.5333e-106	Reject
topic1194	4.1063e-05	0.0001442	1.45018e+06	1.7936e-151	Reject
topic1195	7.7884e-05	6.2519e-05	2.36813e+06	3.2411e-09	Reject
topic1196	0.00012402	0.00010477	2.31557e+06	1.5199e-12	Reject
topic1197	0.00013997	7.4455e-05	1.16532e+06	1.5305e-233	Reject
topic1198	0.0002451	0.00017813	1.96168e+06	1.9466e-50	Reject
topic1199	0.00011937	0.00013965	2.34023e+06	7.6919e-11	Reject
topic1200	0.00011757	0.00011411	2.56151e+06	0.064518	Accept
topic1201	6.3938e-05	1.1869e-05	1.01057e+06	6.2349e-282	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1202	0.00013881	7.9415e-05	1.87852e+06	5.5476e-61	Reject
topic1203	0.00012638	6.1028e-05	1.84328e+06	1.2677e-68	Reject
topic1204	8.7099e-05	6.2438e-05	2.13322e+06	1.1706e-28	Reject
topic1205	7.4433e-05	4.7263e-05	1.92747e+06	6.2515e-55	Reject
topic1206	3.5903e-05	2.0012e-05	1.64065e+06	2.8173e-107	Reject
topic1207	7.9324e-05	6.6238e-05	2.27086e+06	7.628e-16	Reject
topic1208	5.2197e-05	3.8235e-05	2.18346e+06	2.0739e-23	Reject
topic1209	0.00011072	9.5285e-05	2.26648e+06	2.8721e-16	Reject
topic1210	6.9061e-05	7.7365e-05	2.48439e+06	0.00078252	Reject
topic1211	9.0853e-05	8.675e-05	2.48989e+06	0.0010869	Reject
topic1212	9.215e-05	8.1841e-05	2.38351e+06	2.0586e-08	Reject
topic1213	9.5579e-05	4.32e-05	1.19239e+06	4.6343e-225	Reject
topic1214	8.5607e-05	0.00010562	2.276e+06	1.6082e-15	Reject
topic1215	0.00011136	5.3411e-05	1.58756e+06	1.0591e-118	Reject
topic1216	0.00016128	0.00012557	2.12825e+06	3.4704e-29	Reject
topic1217	0.00015268	0.00011229	2.1899e+06	8.7992e-23	Reject
topic1218	0.00014927	7.5759e-05	1.46211e+06	3.1363e-148	Reject
topic1219	6.0699e-05	5.1877e-05	2.3381e+06	4.7842e-11	Reject
topic1220	8.6303e-05	5.0828e-05	1.40607e+06	4.3369e-164	Reject
topic1221	0.00015449	0.00011625	2.04893e+06	2.2757e-38	Reject
topic1222	1.5203e-05	6.5448e-05	1.69641e+06	1.6173e-82	Reject
topic1223	9.7639e-05	7.6764e-05	2.32754e+06	1.3682e-11	Reject
topic1224	7.2295e-05	6.8082e-05	2.48116e+06	0.00051262	Reject
topic1225	0.00013953	4.9101e-05	1.24175e+06	3.1718e-207	Reject
topic1226	7.0655e-05	3.1274e-05	1.78349e+06	5.1913e-79	Reject
topic1227	6.5811e-06	2.2473e-06	2.20936e+06	8.5519e-14	Reject
topic1228	0.00013112	9.6094e-05	2.04515e+06	7.3863e-39	Reject
topic1229	0.00016003	0.00014854	2.50271e+06	0.0043394	Reject
topic1230	5.0474e-05	3.9667e-05	2.35051e+06	3.9341e-10	Reject
topic1231	0.00010469	0.0001043	2.58695e+06	0.17082	Accept
topic1232	0.00015155	0.00010038	2.04424e+06	7.7992e-39	Reject
topic1233	0.00013464	0.00012022	2.38903e+06	4.7029e-08	Reject
topic1234	0.00013759	6.9491e-05	2.16829e+06	4.1931e-25	Reject
topic1235	4.7487e-05	3.5725e-05	2.2371e+06	1.323e-18	Reject
topic1236	0.00010121	7.4872e-05	2.07984e+06	1.7305e-34	Reject
topic1237	8.922e-05	6.6844e-05	2.30482e+06	5.2146e-13	Reject
topic1238	6.9232e-05	6.2788e-05	2.52705e+06	0.011122	Reject
topic1239	4.6595e-05	6.9789e-06	1.04606e+06	5.1532e-265	Reject
topic1240	8.0628e-05	3.7336e-05	1.77842e+06	3.0338e-80	Reject
topic1241	0.00013833	0.00015957	2.33734e+06	6.864e-11	Reject
topic1242	0.00011214	8.3461e-05	1.90254e+06	6.1299e-59	Reject
topic1243	0.00010966	0.00016156	2.28515e+06	8.0597e-15	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1244	6.1741e-05	0.0001202	2.00399e+06	4.094e-44	Reject
topic1245	5.5524e-05	4.3024e-05	2.13251e+06	1.005e-28	Reject
topic1246	3.5036e-05	2.1017e-05	1.86344e+06	6.005e-65	Reject
topic1247	0.00013882	7.0201e-05	1.53857e+06	2.208e-130	Reject
topic1248	0.00012713	7.3333e-05	1.71689e+06	8.2685e-92	Reject
topic1249	0.00011293	9.9578e-05	2.38854e+06	4.4293e-08	Reject
topic1250	0.00013559	8.8673e-05	1.50412e+06	5.2881e-139	Reject
topic1251	0.00012616	0.00010383	2.2141e+06	1.5574e-20	Reject
topic1252	2.1863e-05	2.2991e-05	2.56902e+06	0.20702	Accept
topic1253	0.0001621	0.00011275	2.37446e+06	5.1052e-08	Reject
topic1254	8.9253e-05	3.4967e-05	1.46826e+06	7.3115e-148	Reject
topic1255	0.00013652	0.00011456	2.42414e+06	2.9532e-06	Reject
topic1256	8.249e-05	2.6664e-05	1.89229e+06	2.7434e-56	Reject
topic1257	6.9299e-05	7.5439e-05	2.53689e+06	0.019377	Reject
topic1258	6.2658e-05	6.1938e-05	2.62248e+06	0.43685	Accept
topic1259	0.00011333	8.4946e-05	1.85231e+06	1.4787e-67	Reject
topic1260	6.7268e-05	5.8241e-05	2.39288e+06	1.2734e-07	Reject
topic1261	5.3309e-05	3.9774e-05	2.23163e+06	4.4665e-19	Reject
topic1262	8.1112e-05	4.5588e-05	1.74434e+06	1.1432e-86	Reject
topic1263	5.5606e-05	1.6195e-05	1.34347e+06	1.3372e-170	Reject
topic1264	9.305e-05	6.3353e-05	2.09029e+06	1.7927e-33	Reject
topic1265	9.6259e-05	7.2821e-05	1.97244e+06	9.4267e-49	Reject
topic1266	7.6069e-05	5.887e-05	2.19145e+06	1.2375e-22	Reject
topic1267	0.0001443	0.00011742	2.40468e+06	3.8855e-07	Reject
topic1268	9.5796e-05	7.8335e-05	2.07331e+06	1.6769e-35	Reject
topic1269	8.1824e-05	7.4371e-05	2.29915e+06	1.0521e-13	Reject
topic1270	1.2322e-05	5.9249e-05	1.96982e+06	1.7371e-45	Reject
topic1271	0.00010098	5.2695e-05	1.80574e+06	8.0083e-75	Reject
topic1272	0.00012409	9.6992e-05	2.26836e+06	8.6983e-16	Reject
topic1273	0.00010596	8.7394e-05	2.27313e+06	1.415e-15	Reject
topic1274	1.1455e-05	1.2369e-06	1.40097e+06	1.7088e-159	Reject
topic1275	0.00010328	0.00010772	2.54531e+06	0.031943	Reject
topic1276	2.6059e-05	9.01e-06	1.40986e+06	1.0719e-162	Reject
topic1277	7.0606e-05	3.9541e-05	1.8405e+06	1.4678e-69	Reject
topic1278	8.0676e-05	8.5481e-05	2.53694e+06	0.019426	Reject
topic1279	2.3805e-05	8.0071e-06	1.76413e+06	3.2372e-82	Reject
topic1280	0.0001223	8.241e-05	2.05555e+06	1.1456e-37	Reject
topic1281	0.00018679	0.00012571	1.89019e+06	1.3226e-60	Reject
topic1282	0.00010989	9.8099e-05	2.51399e+06	0.0099822	Reject
topic1283	0.00011262	0.00013376	2.33398e+06	2.2001e-11	Reject
topic1284	7.4319e-05	4.4345e-05	2.07677e+06	1.419e-33	Reject
topic1285	8.1661e-05	7.6601e-05	2.53051e+06	0.015333	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1286	0.00011282	8.2317e-05	2.21916e+06	3.4982e-20	Reject
topic1287	6.9052e-05	1.6593e-05	1.11561e+06	9.0104e-248	Reject
topic1288	0.00021018	0.00017546	2.24591e+06	7.2355e-18	Reject
topic1289	8.5674e-05	3.9938e-05	1.62953e+06	4.91e-110	Reject
topic1290	0.00015024	5.6301e-05	1.37228e+06	2.5148e-172	Reject
topic1291	7.0602e-05	4.1559e-05	2.09208e+06	2.3691e-32	Reject
topic1292	3.2636e-05	2.2125e-05	1.77951e+06	1.5356e-79	Reject
topic1293	0.00011845	0.00010118	2.17849e+06	5.4865e-24	Reject
topic1294	0.00012396	8.5991e-05	1.75407e+06	5.676e-85	Reject
topic1295	0.00017166	0.00010182	1.88585e+06	2.6771e-61	Reject
topic1296	7.2183e-05	7.958e-05	2.44421e+06	2.0002e-05	Reject
topic1297	5.4464e-05	4.8187e-05	2.48617e+06	0.00075173	Reject
topic1298	1.888e-05	3.6613e-06	1.31501e+06	2.3035e-187	Reject
topic1299	4.2908e-05	3.5366e-05	2.26806e+06	5.5802e-16	Reject
topic1300	0.00013683	9.2498e-05	1.74326e+06	3.451e-87	Reject
topic1301	0.00013009	0.00015696	2.41393e+06	9.6948e-07	Reject
topic1302	0.00012156	7.1692e-05	1.90297e+06	3.8792e-58	Reject
topic1303	6.7543e-06	1.7361e-05	1.93744e+06	2.2116e-50	Reject
topic1304	0.00017322	0.00011306	1.94412e+06	6.8216e-52	Reject
topic1305	0.00019723	0.00016159	2.20534e+06	1.9752e-21	Reject
topic1306	8.1382e-05	4.6563e-05	1.91727e+06	4.4521e-57	Reject
topic1307	5.2867e-05	3.4636e-05	2.18433e+06	1.2773e-22	Reject
topic1308	0.00011506	0.00013325	2.34772e+06	1.6535e-10	Reject
topic1309	0.00016338	0.00020156	2.20753e+06	2.5154e-21	Reject
topic1310	0.0001627	0.00020868	2.31826e+06	3.8856e-12	Reject
topic1311	0.00013939	7.1129e-05	1.8758e+06	3.2439e-63	Reject
topic1312	0.00015919	0.00011734	2.07209e+06	4.6207e-35	Reject
