A CORPUS STUDY OF ETHNIC SLURS AND DEROGATORY LANGUAGE ACROSS REDDIT AND YOUTUBE WITH SENTIMENT CONSIDERED

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Linguistics

---

by

Hasan K. Autman

Spring 2016

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Hasan K. Autman:

A Corpus Study of Ethnic Slurs and Derogatory Language across Reddit and

YouTube with Sentiment Considered

Robert Malouf, Chair
Department of Linguistics and Asian/Middle Eastern Languages

Eniko Csomay
Department of Linguistics and Asian/Middle Eastern Languages

Adisa Alkebulan
Department of Africana Studies

4/13/2016
Approval Date

DEDICATION

I dedicate this thesis to Yoko Martinez, who helped me transform from a boy to a man, to my sister Valita Jones, who helped me transform from a warehouse worker to a graduate student, to my best friend Caitlin Mueller, who supported me during all my transformations, and to Stan Lee, who through his imagination, transformed us all.

ABSTRACT OF THE THESIS

A Corpus Study of Ethnic Slurs and Derogatory Language across
Reddit and YouTube with Sentiment Considered
by
Hasan K. Autman
Master of Arts in Linguistics
San Diego State University, 2016

This study sought to create a specialized corpus to determine the frequency of an ethnic slur and animal referents for African Americans, as well as derogatory language aimed at law enforcement agents found on Reddit and YouTube; specifically, when the language inspiring event is ethnically charged and is mirrored on both platforms. Comments inspired by the deaths of unarmed suspects after interactions with law enforcement agents were chosen, and a corpus of 1,883,703 words was created using Enthought Canopy, Python Reddit API Wrapper, and YouTube Comment Scrapper. Antconc was then used to determine the frequency of the target language and next, the online platform, ethnicity of the deceased, gender of the deceased, and the ethnicity of the law enforcement agent were factored in a SPSS Poisson regression test to determine significance. The results indicated that the online platform (YouTube), the ethnicity of the deceased (African American), and the gender of the deceased (female) were predictors of a higher frequency of the ethnic slur and animal referents for African Americans. However, the frequency of derogatory language aimed at law enforcement agents was shown to be higher when the deceased was Caucasian and the agent was Hispanic. Finally, Sentiment Analyzer determined that despite the significant frequency of the target derogatory language, individual YouTube comment sections were classified as neutral in semantic orientation; while individual Reddit comment sections, despite having fewer instances of the target derogatory language, were deemed semantically negative.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

ACKNOWLEDGEMENTS

I owe my deepest appreciation to my committee chair Dr. Robert Malouf for his support during the creation of this thesis. I am infinitely grateful for his role in this process as a Python expert, a source of useful information, a mediator between me and statistics, an intelligence standard to live up to, and one of the most affable professors I have encountered in my academic career. Dr. Robert Malouf might be known by some graduate students for teaching an extremely interesting sociolinguistic topology seminar; however, he will also go down in linguistics history as the host for the coolest Linguistics Student Association (LSA) Halloween party ever.

I would also like to thank my second reader and inspiration for this thesis, Dr. Eniko Csomay. She is the exemplar of how intelligence, expertise, and a great personality can all occupy the same space at the same time. She is also the person that introduced me to the exciting world of corpus linguistics and without her seminar class, this thesis would not have even reached the dream stage. Dr. Eniko Csomay continues to maintain her reputation as the instructor that will make you enjoy learning quantitative research methods, despite statistics being involved.

In addition, I would like to thank Dr. Adisa Alkebulan for agreeing to be my third reader; and honestly, sometimes that is enough.

Furthermore, I would like to thank Mattie Faye Autman, Robert Autman, and Armando Salazar. The first is my incredibly strong mother, who before she passed away, managed to work part-time and successfully raise seven incredibly strong children. The second is my biological father, who provided me with the genetic potential for intelligence. The third is my surrogate father, who encouraged me to be better and made it possible, in too many ways to list, for me to achieve my goals.

Likewise, I would like to acknowledge Paul Justice, who stands as my blueprint for becoming an excellent teacher. Thank you Master Linguistics Jedi Justice, from your loyal Padawan, Hasan.

Last but not least, I would like to thank my beautiful girlfriend, travel partner, and Netflix and chill partner for the last seven years, Amanda Black. I look forward to the day when we only marathon watch Star Trek: The Next Generation episodes on Netflix, without the fear of twenty page research papers due for graduate school.