topic1313	0.00010452	6.9979e-05	2.18176e+06	1.105e-22	Reject
topic1314	0.00013173	9.1185e-05	1.73225e+06	3.8615e-89	Reject
topic1315	1.2046e-05	2.9048e-05	1.69395e+06	2.1927e-77	Reject
topic1316	9.1446e-05	9.0365e-05	2.55169e+06	0.04817	Reject
topic1317	0.0001346	0.00010641	2.30661e+06	3.0132e-13	Reject
topic1318	4.3319e-05	2.8633e-05	1.66114e+06	1.1584e-103	Reject
topic1319	0.00014556	0.00011614	2.38768e+06	5.867e-08	Reject
topic1320	2.5557e-05	3.4161e-05	2.30096e+06	4.6648e-13	Reject
topic1321	0.00013691	7.4499e-05	1.93424e+06	4.5456e-54	Reject
topic1322	0.00012073	7.4729e-05	2.15812e+06	6.2083e-26	Reject
topic1323	3.6289e-05	9.3904e-06	1.3383e+06	2.7272e-182	Reject
topic1324	0.00014524	0.00016304	2.36685e+06	2.7325e-09	Reject
topic1325	0.00011897	0.00013166	2.44376e+06	1.9151e-05	Reject
topic1326	3.1728e-05	5.3963e-05	2.03e+06	7.9277e-40	Reject
topic1327	6.2699e-08	9.7696e-08	895492	2.5491e-19	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1328	0.00010836	0.00014728	2.06441e+06	1.8922e-36	Reject
topic1329	0.00012857	8.0085e-05	1.88271e+06	6.8615e-61	Reject
topic1330	7.6483e-05	7.056e-05	2.4875e+06	0.00098325	Reject
topic1331	7.5555e-05	5.6587e-05	2.23619e+06	1.6538e-18	Reject
topic1332	0.00010177	0.00010347	2.61590e+06	0.38	Accept
topic1333	0.00012337	9.2748e-05	1.98669e+06	1.3094e-46	Reject
topic1334	4.1147e-05	4.5362e-05	2.59152e+06	0.19791	Accept
topic1335	6.6412e-05	6.3832e-05	2.59054e+06	0.19188	Accept
topic1336	6.3199e-05	7.0921e-05	2.51416e+06	0.0076554	Reject
topic1337	0.00010719	6.0724e-05	1.54976e+06	5.924e-128	Reject
topic1338	0.00013325	0.00010876	2.24035e+06	2.5047e-18	Reject
topic1339	2.2604e-05	3.2662e-06	1.78434e+06	2.8452e-75	Reject
topic1340	0.00013359	0.00011426	2.33589e+06	4.7831e-11	Reject
topic1341	5.2185e-05	3.4359e-05	2.02501e+06	1.0662e-41	Reject
topic1342	0.00012956	9.88e-05	2.01723e+06	2.4166e-42	Reject
topic1343	3.4114e-05	5.377e-06	907218	0	Reject
topic1344	0.00011618	6.923e-05	1.72476e+06	1.8624e-90	Reject
topic1345	0.00017116	0.00014329	2.17294e+06	1.5192e-24	Reject
topic1346	0.00010784	5.8033e-05	1.43044e+06	1.6678e-157	Reject
topic1347	0.00010171	3.1691e-05	1.32111e+06	3.888e-183	Reject
topic1348	8.2692e-05	1.0215e-06	786168	0	Reject
topic1349	8.4126e-05	6.7605e-05	2.30514e+06	3.3831e-13	Reject
topic1350	0.0001072	8.512e-05	2.24096e+06	6.2371e-18	Reject
topic1351	0.00014789	0.00014086	2.49545e+06	0.001788	Reject
topic1352	8.6383e-05	5.8232e-05	2.04553e+06	8.4866e-39	Reject
topic1353	6.1531e-05	1.0309e-05	1.07667e+06	2.4587e-261	Reject
topic1354	0.00011458	6.5747e-05	1.67873e+06	1.735e-99	Reject
topic1355	0.00011475	7.8994e-05	2.01475e+06	1.9722e-42	Reject
topic1356	6.0415e-05	5.054e-05	2.19235e+06	9.4973e-23	Reject
topic1357	3.314e-05	2.1221e-05	2.09393e+06	4.649e-33	Reject
topic1358	0.00014167	0.00012736	2.33175e+06	2.5083e-11	Reject
topic1359	8.7659e-05	6.4673e-05	2.14012e+06	5.089e-28	Reject
topic1360	0.00012704	0.00013	2.571e+06	0.10023	Accept
topic1361	0.0001123	7.2484e-05	2.00583e+06	2.9678e-44	Reject
topic1362	0.00012128	8.9868e-05	2.01438e+06	5.6732e-43	Reject
topic1363	0.00011162	7.4966e-05	1.87532e+06	9.6754e-64	Reject
topic1364	0.00018089	9.6079e-05	1.70622e+06	8.0221e-92	Reject
topic1365	0.0002418	0.00019953	2.10136e+06	3.4256e-32	Reject
topic1366	0.00011309	9.324e-05	2.28574e+06	1.5406e-14	Reject
topic1367	0.00013181	9.2984e-05	2.0469e+06	1.624e-37	Reject
topic1368	6.2492e-05	4.7327e-05	2.24224e+06	1e-17	Reject
topic1369	0.00012045	9.7932e-05	2.02304e+06	7.9451e-42	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1370	0.00010286	7.082e-05	1.68627e+06	4.1594e-98	Reject
topic1371	0.0001286	0.00014125	2.52911e+06	0.013404	Reject
topic1372	9.6727e-05	7.6965e-05	2.02497e+06	1.4214e-41	Reject
topic1373	5.2432e-05	1.1335e-05	1.12945e+06	7.5446e-243	Reject
topic1374	0.00018215	0.00017345	2.50049e+06	0.0019998	Reject
topic1375	6.923e-05	4.1255e-05	1.91718e+06	4.3643e-56	Reject
topic1376	9.8963e-05	0.00012614	2.23106e+06	3.9863e-19	Reject
topic1377	3.7623e-05	2.2215e-05	1.74628e+06	2.8877e-86	Reject
topic1378	0.00014705	6.5067e-05	1.77823e+06	1.1405e-78	Reject
topic1379	0.00012067	7.1257e-05	1.59008e+06	3.5744e-118	Reject
topic1380	9.9349e-05	0.00013671	2.11499e+06	1.9796e-30	Reject
topic1381	4.1805e-05	2.7552e-05	1.81746e+06	3.0404e-73	Reject
topic1382	0.00012818	0.00014691	2.34874e+06	1.9152e-10	Reject
topic1383	0.00013069	0.00012095	2.4354e+06	8.3923e-06	Reject
topic1384	0.00010904	7.8594e-05	2.09913e+06	1.4523e-32	Reject
topic1385	8.0982e-06	8.6633e-06	2.50757e+06	0.0077868	Reject
topic1386	0.00012759	0.00010252	2.24614e+06	6.2718e-18	Reject
topic1387	1.7207e-05	1.1555e-05	2.13963e+06	7.4991e-28	Reject
topic1388	0.00010056	0.00011434	2.43486e+06	7.086e-06	Reject
topic1389	0.0001004	4.3183e-05	1.49447e+06	8.4766e-140	Reject
topic1390	6.8044e-05	5.2252e-05	2.38747e+06	3.8798e-08	Reject
topic1391	9.8835e-05	6.0599e-05	2.01069e+06	6.169e-43	Reject
topic1392	0.00011926	8.6858e-05	2.1528e+06	2.8243e-26	Reject
topic1393	7.7636e-05	6.8054e-05	2.36181e+06	1.1903e-09	Reject
topic1394	0.00015255	0.00010318	1.98145e+06	5.7452e-47	Reject
topic1395	8.4555e-05	8.103e-05	2.56563e+06	0.080753	Accept
topic1396	5.4009e-05	4.2961e-05	2.39919e+06	4.3337e-07	Reject
topic1397	0.0001083	4.5449e-05	1.26076e+06	1.4497e-204	Reject
topic1398	2.4179e-05	1.1758e-05	1.72548e+06	3.8501e-90	Reject
topic1399	8.0988e-05	9.5838e-05	2.4535e+06	5.2865e-05	Reject
topic1400	0.0001526	0.00019454	2.13479e+06	2.8788e-28	Reject
topic1401	9.2337e-05	8.0758e-05	2.207e+06	2.8085e-21	Reject
topic1402	0.00012128	0.00010161	2.1932e+06	1.4426e-22	Reject
topic1403	3.9849e-05	1.8527e-05	1.49218e+06	5.8275e-141	Reject
topic1404	7.9301e-05	8.7175e-05	2.5021e+06	0.0035519	Reject
topic1405	8.4449e-05	4.0337e-05	1.62391e+06	1.0826e-110	Reject
topic1406	0.00012385	0.0001424	2.43337e+06	6.8367e-06	Reject
topic1407	0.00011895	8.4642e-05	2.01373e+06	1.483e-42	Reject
topic1408	3.5646e-05	1.611e-05	1.3935e+06	3.1898e-167	Reject
topic1409	0.00014336	7.1308e-05	1.43528e+06	1.1186e-155	Reject
topic1410	3.3077e-05	1.8479e-05	1.66185e+06	1.632e-103	Reject
topic1411	0.0001506	0.00013819	2.44508e+06	2.176e-05	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1412	0.00011498	0.00014672	2.43739e+06	1.5729e-05	Reject
topic1413	0.00015678	0.00014997	2.51038e+06	0.0039338	Reject
topic1414	0.00013237	9.7066e-05	2.15411e+06	3.0802e-26	Reject
topic1415	4.8294e-05	3.347e-05	1.8957e+06	3.6151e-60	Reject
topic1416	3.8206e-05	4.5047e-05	2.48737e+06	0.00083021	Reject
topic1417	0.00012094	8.9979e-05	1.87833e+06	6.0725e-63	Reject
topic1418	0.00017824	0.00010326	1.95348e+06	1.1535e-50	Reject
topic1419	6.3846e-05	2.6245e-05	1.35795e+06	1.2242e-176	Reject
topic1420	8.969e-05	7.2178e-05	2.3615e+06	2.055e-09	Reject
topic1421	0.00014939	0.00021788	1.84776e+06	2.524e-68	Reject
topic1422	3.9891e-05	4.0805e-05	2.62232e+06	0.47512	Accept
topic1423	0.00013513	0.00013019	2.57788e+06	0.15304	Accept
topic1424	8.1884e-05	2.8368e-05	2.05714e+06	7.0138e-37	Reject
topic1425	3.7169e-05	3.1573e-05	2.21688e+06	2.7719e-20	Reject
topic1426	1.9378e-05	1.8244e-05	2.46812e+06	0.001769	Reject
topic1427	0.0001659	0.00016761	2.60352e+06	0.28046	Accept
topic1428	2.1888e-05	1.0941e-05	1.94812e+06	3.273e-50	Reject
topic1429	0.00012095	0.00010518	2.34767e+06	1.8967e-10	Reject
topic1430	0.00010905	5.7376e-05	1.76867e+06	4.7423e-81	Reject
topic1431	9.7969e-05	7.4605e-05	2.3131e+06	1.0259e-12	Reject
topic1432	0.00013804	7.216e-05	1.45068e+06	4.2146e-152	Reject
topic1433	0.00014619	7.5011e-05	1.65112e+06	1.607e-104	Reject
topic1434	6.5514e-05	4.8598e-05	1.98556e+06	6.6409e-47	Reject
topic1435	9.1648e-05	8.6603e-05	2.55254e+06	0.042894	Reject
topic1436	0.00012463	7.0775e-05	1.66429e+06	5.8159e-102	Reject
topic1437	0.00014747	0.00012219	2.38271e+06	6.4043e-08	Reject
topic1438	6.5918e-05	1.9408e-05	1.21870e+06	4.3517e-216	Reject
topic1439	9.4979e-06	1.3177e-05	2.282e+06	8.0657e-15	Reject
topic1440	7.7781e-05	3.8841e-05	1.78138e+06	2.0641e-79	Reject
topic1441	0.0001004	9.1743e-05	2.49279e+06	0.0014759	Reject
topic1442	0.00012961	0.00010106	2.01139e+06	3.0672e-43	Reject
topic1443	0.00012624	9.5973e-05	2.09725e+06	1.478e-32	Reject
topic1444	0.00010659	8.8826e-05	2.40481e+06	4.4867e-07	Reject
topic1445	0.00010551	8.9062e-05	2.49555e+06	0.001788	Reject
topic1446	5.8873e-05	2.6713e-05	1.55572e+06	2.2502e-126	Reject
topic1447	2.6745e-05	2.9209e-06	1.24173e+06	2.1177e-209	Reject
topic1448	0.0001668	0.00017383	2.53844e+06	0.02384	Reject
topic1449	2.8785e-05	4.518e-06	1.08716e+06	1.7951e-256	Reject
topic1450	0.00011389	0.00010411	2.46023e+06	8.7596e-05	Reject
topic1451	9.7953e-05	8.2953e-05	2.28382e+06	6.3866e-15	Reject
topic1452	0.00015819	7.4406e-05	1.60089e+06	3.1277e-114	Reject
topic1453	5.4128e-06	6.9308e-06	2.60307e+06	0.28604	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1454	0.00011483	8.1853e-05	1.99482e+06	9.3427e-46	Reject
topic1455	0.00011128	4.8081e-05	1.46508e+06	2.9628e-148	Reject
topic1456	9.3834e-05	2.7988e-05	1.26234e+06	1.1564e-203	Reject
topic1457	0.00010887	6.471e-05	1.86102e+06	2.4678e-65	Reject
topic1458	0.00011784	8.2046e-05	2.0265e+06	3.045e-40	Reject
topic1459	0.00014168	0.00011662	2.20865e+06	3.1853e-21	Reject
topic1460	7.6942e-05	4.1973e-05	1.98708e+06	1.0814e-46	Reject
topic1461	0.00011178	7.2028e-05	2.00202e+06	2.855e-44	Reject
topic1462	0.0001202	8.8713e-05	1.97344e+06	1.3083e-48	Reject
topic1463	0.00012893	0.00014376	2.37418e+06	6.2172e-09	Reject
topic1464	0.00013026	0.00014624	2.48289e+06	0.00058283	Reject
topic1465	7.5827e-05	1.8762e-05	1.26304e+06	2.6143e-193	Reject
topic1466	0.00011972	8.5738e-05	1.95961e+06	1.871e-50	Reject
topic1467	0.00012283	0.00011365	2.47364e+06	0.00030674	Reject
topic1468	9.1305e-05	6.8961e-05	1.89543e+06	1.6823e-60	Reject
topic1469	0.00010801	7.5537e-05	2.12797e+06	5.2963e-29	Reject
topic1470	0.00015263	0.00014021	2.54929e+06	0.03884	Reject
topic1471	0.00010587	5.6203e-05	1.61736e+06	7.536e-113	Reject
topic1472	0.00010057	4.3877e-05	1.58725e+06	5.5672e-119	Reject
topic1473	0.00011433	6.7318e-05	1.85353e+06	4.7228e-67	Reject
topic1474	7.604e-08	1.5992e-07	415900	3.8439e-19	Reject
topic1475	0.00011213	0.00010192	2.56158e+06	0.079061	Accept
topic1476	5.1467e-05	1.3575e-05	1.05983e+06	5.1779e-267	Reject
topic1477	0.00011668	0.0001249	2.47433e+06	0.00026878	Reject
topic1478	0.00014917	8.0153e-05	1.76026e+06	1.8961e-83	Reject
topic1479	8.5296e-05	5.1287e-05	2.08874e+06	1.1804e-33	Reject
topic1480	7.8047e-05	7.372e-05	2.4824e+06	0.00051618	Reject
topic1481	8.2998e-05	6.5311e-05	2.00196e+06	2.1779e-44	Reject
topic1482	2.1032e-05	9.8033e-06	2.02062e+06	1.2389e-41	Reject
topic1483	9.9625e-05	5.1402e-05	1.80762e+06	1.6656e-74	Reject
topic1484	7.7398e-05	6.6309e-05	2.44034e+06	1.7128e-05	Reject
topic1485	0.00012987	0.00010896	2.13196e+06	8.7611e-29	Reject
topic1486	0.00013643	0.00010254	2.19743e+06	5.4868e-22	Reject
topic1487	0.00015936	0.00013408	2.32376e+06	4.626e-12	Reject
topic1488	1.2966e-05	1.8869e-05	2.16672e+06	1.1431e-23	Reject
topic1489	7.424e-05	7.7702e-05	2.56732e+06	0.082497	Accept
topic1490	0.00013391	0.00011465	2.21085e+06	1.2075e-20	Reject
topic1491	0.00016441	0.00017461	2.52726e+06	0.011259	Reject
topic1492	0.00010878	8.0675e-05	2.35328e+06	6.0013e-08	Reject
topic1493	0.00014086	0.00012752	2.48476e+06	0.00081129	Reject
topic1494	3.3447e-05	1.9825e-05	1.84812e+06	4.0317e-68	Reject
topic1495	0.00014434	7.7436e-05	2.02029e+06	3.3333e-40	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1496	0.00014317	9.095e-05	2.06172e+06	2.5621e-36	Reject
topic1497	0.00014235	6.5307e-05	1.62401e+06	1.0923e-110	Reject
topic1498	0.0001607	0.00016428	2.62716e+06	0.47819	Accept
topic1499	0.00015649	0.00010241	2.06601e+06	2.8915e-36	Reject
topic1500	9.1585e-05	5.208e-05	1.94738e+06	5.709e-52	Reject
topic1501	0.00011079	7.8952e-05	2.04482e+06	5.0099e-39	Reject
topic1502	5.9245e-05	3.5486e-05	1.60928e+06	1.267e-114	Reject
topic1503	0.00013093	0.00010952	2.19741e+06	2.8665e-22	Reject
topic1504	0.00012825	5.5483e-05	1.82683e+06	1.5787e-70	Reject
topic1505	9.4858e-05	9.8052e-05	2.60541e+06	0.29482	Accept
topic1506	0.00010812	0.00013103	2.28821e+06	1.9569e-14	Reject
topic1507	6.3975e-06	4.8886e-06	2.34206e+06	1.4937e-07	Reject
topic1508	4.3345e-05	3.8766e-05	2.2422e+06	3.5262e-18	Reject
topic1509	0.00012722	0.0001228	2.53818e+06	0.020778	Reject
topic1510	0.00012069	0.00017888	1.87096e+06	1.8634e-64	Reject
topic1511	0.00011318	8.9448e-05	2.01156e+06	1.7562e-43	Reject
topic1512	6.7482e-05	5.6068e-05	2.29703e+06	8.7165e-14	Reject
topic1513	9.3703e-05	8.7813e-05	2.52268e+06	0.0091418	Reject
topic1514	0.00015511	0.00014008	2.36308e+06	1.8859e-09	Reject
topic1515	0.00013955	7.2501e-05	1.8439e+06	9.6512e-68	Reject
topic1516	0.00013484	0.00011715	2.24986e+06	2.321e-17	Reject
topic1517	9.1551e-05	3.0493e-05	1.244e+06	1.2639e-207	Reject
topic1518	0.0001481	0.00010911	1.96647e+06	9.5601e-50	Reject
topic1519	0.0001548	0.00017619	2.3832e+06	2.2516e-08	Reject
topic1520	0.00011802	8.2554e-05	2.16034e+06	1.0453e-25	Reject
topic1521	0.00011103	9.161e-05	2.29136e+06	4.8381e-14	Reject
topic1522	8.3971e-05	4.4851e-05	1.43861e+06	2.0434e-155	Reject
topic1523	3.8713e-05	3.1677e-05	2.12802e+06	2.525e-29	Reject
topic1524	0.00013429	0.00010791	2.16108e+06	1.5864e-25	Reject
topic1525	6.156e-05	4.8097e-05	2.22639e+06	1.2566e-19	Reject
topic1526	0.00012328	0.00014385	2.31256e+06	9.4097e-13	Reject
topic1527	0.0001178	9.3825e-05	1.90889e+06	3.1329e-58	Reject
topic1528	0.00014204	0.00010852	2.20083e+06	1.7853e-21	Reject
topic1529	7.4478e-05	9.374e-05	2.2537e+06	2.6582e-17	Reject
topic1530	9.5656e-05	0.00013247	2.3678e+06	3.592e-09	Reject
topic1531	0.00020563	0.00014063	2.16211e+06	2.0615e-25	Reject
topic1532	0.00012098	9.9286e-05	2.1637e+06	1.4435e-25	Reject
topic1533	9.0778e-05	7.1306e-05	2.15723e+06	5.0308e-26	Reject
topic1534	0.00015017	0.00011235	2.08158e+06	2.9265e-34	Reject
topic1535	0.00011358	6.7135e-05	1.62817e+06	6.1316e-110	Reject
topic1536	0.00018142	0.00011634	1.94234e+06	1.0271e-52	Reject
topic1537	0.00010603	7.7958e-05	2.22043e+06	4.5392e-20	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1538	0.0001064	2.2968e-05	1.0762e+06	4.7604e-259	Reject
topic1539	5.239e-05	1.8772e-05	1.5261e+06	1.5634e-133	Reject
topic1540	0.0001263	7.7086e-05	2.06359e+06	1.1296e-36	Reject
topic1541	7.0262e-05	3.5712e-05	1.83084e+06	1.0452e-69	Reject
topic1542	1.8719e-05	1.8511e-05	2.61609e+06	0.39177	Accept
topic1543	0.0001445	0.00015134	2.55448e+06	0.049623	Reject
topic1544	0.00012577	0.00010443	2.07301e+06	1.1941e-35	Reject
topic1545	0.00030248	0.0002288	1.96996e+06	3.0278e-49	Reject
topic1546	0.000137	0.00010959	2.12287e+06	2.4671e-29	Reject
topic1547	0.00010958	8.848e-05	2.21128e+06	1.5815e-20	Reject
topic1548	0.00015586	0.00010316	1.86758e+06	9.7763e-63	Reject
topic1549	5.4833e-05	5.5506e-05	2.62387e+06	0.46934	Accept
topic1550	2.6688e-05	1.9392e-06	787854	0	Reject
topic1551	0.00011197	0.00010221	2.34421e+06	9.968e-11	Reject
topic1552	0.00010051	0.00011623	2.24412e+06	5.2111e-18	Reject
topic1553	0.00014882	0.00010369	2.24839e+06	2.1501e-17	Reject
topic1554	9.791e-05	9.6522e-05	2.54221e+06	0.028824	Reject
topic1555	5.8409e-05	4.1988e-05	2.23523e+06	1.7341e-16	Reject
topic1556	0.00016296	0.00019777	2.22446e+06	8.5005e-20	Reject
topic1557	0.00010239	9.4354e-05	2.48762e+06	0.00084662	Reject
topic1558	5.3428e-05	3.9661e-05	2.09567e+06	1.6687e-32	Reject
topic1559	0.0001334	0.00010381	2.25559e+06	5.6896e-17	Reject
topic1560	9.4558e-05	7.4485e-05	2.12849e+06	4.6568e-29	Reject
topic1561	0.00013514	9.5832e-05	1.88401e+06	6.9209e-62	Reject
topic1562	7.4804e-05	3.8731e-05	1.889e+06	1.2478e-60	Reject
topic1563	9.0007e-05	6.9871e-05	2.27113e+06	8.1337e-16	Reject
topic1564	0.00012325	0.00011608	2.51641e+06	0.0066849	Reject
topic1565	4.2047e-05	1.4773e-05	1.5969e+06	3.3206e-114	Reject
topic1566	0.00014918	0.00014789	2.61056e+06	0.35433	Accept
topic1567	0.0001343	0.00012041	2.45179e+06	4.5174e-05	Reject
topic1568	9.3584e-05	5.7591e-05	2.06462e+06	2.0087e-36	Reject
topic1569	4.5616e-05	6.1393e-05	2.27260e+06	1.0615e-15	Reject
topic1570	8.4864e-05	4.9004e-05	1.88434e+06	5.695e-62	Reject
topic1571	0.00012224	4.5543e-05	1.30944e+06	1.4379e-189	Reject
topic1572	9.1884e-05	6.4234e-05	2.26659e+06	4.3453e-16	Reject
topic1573	5.9869e-05	3.3809e-05	1.99486e+06	1.4602e-44	Reject
topic1574	4.2821e-05	1.3695e-05	1.41023e+06	3.3426e-162	Reject
topic1575	4.6254e-05	1.7834e-05	1.57666e+06	1.8768e-120	Reject
topic1576	0.00012545	5.7304e-05	1.26014e+06	1.1069e-203	Reject
topic1577	0.00011104	0.00011417	2.62053e+06	0.45916	Accept
topic1578	0.00012967	6.1573e-05	1.43322e+06	1.4047e-156	Reject
topic1579	0.0001443	0.00010924	2.0251e+06	1.4796e-41	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1580	0.0001257	0.0001104	2.46698e+06	0.00015929	Reject
topic1581	9.8405e-05	9.258e-05	2.48097e+06	0.00054868	Reject
topic1582	9.8852e-05	8.7522e-05	2.45192e+06	5.0335e-05	Reject
topic1583	0.00010152	5.1789e-05	2.12537e+06	1.683e-29	Reject
topic1584	9.4452e-05	5.9771e-05	1.9778e+06	7.4572e-48	Reject
topic1585	3.2984e-08	4.5945e-08	811416	2.9102e-05	Reject
topic1586	6.0108e-05	4.5929e-05	2.23921e+06	1.6243e-18	Reject
topic1587	9.2888e-05	8.2733e-05	2.39968e+06	1.6765e-07	Reject
topic1588	7.3348e-05	6.4015e-05	2.44763e+06	2.7495e-05	Reject
topic1589	6.6051e-05	7.3194e-05	2.51095e+06	0.0047517	Reject
topic1590	0.00013968	9.7113e-05	2.05372e+06	1.6079e-37	Reject
topic1591	9.0924e-05	6.9208e-05	2.22112e+06	1.001e-19	Reject
topic1592	1.2928e-05	1.1812e-05	2.44746e+06	2.7316e-05	Reject
topic1593	5.3725e-08	7.518e-08	1.57138e+06	6.0738e-11	Reject
topic1594	0.00012291	9.9569e-05	2.23648e+06	1.1717e-18	Reject
topic1595	0.00014831	7.7502e-05	1.40121e+06	9.367e-164	Reject
topic1596	6.4097e-05	6.4241e-05	2.627e+06	0.47679	Accept
topic1597	0.00013912	5.2288e-05	1.44439e+06	6.5341e-154	Reject
topic1598	0.00011102	6.7778e-05	1.97606e+06	4.2247e-48	Reject
topic1599	0.00015576	0.00015836	2.57465e+06	0.13093	Accept
topic1600	0.00014609	0.00011908	2.11373e+06	1.1247e-30	Reject
topic1601	4.8858e-05	4.0751e-05	2.26283e+06	2.1447e-16	Reject
topic1602	0.00013364	0.00010356	1.94637e+06	1.5532e-52	Reject
topic1603	1.75e-08	8.2163e-08	412118	9.9677e-67	Reject
topic1604	6.7622e-05	6.6638e-05	2.62032e+06	0.44944	Accept
topic1605	0.00011402	7.6446e-05	2.2057e+06	4.1702e-21	Reject
topic1606	1.7846e-05	3.8171e-06	1.582e+06	8.1879e-118	Reject
topic1607	9.3937e-05	8.6961e-05	2.48733e+06	0.0020856	Reject
topic1608	8.1419e-05	6.0972e-05	2.37945e+06	9.4962e-08	Reject
topic1609	7.3969e-05	3.8448e-05	1.39429e+06	2.2034e-166	Reject
topic1610	8.6092e-05	4.9818e-05	1.49687e+06	8.9939e-141	Reject
topic1611	8.2609e-05	5.9872e-05	1.99169e+06	6.4533e-46	Reject
topic1612	0.00012893	8.0978e-05	1.83664e+06	6.68e-70	Reject
topic1613	0.00012552	7.7908e-05	1.86611e+06	4.235e-65	Reject
topic1614	0.0001606	0.0001365	2.24661e+06	8.4264e-18	Reject
topic1615	0.0001079	8.3341e-05	2.25623e+06	1.6511e-16	Reject
topic1616	0.00011545	0.00010195	2.30625e+06	3.3944e-13	Reject
topic1617	6.1202e-05	7.3844e-05	2.48726e+06	0.0010608	Reject
topic1618	8.1293e-05	3.78e-05	1.67552e+06	3.8071e-100	Reject
topic1619	8.3849e-05	2.6753e-05	2.03568e+06	4.5834e-40	Reject
topic1620	1.9975e-05	6.7095e-06	1.67772e+06	6.8444e-100	Reject
topic1621	4.9307e-05	3.9366e-05	2.27708e+06	2.8562e-15	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1622	5.4085e-05	4.2926e-05	2.18995e+06	7.0683e-23	Reject
topic1623	0.00012542	0.00010982	2.43507e+06	1.9905e-05	Reject
topic1624	3.3011e-05	1.4745e-05	1.9868e+06	4.3692e-46	Reject
topic1625	0.00012192	9.8909e-05	2.35702e+06	8.3515e-10	Reject
topic1626	4.5609e-05	2.3709e-05	1.44869e+06	1.3135e-152	Reject
topic1627	8.1858e-05	6.8466e-05	2.50618e+06	0.0037225	Reject
topic1628	0.00012509	0.00011167	2.5121e+06	0.006307	Reject
topic1629	0.00013353	7.8187e-05	1.96227e+06	3.7631e-46	Reject
topic1630	7.5258e-05	9.0145e-05	2.35847e+06	8.7595e-10	Reject
topic1631	0.00017398	0.00015023	2.31066e+06	5.8169e-13	Reject
topic1632	9.5288e-05	8.2707e-05	2.43608e+06	1.3965e-05	Reject
topic1633	0.00012606	0.00011647	2.46553e+06	0.00018713	Reject
topic1634	0.00014375	0.00010858	2.02021e+06	3.2785e-42	Reject
topic1635	0.00016331	0.00013523	2.27259e+06	1.0402e-15	Reject
topic1636	0.00014037	7.8554e-05	1.88369e+06	5.0921e-60	Reject
topic1637	0.0001286	9.3103e-05	2.08386e+06	4.1337e-34	Reject
topic1638	1.822e-05	1.0572e-05	1.82524e+06	2.9716e-71	Reject
topic1639	0.00010148	3.339e-05	1.56386e+06	1.5764e-123	Reject
topic1640	0.00013446	0.00010362	2.01741e+06	1.3966e-42	Reject
topic1641	7.5341e-05	4.9627e-05	1.60865e+06	9.2103e-115	Reject
topic1642	0.00010199	1.4332e-05	1.15717e+06	3.9847e-236	Reject
topic1643	9.6072e-05	9.4218e-05	2.59606e+06	0.27722	Accept
topic1644	0.00012667	0.00010959	2.23938e+06	3.0854e-18	Reject
topic1645	9.153e-05	9.7725e-05	2.51530e+06	0.0061988	Reject
topic1646	4.9595e-05	3.9726e-05	2.19493e+06	2.102e-22	Reject
topic1647	3.6627e-05	3.5773e-05	2.53381e+06	0.018609	Reject
topic1648	0.00014703	0.00010963	2.08749e+06	8.4154e-34	Reject
topic1649	0.00014521	0.00012441	2.31221e+06	7.4695e-13	Reject
topic1650	0.00017819	0.00015752	2.32361e+06	4.5196e-12	Reject
topic1651	0.00012846	9.6392e-05	2.04456e+06	8.1224e-39	Reject
topic1652	0.00012595	6.0088e-05	1.63895e+06	4.3749e-107	Reject
topic1653	0.00011747	8.7545e-05	2.02478e+06	1.8151e-41	Reject
topic1654	0.00013533	0.00014248	2.54349e+06	0.029142	Reject
topic1655	0.0001448	9.148e-05	1.8265e+06	2.5042e-71	Reject
topic1656	5.4414e-05	3.5182e-05	2.02859e+06	4.1152e-41	Reject
topic1657	0.00019173	0.00013705	2.08103e+06	2.3996e-34	Reject
topic1658	0.00016476	0.00013758	2.30473e+06	2.6442e-13	Reject
topic1659	0.00013977	8.4299e-05	1.8029e+06	2.5929e-75	Reject
topic1660	0.00010048	2.048e-05	1.66274e+06	6.6422e-102	Reject
topic1661	0.00015419	0.00017933	2.27831e+06	2.4246e-15	Reject
topic1662	0.00020041	0.00013711	1.62119e+06	4.9499e-112	Reject
topic1663	4.8754e-05	2.3521e-05	2.02034e+06	2.0515e-39	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1664	4.6694e-05	4.734e-05	2.60898e+06	0.32282	Accept
topic1665	2.0179e-05	7.0294e-05	1.61634e+06	3.8474e-109	Reject
topic1666	0.00016246	0.00012474	2.24972e+06	4.9894e-17	Reject
topic1667	0.00014613	0.00010736	2.16455e+06	4.6535e-24	Reject
topic1668	9.7692e-05	7.6293e-05	2.17342e+06	1.6989e-24	Reject
topic1669	0.00010401	6.4071e-05	1.83838e+06	5.1955e-69	Reject
topic1670	7.3169e-05	7.0777e-05	2.6055e+06	0.29552	Accept
topic1671	0.00010481	7.6658e-05	1.87434e+06	6.691e-64	Reject
topic1672	4.5749e-05	2.8879e-05	2.2951e+06	1.3036e-13	Reject
topic1673	0.00013944	8.5886e-05	2.01738e+06	1.1137e-41	Reject
topic1674	9.8032e-05	8.9951e-05	2.41812e+06	1.2128e-06	Reject
topic1675	0.00014142	0.00014205	2.61458e+06	0.36879	Accept
topic1676	2.6222e-05	2.5552e-05	2.60623e+06	0.34643	Accept
topic1677	6.9533e-05	0.00013259	2.18565e+06	5.4835e-23	Reject
topic1678	4.6518e-05	2.8524e-05	2.12587e+06	3.2027e-29	Reject
topic1679	2.2067e-05	7.518e-06	1.27557e+06	4.8936e-200	Reject
topic1680	9.8699e-05	7.2115e-05	2.08454e+06	3.7798e-34	Reject
topic1681	0.00012211	9.8309e-05	2.08656e+06	4.9637e-34	Reject
topic1682	9.8332e-05	7.3011e-05	2.21909e+06	3.5146e-20	Reject
topic1683	4.014e-05	1.3168e-05	2.1566e+06	8.8621e-26	Reject
topic1684	0.00014208	8.09e-05	1.72343e+06	1.5931e-90	Reject
topic1685	0.00010201	0.00011999	2.2288e+06	2.048e-19	Reject
topic1686	0.00013633	0.00017944	2.26014e+06	1.3275e-16	Reject
topic1687	4.9881e-05	2.3821e-05	2.18441e+06	3.1688e-23	Reject
topic1688	4.6975e-05	2.8434e-05	1.89668e+06	2.0027e-59	Reject
topic1689	0.00014863	9.9549e-05	1.96605e+06	1.1502e-49	Reject
topic1690	4.7462e-05	4.8012e-05	2.53041e+06	0.025006	Reject
topic1691	0.00016421	0.00012243	1.996e+06	2.5237e-45	Reject
topic1692	2.7439e-06	1.2388e-05	1.99636e+06	3.1375e-43	Reject
topic1693	0.00012971	0.00013377	2.59691e+06	0.24122	Accept
topic1694	2.2055e-05	1.7869e-05	2.37748e+06	1.4452e-08	Reject
topic1695	1.1951e-07	1.3419e-07	2.24572e+06	0.38819	Accept
topic1696	0.00010495	7.1857e-05	2.21433e+06	1.3118e-20	Reject
topic1697	4.8057e-05	2.6013e-05	1.566e+06	5.0767e-124	Reject
topic1698	0.00014338	9.5407e-05	2.10905e+06	4.2745e-31	Reject
topic1699	0.00013013	0.00016868	2.1811e+06	1.2227e-23	Reject
topic1700	0.00015718	7.2547e-05	1.23566e+06	3.2042e-211	Reject
topic1701	0.00011245	9.6063e-05	2.29075e+06	2.1199e-14	Reject
topic1702	0.00017027	0.00011345	1.67576e+06	1.2708e-100	Reject
topic1703	9.1337e-05	5.8552e-05	1.77037e+06	3.0579e-81	Reject
topic1704	5.3937e-05	3.6372e-05	2.31926e+06	1.2127e-11	Reject
topic1705	0.00011227	7.6869e-05	2.08811e+06	7.556e-34	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1706	2.6229e-05	1.0992e-05	2.03703e+06	3.9859e-39	Reject
topic1707	0.00011993	0.00013005	2.52125e+06	0.010154	Reject
topic1708	0.00013763	7.4999e-05	1.51161e+06	8.2419e-137	Reject
topic1709	2.2499e-05	3.942e-05	1.93747e+06	9.2908e-43	Reject
topic1710	0.00018885	0.00014657	2.08961e+06	1.135e-33	Reject
topic1711	0.00016757	0.00016346	2.59312e+06	0.22346	Accept
topic1712	0.00025307	0.00017196	1.96696e+06	8.6169e-48	Reject
topic1713	1.8266e-05	8.6814e-06	1.77586e+06	4.7261e-81	Reject
topic1714	5.3242e-05	4.4012e-05	2.22083e+06	6.2424e-20	Reject
topic1715	0.00014278	0.00015533	2.51046e+06	0.0039552	Reject
topic1716	0.00013141	9.7551e-05	2.1604e+06	1.085e-25	Reject
topic1717	0.00013604	7.1925e-05	1.4654e+06	1.2353e-146	Reject
topic1718	0.00013112	0.00010754	2.36977e+06	4.6617e-09	Reject
topic1719	0.00016473	0.00012422	2.06275e+06	1.1838e-36	Reject
topic1720	8.4774e-05	5.2816e-05	2.08575e+06	5.2496e-34	Reject
topic1721	0.00013832	0.00012138	2.36069e+06	1.1872e-09	Reject
topic1722	6.8308e-05	3.4642e-05	1.74908e+06	9.0728e-85	Reject
topic1723	4.8329e-05	4.2155e-05	2.47246e+06	0.00027373	Reject
topic1724	4.0737e-05	2.4271e-05	1.55808e+06	5.306e-126	Reject
topic1725	0.00019196	0.00014897	2.08655e+06	6.5211e-34	Reject
topic1726	5.452e-05	4.4376e-05	2.28747e+06	7.408e-14	Reject
topic1727	0.00016121	0.00014465	2.36630e+06	2.5393e-09	Reject
topic1728	5.4573e-05	4.216e-05	2.23525e+06	7.4374e-19	Reject
topic1729	8.2877e-05	6.2701e-05	2.11934e+06	2.8014e-30	Reject
topic1730	0.00012051	0.00016915	2.09136e+06	2.3897e-33	Reject
topic1731	8.6695e-05	4.9117e-05	1.91974e+06	2.248e-55	Reject
topic1732	8.3616e-05	4.7127e-05	1.59020e+06	7.1972e-119	Reject
topic1733	0.00013315	0.00010853	2.1899e+06	8.8128e-23	Reject
topic1734	0.00014323	0.00011603	2.22421e+06	9.8243e-20	Reject
topic1735	3.7876e-05	2.4373e-05	2.21886e+06	3.3396e-19	Reject
topic1736	0.00014249	0.00012847	2.30916e+06	5.447e-13	Reject
topic1737	0.00011002	7.8551e-05	2.1762e+06	4.0551e-24	Reject
topic1738	0.00012248	8.4292e-05	1.74833e+06	9.8038e-86	Reject
topic1739	1.6602e-05	2.6911e-05	2.08412e+06	2.0833e-25	Reject
topic1740	0.0001261	0.00010222	2.37408e+06	8.0909e-09	Reject
topic1741	0.0001267	8.3412e-05	2.1393e+06	6.9191e-28	Reject
topic1742	0.00014378	7.4213e-05	1.47917e+06	1.4938e-144	Reject
topic1743	0.00012875	0.00010891	2.26599e+06	2.6265e-16	Reject
topic1744	0.00010223	7.4792e-05	2.10818e+06	2.6782e-31	Reject
topic1745	0.00013107	9.2782e-05	1.81672e+06	1.6096e-73	Reject
topic1746	3.6721e-05	1.6474e-05	1.63543e+06	9.1925e-109	Reject
topic1747	0.00013525	0.00013306	2.60816e+06	0.35413	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1748	3.5325e-05	1.3002e-05	1.47463e+06	2.8343e-142	Reject
topic1749	0.00013117	0.00010065	2.2581e+06	7.3176e-17	Reject
topic1750	9.6452e-05	4.1008e-05	1.4418e+06	9.3202e-154	Reject
topic1751	0.00021534	0.00018274	2.18171e+06	8.9465e-24	Reject
topic1752	0.00017186	0.00021929	2.13937e+06	4.2387e-28	Reject
topic1753	0.00010528	8.2197e-05	2.1902e+06	9.2157e-23	Reject
topic1754	8.9198e-05	5.3873e-05	2.12113e+06	1.2274e-29	Reject
topic1755	9.821e-05	0.00010251	2.61277e+06	0.36366	Accept
topic1756	0.00015337	0.00010949	1.92891e+06	1.3572e-54	Reject
topic1757	2.4217e-05	1.7939e-05	2.33185e+06	3.5864e-11	Reject
topic1758	0.00011775	8.5005e-05	1.89416e+06	2.0575e-60	Reject
topic1759	0.00010285	5.6436e-05	1.43675e+06	2.8021e-155	Reject
topic1760	3.0809e-05	1.8907e-05	1.54678e+06	7.5941e-129	Reject
topic1761	0.00013027	0.00012131	2.41858e+06	2.0573e-06	Reject
topic1762	8.5731e-05	7.0058e-05	2.25092e+06	1.5684e-17	Reject
topic1763	6.1559e-05	5.9675e-05	2.53978e+06	0.024058	Reject
topic1764	0.00011734	5.0294e-05	1.57798e+06	4.2982e-121	Reject
topic1765	8.9736e-05	8.2767e-05	2.47000e+06	0.00027012	Reject
topic1766	0.00012507	0.00012003	2.52087e+06	0.0094827	Reject
topic1767	1.1493e-05	1.6325e-05	2.31642e+06	2.0655e-12	Reject
topic1768	0.00013349	0.00011238	2.33186e+06	2.227e-11	Reject
topic1769	0.0001191	0.00015207	2.1886e+06	8.168e-23	Reject
topic1770	0.00012582	7.2404e-05	1.5616e+06	1.258e-124	Reject
topic1771	3.9635e-05	1.0958e-05	1.49056e+06	1.6147e-141	Reject
topic1772	8.0904e-05	8.068e-05	2.62054e+06	0.4295	Accept
topic1773	8.5792e-05	4.1018e-05	1.48803e+06	2.4226e-142	Reject
topic1774	3.503e-05	8.4333e-06	1.10132e+06	6.1467e-253	Reject
topic1775	0.00010929	0.00011432	2.55303e+06	0.046184	Reject
topic1776	0.00015583	0.00015876	2.62219e+06	0.43435	Accept
topic1777	5.2639e-05	2.9534e-05	1.90389e+06	1.1453e-57	Reject
topic1778	0.0001218	3.3248e-05	1.35277e+06	1.2695e-176	Reject
topic1779	0.00010901	7.5318e-05	2.10404e+06	6.946e-32	Reject
topic1780	8.9199e-05	0.00014377	103196	1.8557e-10	Reject
topic1781	0.00014646	9.8551e-05	1.96473e+06	1.9557e-49	Reject
topic1782	0.00011428	5.4916e-05	1.65186e+06	5.5652e-104	Reject
topic1783	6.4188e-05	2.5516e-05	1.37484e+06	4.9452e-172	Reject
topic1784	5.4566e-05	2.4982e-05	1.66915e+06	8.2492e-102	Reject
topic1785	7.3599e-05	3.7734e-05	2.02561e+06	4.1245e-41	Reject
topic1786	4.2087e-05	2.1728e-05	1.84377e+06	5.7771e-66	Reject
topic1787	2.9941e-05	1.5871e-05	1.62656e+06	7.3007e-111	Reject
topic1788	0.00013701	0.00013015	2.5258e+06	0.013471	Reject
topic1789	9.5753e-05	8.1995e-05	2.41633e+06	9.9657e-07	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1790	0.00013581	0.00010304	2.17617e+06	2.5541e-24	Reject
topic1791	0.00012963	0.0001161	2.46851e+06	0.00016464	Reject
topic1792	0.00017213	0.00020413	2.29213e+06	3.8807e-14	Reject
topic1793	6.9616e-05	9.848e-05	1.91994e+06	1.6099e-56	Reject
topic1794	0.00010149	8.0269e-05	2.11259e+06	1.0426e-30	Reject
topic1795	9.2268e-05	0.00015117	1.77945e+06	2.1749e-80	Reject
topic1796	7.5302e-05	2.7355e-05	1.26194e+06	1.9525e-204	Reject
topic1797	0.00017248	0.00010966	1.99498e+06	2.4978e-45	Reject
topic1798	0.00015313	0.00013399	2.22392e+06	9.2631e-20	Reject
topic1799	0.00014523	0.00012563	2.36742e+06	3.9077e-09	Reject
topic1800	2.5321e-05	1.9053e-05	2.47592e+06	0.00036944	Reject
topic1801	0.00013141	9.8438e-05	2.22986e+06	2.0261e-18	Reject
topic1802	0.00013812	7.6903e-05	1.96943e+06	6.4661e-49	Reject
topic1803	0.00014366	9.8484e-05	1.82215e+06	2.0143e-72	Reject
topic1804	0.00010079	8.2782e-05	2.29492e+06	5.1786e-14	Reject
topic1805	0.00013446	0.00011738	2.27247e+06	8.5367e-16	Reject
topic1806	0.00013807	0.00011291	2.34561e+06	1.9635e-10	Reject
topic1807	0.00011431	0.00010151	2.37593e+06	1.8167e-08	Reject
topic1808	2.2016e-05	5.0704e-05	1.95519e+06	1.2768e-34	Reject
topic1809	7.0444e-05	6.4269e-05	2.40678e+06	3.3964e-07	Reject
topic1810	0.00014534	0.00011233	2.07464e+06	3.2103e-35	Reject
topic1811	0.00012923	0.00011087	2.15385e+06	1.8067e-26	Reject
topic1812	9.1506e-05	4.9376e-05	1.59206e+06	7.0056e-118	Reject
topic1813	0.00011056	4.4265e-05	1.43855e+06	1.4617e-152	Reject
topic1814	8.114e-05	4.7114e-05	1.78866e+06	1.0513e-78	Reject
topic1815	0.00014731	0.00010862	2.15099e+06	1.9191e-26	Reject
topic1816	5.6755e-05	5.0275e-05	2.3495e+06	2.4958e-10	Reject
topic1817	6.3439e-06	2.4395e-05	1.51193e+06	7.7808e-128	Reject
topic1818	5.072e-05	3.359e-05	2.10657e+06	3.7279e-31	Reject
topic1819	0.00011664	7.7422e-05	1.7953e+06	1.6803e-77	Reject
topic1820	8.7383e-05	7.2792e-05	2.42283e+06	4.4641e-06	Reject
topic1821	4.5015e-05	3.1153e-05	2.25803e+06	8.8251e-17	Reject
topic1822	9.1123e-05	7.6887e-05	2.42893e+06	3.8511e-06	Reject
topic1823	0.00012883	7.4185e-05	1.73205e+06	1.1731e-88	Reject
topic1824	0.00013539	8.6501e-05	1.83504e+06	4.9762e-70	Reject
topic1825	0.00010259	7.0501e-05	2.10842e+06	2.1307e-31	Reject
topic1826	0.00015705	0.00016128	2.57545e+06	0.11882	Accept
topic1827	0.00015102	0.00010217	2.02376e+06	2.2977e-41	Reject
topic1828	0.00010357	0.00010537	2.59027e+06	0.19024	Accept
topic1829	2.1528e-05	1.0898e-05	1.96431e+06	4.9716e-42	Reject
topic1830	0.00011316	0.0001124	2.59167e+06	0.22831	Accept
topic1831	0.00012702	9.507e-05	2.0266e+06	3.136e-41	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1832	1.9149e-05	3.1561e-05	2.1944e+06	5.7455e-22	Reject
topic1833	0.00015343	0.0001656	2.46094e+06	8.5037e-05	Reject
topic1834	9.1235e-05	6.3067e-05	2.15814e+06	1.0166e-25	Reject
topic1835	9.8533e-05	8.9341e-05	2.50092e+06	0.0022352	Reject
topic1836	0.00012246	0.00012644	2.53752e+06	0.020049	Reject
topic1837	1.031e-05	1.9541e-05	1.96939e+06	3.2607e-43	Reject
topic1838	1.1291e-05	1.0389e-05	2.55141e+06	0.13687	Accept
topic1839	0.00020795	0.00017015	2.14283e+06	9.8784e-28	Reject
topic1840	7.5457e-05	6.2655e-05	2.34394e+06	9.582e-11	Reject
topic1841	2.19e-05	5.9402e-05	1.69912e+06	1.0589e-95	Reject
topic1842	0.00012915	0.00012727	2.57026e+06	0.092917	Accept
topic1843	0.00012693	0.00016538	2.18427e+06	2.482e-23	Reject
topic1844	0.00012313	0.00011855	2.51426e+06	0.0050679	Reject
topic1845	7.2131e-05	4.5611e-05	1.76035e+06	1.2836e-83	Reject
topic1846	9.4485e-05	6.049e-05	2.08192e+06	1.3995e-34	Reject
topic1847	3.9253e-05	2.475e-05	1.78481e+06	4.19e-79	Reject
topic1848	0.00015126	0.00012697	2.24908e+06	1.6614e-17	Reject
topic1849	4.7443e-05	1.4363e-05	1.28686e+06	9.4888e-197	Reject
topic1850	9.424e-05	4.6101e-05	1.8938e+06	9.2491e-61	Reject
topic1851	3.3978e-05	1.5254e-05	1.92e+06	7.4328e-53	Reject
topic1852	9.177e-05	0.00012683	2.03740e+06	4.3396e-40	Reject
topic1853	0.00011772	7.5703e-05	2.02351e+06	3.8823e-41	Reject
topic1854	6.9455e-05	4.3054e-05	1.81462e+06	6.8712e-74	Reject
topic1855	0.00012475	5.7642e-05	1.557e+06	4.4712e-126	Reject
topic1856	0.00015457	0.00012792	2.29396e+06	4.3278e-14	Reject
topic1857	8.4088e-05	6.4037e-05	2.25642e+06	5.4323e-17	Reject
topic1858	9.3418e-05	4.368e-05	1.64827e+06	7.3948e-106	Reject
topic1859	9.4457e-05	4.6411e-05	1.41673e+06	6.8728e-161	Reject
topic1860	5.4603e-05	9.9383e-06	1.7564e+06	1.0833e-79	Reject
topic1861	8.4731e-05	5.0084e-05	1.68197e+06	8.3209e-99	Reject
topic1862	0.00013548	7.9346e-05	1.62179e+06	2.4331e-111	Reject
topic1863	7.1746e-05	6.3763e-05	2.5169e+06	0.0084866	Reject
topic1864	0.00012731	8.8266e-05	1.98814e+06	7.0812e-46	Reject
topic1865	4.9031e-05	3.2876e-05	1.91078e+06	6.171e-58	Reject
topic1866	5.8559e-05	6.3932e-05	2.50969e+06	0.0037577	Reject
topic1867	1.9298e-05	1.2359e-05	1.78591e+06	7.152e-79	Reject
topic1868	2.8034e-05	8.6433e-06	1.26493e+06	2.3443e-203	Reject
topic1869	9.7743e-05	8.418e-05	2.39914e+06	1.3936e-07	Reject
topic1870	0.0001992	0.00013009	1.54535e+06	3.53e-129	Reject
topic1871	4.9447e-05	3.5082e-05	2.10196e+06	6.8995e-31	Reject
topic1872	0.0001264	8.3799e-05	1.75027e+06	1.119e-85	Reject
topic1873	0.00015931	0.00020106	2.28806e+06	1.3346e-14	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1874	3.7938e-05	1.9747e-05	1.4313e+06	2.7818e-157	Reject
topic1875	0.0001399	0.00014334	2.5462e+06	0.033215	Reject
topic1876	0.00015609	0.0001079	2.4382e+06	3.903e-05	Reject
topic1877	0.00014267	0.00011529	2.23036e+06	7.6722e-19	Reject
topic1878	9.0405e-05	0.00010544	2.38936e+06	6.2888e-08	Reject
topic1879	0.00011442	6.2738e-05	1.56822e+06	1.7895e-123	Reject
topic1880	0.0001252	0.00013916	2.49031e+06	0.0011302	Reject
topic1881	0.00012418	7.3476e-05	1.72558e+06	6.2423e-90	Reject
topic1882	9.7227e-05	3.0217e-05	1.87124e+06	1.8965e-62	Reject
topic1883	7.7423e-05	7.7141e-05	2.61917e+06	0.41747	Accept
topic1884	0.00013413	0.00010206	2.14457e+06	1.9395e-27	Reject
topic1885	0.00033621	0.0002604	1.9004e+06	1.0336e-59	Reject
topic1886	0.00018958	0.00014081	1.90554e+06	9.1181e-59	Reject
topic1887	8.8098e-05	9.8326e-05	2.43201e+06	5.2985e-06	Reject
topic1888	9.4838e-05	3.931e-05	1.69027e+06	2.7131e-97	Reject
topic1889	0.00011672	0.00010308	2.49828e+06	0.0055984	Reject
topic1890	0.00013134	9.7773e-05	1.98764e+06	1.2976e-46	Reject
topic1891	0.00015042	0.00013567	2.41040e+06	5.1379e-07	Reject
topic1892	7.7582e-05	0.0001061	2.202e+06	9.6792e-22	Reject
topic1893	0.0001062	6.1442e-05	2.06188e+06	5.7853e-35	Reject
topic1894	0.00014291	9.3573e-05	1.97246e+06	6.8883e-49	Reject
topic1895	0.00017972	0.00014863	2.27068e+06	8.9533e-16	Reject
topic1896	8.4635e-05	5.7217e-05	1.73717e+06	3.3425e-88	Reject
topic1897	3.3105e-05	1.9447e-05	1.59273e+06	1.7078e-118	Reject
topic1898	7.5706e-05	2.309e-05	801742	0	Reject
topic1899	5.2596e-05	2.1272e-05	1.66515e+06	8.7717e-102	Reject
topic1900	3.0871e-05	1.4089e-05	1.68865e+06	2.7887e-97	Reject
topic1901	0.00011377	9.0059e-05	2.21014e+06	1.0202e-20	Reject
topic1902	0.00023952	0.00027688	2.43004e+06	6.7424e-06	Reject
topic1903	8.0058e-05	8.3899e-05	2.50426e+06	0.0026008	Reject
topic1904	0.0001412	0.00013175	2.45429e+06	4.6528e-05	Reject
topic1905	8.9809e-05	0.00011557	2.26994e+06	5.3972e-16	Reject
topic1906	0.00024619	0.00024841	2.60132e+06	0.2729	Accept
topic1907	0.00012702	0.00012723	2.57797e+06	0.12486	Accept
topic1908	0.00014371	0.00012029	2.19058e+06	8.1088e-23	Reject
topic1909	7.0099e-05	2.6868e-05	972529	1.634e-298	Reject
topic1910	9.7043e-05	7.5753e-05	2.2177e+06	3.9304e-20	Reject
topic1911	5.9547e-05	0.00021324	1.45934e+06	7.7337e-140	Reject
topic1912	6.1291e-05	3.5298e-05	2.17852e+06	2.6321e-23	Reject
topic1913	9.824e-05	7.3117e-05	2.31418e+06	2.3606e-12	Reject
topic1914	0.00010404	9.4702e-05	2.43534e+06	7.437e-06	Reject
topic1915	8.8059e-05	8.4903e-05	2.60187e+06	0.27621	Accept

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1916	0.00010596	0.00012125	2.39503e+06	1.1144e-07	Reject
topic1917	0.00014094	9.7017e-05	2.00159e+06	1.4634e-44	Reject
topic1918	0.00014258	9.058e-05	1.69071e+06	1.0984e-96	Reject
topic1919	8.5467e-05	7.5575e-05	2.44449e+06	2.2919e-05	Reject
topic1920	0.00010637	6.7687e-05	1.77986e+06	7.9158e-80	Reject
topic1921	0.00015518	0.00012552	2.12401e+06	1.166e-29	Reject
topic1922	0.00011038	8.3488e-05	2.33451e+06	3.2614e-11	Reject
topic1923	7.6605e-05	6.867e-05	2.44301e+06	1.7805e-05	Reject
topic1924	6.1061e-05	1.058e-05	1.08622e+06	1.275e-258	Reject
topic1925	5.359e-06	4.7687e-06	2.42545e+06	0.0055845	Reject
topic1926	0.00020425	0.00024073	2.28147e+06	5.038e-15	Reject
topic1927	0.0001482	0.00014224	2.56446e+06	0.084448	Accept
topic1928	9.5769e-05	7.4047e-05	2.29691e+06	1.0247e-13	Reject
topic1929	0.00013349	0.00013858	2.45719e+06	6.0655e-05	Reject
topic1930	0.00010484	8.6672e-05	2.24214e+06	4.3712e-18	Reject
topic1931	0.00013027	8.1072e-05	1.80912e+06	3.1808e-74	Reject
topic1932	0.00012484	6.9297e-05	1.71084e+06	5.076e-93	Reject
topic1933	0.00015989	0.00011434	1.91552e+06	4.7183e-57	Reject
topic1934	0.00011974	9.5259e-05	2.02473e+06	1.3241e-41	Reject
topic1935	0.00012624	8.1497e-05	2.08628e+06	7.7842e-34	Reject
topic1936	6.2192e-05	8.1311e-05	2.25256e+06	4.8231e-17	Reject
topic1937	8.7099e-05	9.0362e-05	2.55136e+06	0.047876	Reject
topic1938	0.00013575	6.5263e-05	1.55783e+06	6.9332e-126	Reject
topic1939	0.00012272	0.00011058	2.55167e+06	0.041171	Reject
topic1940	3.5067e-05	2.221e-05	2.28426e+06	9.8937e-15	Reject
topic1941	0.00012125	0.00011279	2.57968e+06	0.15631	Accept
topic1942	6.9629e-05	6.5217e-05	2.49462e+06	0.0018289	Reject
topic1943	3.0366e-05	2.1999e-05	2.20469e+06	1.72e-21	Reject
topic1944	1.1577e-05	2.0702e-05	2.00898e+06	6.3383e-40	Reject
topic1945	0.00015234	0.00012438	2.27971e+06	4.4708e-15	Reject
topic1946	5.6514e-05	4.2778e-05	2.10046e+06	1.202e-31	Reject
topic1947	2.5498e-05	6.0245e-06	1.03799e+06	1.2188e-274	Reject
topic1948	0.00015386	0.00015453	2.61010e+06	0.33186	Accept
topic1949	7.3499e-05	5.3118e-05	2.02674e+06	3.1849e-41	Reject
topic1950	8.9294e-05	4.0184e-05	1.49274e+06	3.4951e-141	Reject
topic1951	0.00015629	0.00013897	2.47169e+06	0.00048935	Reject
topic1952	0.00015023	8.8149e-05	1.68780e+06	1.8765e-97	Reject
topic1953	0.00015572	0.00012511	2.20042e+06	8.4802e-22	Reject
topic1954	9.6457e-05	4.386e-05	1.82051e+06	2.2707e-72	Reject
topic1955	9.1242e-05	2.6056e-05	1.6971e+06	9.7984e-96	Reject
topic1956	6.282e-05	1.4559e-05	1.05331e+06	1.8755e-269	Reject
topic1957	6.2023e-06	2.7806e-05	1.64827e+06	4.3769e-32	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	EN median	FL median	u-value	p-value	H_0
topic1958	0.00020408	0.00015501	1.88522e+06	3.909e-62	Reject
topic1959	8.5876e-05	7.0646e-05	2.25229e+06	3.0018e-17	Reject
topic1960	1.5997e-06	8.9118e-06	1.67788e+06	2.7359e-64	Reject
topic1961	0.00015774	0.00011817	1.98554e+06	8.7945e-47	Reject
topic1962	6.6562e-05	4.6436e-05	1.98485e+06	6.0069e-46	Reject
topic1963	0.00010747	6.7012e-05	2.14721e+06	9.7484e-27	Reject
topic1964	7.1466e-05	8.1509e-05	2.45534e+06	5.1227e-05	Reject
topic1965	6.2461e-05	4.6536e-05	2.02035e+06	3.5084e-42	Reject
topic1966	6.4129e-05	5.2264e-05	2.27404e+06	1.1319e-15	Reject
topic1967	0.00015806	0.00011075	1.85964e+06	1.0199e-65	Reject
topic1968	0.00010582	7.6344e-05	2.14112e+06	8.3921e-28	Reject
topic1969	0.00010293	6.9275e-05	1.96985e+06	7.211e-49	Reject
topic1970	0.00018583	0.00016343	2.29363e+06	3.4613e-14	Reject
topic1971	9.283e-05	8.716e-05	2.44116e+06	1.3299e-05	Reject
topic1972	0.00011137	9.5248e-05	2.31476e+06	1.3358e-12	Reject
topic1973	6.3132e-05	4.6866e-05	2.08642e+06	4.0803e-33	Reject
topic1974	0.00015752	0.00017243	2.48703e+06	0.00094947	Reject
topic1975	2.8016e-05	0.00010521	1.59874e+06	4.9379e-75	Reject
topic1976	8.5029e-05	5.0896e-05	1.70426e+06	1.7115e-94	Reject
topic1977	0.00012079	8.5383e-05	2.1498e+06	1.4134e-26	Reject
topic1978	0.00011448	9.0594e-05	2.11888e+06	4.0151e-30	Reject
topic1979	0.00011168	6.5824e-05	1.89261e+06	2.66e-59	Reject
topic1980	4.6726e-05	3.169e-05	2.00123e+06	9.3379e-45	Reject
topic1981	9.7957e-05	7.9304e-05	2.03213e+06	1.6377e-40	Reject
topic1982	9.1961e-05	7.7854e-05	2.52103e+06	0.0089284	Reject
topic1983	4.7049e-05	5.878e-05	2.25765e+06	5.6032e-17	Reject
topic1984	0.00020754	0.00015533	2.00152e+06	7.7192e-45	Reject
topic1985	0.00010205	7.1105e-05	2.2256e+06	2.0398e-19	Reject
topic1986	1.0003e-05	7.6556e-06	2.06632e+06	1.4061e-28	Reject
topic1987	0.00013503	8.8629e-05	1.93941e+06	2.0004e-53	Reject
topic1988	8.1449e-05	6.8648e-05	2.34182e+06	1.1077e-10	Reject
topic1989	0.0001514	0.00014748	2.55897e+06	0.098484	Accept
topic1990	0.00013717	0.00011206	2.2513e+06	1.6849e-17	Reject
topic1991	7.6038e-08	2.1119e-07	578026	1.6993e-29	Reject
topic1992	0.00012389	6.5241e-05	2.17571e+06	3.6502e-23	Reject
topic1993	0.00010935	7.0551e-05	1.66499e+06	1.125e-102	Reject
topic1994	0.00013247	0.00012317	2.52541e+06	0.011449	Reject
topic1995	0.00013808	0.00012293	2.46814e+06	0.00023114	Reject
topic1996	0.00012758	0.00011738	2.4522e+06	4.732e-05	Reject
topic1997	5.1255e-05	4.6133e-05	2.49639e+06	0.0059384	Reject
topic1998	4.4144e-05	2.7293e-05	2.11704e+06	5.7813e-30	Reject
topic1999	7.5208e-05	6.0867e-05	2.32746e+06	1.5961e-11	Reject

Continued on next page

Table A1 – *Continued from previous page*

Feature	**EN median**	**FL median**	**u-value**	**p-value**	**H_0**
gender	1	1	2.35098e+06	1.6547e-10	Reject
age	23	25	1.95898e+06	2.9456e-15	Reject
neg. sent.	2	1	624502	0.0066989	Reject
hapax	739	586	1.85977e+06	2.6195e-66	Reject
dis	160	116	1.70626e+06	1.9604e-94	Reject
M1	3963	3091	2.1844e+06	1.6349e-23	Reject
M2	166660	123880	2.47331e+06	0.00024699	Reject
yule	0.10891	0.087471	1.52566e+06	5.1578e-134	Reject
sichel	0.13771	0.12848	1.80145e+06	2.1448e-76	Reject
brunet	4017.1	3126.2	2.18192e+06	9.3735e-24	Reject
honore	1026.5	1006.5	2.3479e+06	1.6979e-10	Reject

CURRICULUM VITAE

NAME: Kristopher Reese

ADDRESS: Department of Computer Science and Engineering
University of Louisville
Louisville, KY 40292

EDUCATION:

M.S., Computer Science
University of Louisville
2011

B.S., Computer Science
Hood College
2009

B.A., Music Performance
Hood College
2009

TEACHING:

University of Louisville

- Program Design in C (2009/2010)
- CIS Development Project (2013)
- Introduction to Computer Science (2013)

Elizabethtown Community and Technical College

- Introduction to Computers: Instructor (2010-2012)
- Computer Maintenance Essentials: Instructor (2010-2012)
- Advanced Computer Maintenance: Instructor (2010)
- Program Design and Development: Instructor (2011-2012)

- Introduction to Database Design: Instructor (2011)
- Introduction to JavaScript: Instructor (2011)

PROFESSIONAL EXPERIENCE:

U.S. Government Dec. 2013 - Present
Technology Researcher & Analyst

- Research in Biometrics, Identity Sciences, Materials Science, Computer Vision, and Computational Social Sciences.

University of Louisville Aug. 2012 - Dec. 2013
Enterprise Systems Developer III

- Acting as Lead Application Architect and Developer.
- Creating highly interactive Data Visualizations with thousands of data points using D3.js and other JavaScript frameworks for Healthcare analytics.
- Coding a highly secure and massive scale application for Healthcare data using Zend Framework 2.0, PHP, HTML5 and other Web Based Interface Development tools.
- Creation of an OpenID Provider and Identity Management System for user self-service with OpenID endpoints for integration into main application system.
- Extensive use of Oracle SQL, PL/SQL, MySQL, and creation of RESTful APIs for enterprise application architectures.
- Testing of applications with PHPUnit for unit tests.

Alcorn State University Feb. 2012 - Aug. 2012
Visiting Researcher

- Support for biometrics research funded under the DoD and DHS.
- Research and develop new methods to enhance work in thermal facial recognition and night vision recolorization.
- Test and integrate all algorithms and components in the MATLAB prototype system.
- Write technical reports and publish research achievements.

Yakabod, Inc. Feb. 2008 - Aug. 2009
Web Programmer

- Worked on a Highly Secure application (The Yakabox), which is used by over 20,000 users worldwide and by government agencies such as the NSA.
- Researched and Developed Server- and Client- side systems for Large-Scale Information Retrieval and Enterprise Social Networking using various web-based languages, frameworks, Web Services, and APIs.
- Automated the updating process for the Yakabox.

PUBLICATIONS

Conference Papers & Posters:

Reese, K., R. Bessette, P. Hancock. 2013. "KnowYourColors: Visual Dashboards for Blood Metrics and Healthcare Analytics." In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). pp. - . Athens, Greece. October 2013

Ouch, R., K. Reese, R. Yampolskiy. 2013. "Hybrid genetic algorithm for the maximum clique problem combining sharing and migration." In Proceedings of the 24th Midwest Artificial Intelligence and Cognitive Sciences Conference (MAICS '13). pp. - . New Albany, Indiana. April 2013

Zheng, Y., K. Reese, E. Blasch, P. McManamon. 2013. "Qualitative evaluations and comparisons of six night-vision colorization methods." In Proceedings of the International Society for Optics and Phototonics (SPIE). pp. - . Baltimore, Maryland. April 2013

Zheng, Y., A. Elmaghraby, K. Reese. 2012. "Performance Improvement of Face Recognition using Multispectral Images and Stereo Images." In Proceedings of the IEEE Symposium on Signal Processing and Information Technology (ISSPIT). pp. 280-285. Ho Chi Minh City, Vietnam. December 2012

Reese, K., R. Yampolskiy, A. Elmaghraby. 2012. "A Framework for Interactive Generation of Music for Games." In Proceedings of the 17th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES '12). pp. 131-137. Louisville, Kentucky. August 2012

Reese, K., Y. Zheng, A. Elmaghraby. 2012. "A Comparison of Face Detection Algorithms in the Visible and Thermal Spectrums." In Proceedings of the International Conference on Advances in Computer Science and Application (CSA '12). pp. 49-53. Amsterdam, Netherlands. June 2012

Reese, K., A. Salem, G. Dimitoglou. 2009. "Gaming Concepts in Accessible HCI for Bare-Hand Computer Interaction." In Proceedings of the 14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES '09). pp. 40-46. Louisville, KY. August 2009

Reese, K., A. Salem, G. Dimitoglou. 2008. "Using Standard Deviation in Signal Strength Detection to Determine Jamming in Wireless Networks." In Proceedings of the 21st International Conference on Computer Applications in Industry and Engineering (CAINE '08). pp. 250-254. Honolulu, HI. December 2008

Reese, K., G. Dimitoglou. 2008. "A Survey of Path Planning Algorithms for Autonomous Robotics." In Proceedings of the Consortium for Computer Sciences in Colleges, Eastern Conference 2008 (CCSCE '08). Frederick, MD. October 2008

Journal Papers:

Reese, K., M. Shields, (name omitted), (name omitted), A. Elmaghraby "Computational Behavioral Analytics: Emotion Analysis in Foreign Languages and Effects of Translations," Journal of IC Research & Development (submitted), 2020.

Reese, K., M. Shields, (name omitted), D. Woodard, A. Elmaghraby, "Computational Behavioral Analytics: The Effects of Translations on Estimation of Personality Traits," Journal of IC Research & Development (submitted), 2020.

Reese, K., A. Salem. 2009. "A Survey on Jamming Avoidance in Wireless Ad-Hoc Sensory Networks." Journal of Computing Sciences in Colleges (CCSCE '08). Vol. 24. Iss. 3. pp. 93-98. Frederick, MD. January 2009

Theses:

Reese, K., 2011. "Computationally Generated Music using Reinforcement Learning." University of Louisville (UL). Louisville, KY. May 2011

PRESENTATIONS:

Florida Institute for CyberSecurity Conference, February 28, 2018. "The Future of Identity Science." University of Florida, Gainesville, FL.

Joint DoD and DHS workshop on Image Analysis II, June 8, 2012. "A Technical Survey of Facial Recognition Techniques - Past Achievements and Future Applications." Alcorn State University, Alcorn, MS.

Doctoral Seminar, March 27, 2011. "Generative Chord Progressions using Reinforcement Learning." University of Louisville, Louisville, KY.

CONFERENCE REVIEWER:

- 2020 IEEE/IAPR International Joint Conference on Biometrics (IJCB)
- 2019 IEEE Biometrics: Theory, Applications, and Systems (BTAS)

AWARDS:

- Computer Science Undergraduate Achievement Award (2009), Hood College
- Oracle Open World Fusion Middleware Innovation Award (2012), Oracle

- Exceptional Performance Awards from employers for research related efforts (2016, 2016, 2016, 2017, 2017, 2018, 2019)

GRANTS: Summer Research Institute
Co-Primary Investigator
Hood College
Frederick, MD, 2008

PROFESSIONAL ASSOCIATIONS:

Association for Computing Machinery (2007 - 2017)
Institute for Electrical and Electronics Engineers (2008 - Present)
American Mathematical Society (2010 - 2015)
Golden Key International Honor Society
International Neural Network Society (2015)