CHAPTER 1

INTRODUCTION

In 1967, Dr. Martin Luther King Jr. wrote a book entitled: *Where do we go from Here: Chaos or Community,* which attempted to analyze the racial tensions between the Negro community and the White community, while offering solutions for the peaceful integration of both Negroes and Whites into one harmonious community. Opposition to Dr. Martin Luther King Jr.'s goal of peaceful integration came in the form of racist rhetoric from members of the White community and was fueled by The Civil Rights Act of 1964, the actions of racial fear mongering political leaders such as Governor George Wallace (e.g. Stand in the Schoolhouse Door), Martin Luther King Jr's increased media coverage, and the death of Negro men, women, and children after interactions with law enforcement agents (Feagin, 2013; Maddow, 2016). In fact, discourse generated by the race related events began to appear as headlines for newspapers across the United States of America, and conversations regarding racial tension were heard throughout all levels the American socioeconomic spectrum (Martindale, 1990). However, during the 1960's, few sought to analyze the discourse inspired by race related events to discern specific language used to refer to Negroes or Whites, and how that language may have contributed to an altered perception of these two groups in the eyes of the general public (Martindale, 1990). Furthermore, in his final published work before his assassination, Martin Luther King Jr. (1967) argued that certain structural changes in American society would generate a new phase of resistance and an increase in racial rhetoric from those seeking to derogate and dehumanize ethnic groups to maintain the status quo. To explain, King Jr. was referring to controversial societal changes such as the integration of Negroes and Whites into public schools, private workplaces, and service locations (e.g. restaurants); however, other racial rhetoric fueling controversial societal changes such as the advent of the internet, democratization of technology, and the ability to easily contribute to online commentary about race related events were understandably absent from King Jr's prophecy.

It is now forty nine years later, and the current American political and social climate resembles that of the 1960's, with a few exceptions due to intellectual and technological advancements (Maddow, 2016). For example, the terms Negro and White have been replaced in societal discourse with the more acceptable labels of African American and Caucasian; and likewise, the term "race", when used to categorize different human ethnic groups has also come under scrutiny, and has been disregarded in anthropological academia due to its lack of scientific value (Meneses, 1994; O'Neil, 2006). However, similar to the 1960's, America now faces ethnic tension and discourse sparked by ethnically charged events such as immigration, ethnic fear mongering political leaders (e.g. Donald Trump), and the amount of deaths of unarmed (mostly African American) citizens in the wake of interactions with law enforcement agents (Brunson, 2007; Kindy & Fisher, 2016; Maddow, 2016). Previously, this discourse would have possibly been relegated to polite dinner conversation, newspaper headlines, or to a controlled section of a local newspaper entitled: *Letters to the Editor*. However, as a result of the advent of the internet and democratization of technology, online discourse communities are inspiring enormous amounts of discourse related to the aforementioned ethically charged events. The generation of this online discourse is particularly important because anonymous unregulated members of the specific online discourse communities, as opposed to traditional regulated news media sources, often are the ones who initiate and contribute to dialogue related to the ethnically charged event. In addition, online commenters and the forums they inhabit represent a paradigm shift in the way the general public accesses and interacts with controversial ethnically charged events. Furthermore, these ethnically charged events may be found as topics for academic papers on university campuses nationwide; therefore, it is important for instructors to be aware of the kind of online discourse that may influence or motivate their students beliefs, so that the instructors may serve as an informed mediator between the complex issues related to the ethnically charged event (the topic) and the students output (the academic paper).

Ethnically charged events, specifically those related to the deaths of unarmed citizens, generate large amounts of language date; therefore, the goal of this study was to gather online usages of that language, and to categorize language events within the data by way of specialized corpus creation and analysis. As the basis for specialized corpus creation, a convenience sample of parallel Reddit and YouTube ethnically charged events, and comments inspired by those events, were selected and compiled using Python based computational linguistic tools. After corpus creation, target words were searched and the frequency of those words, dependent on online forum (Reddit; YouTube), ethnicity of the deceased (Caucasian, African American, or Hispanic), gender of the deceased (male or female), and ethnicity of the law enforcement agent (Caucasian, African American, or Hispanic), were noted as results. To explain, the target words specifically consisted of an ethnic slur for African Americans (e.g. nigger), animal referents for African Americans (e.g. monkey, chimpanzee, gorilla, and ape), a derogatory word used to refer to law enforcement agents (e.g. pigs), and derogatory collocations related to law enforcement agents (e.g. fuck cops, fucking cop, fucking cops, pussy cop, and pussy cops). Finally, all relevant data and variables were entered into statistics software and a Poisson regression was initiated to discern the significance of any given independent variables in relation to the dependent variables of online forum, ethnicity of the deceased, gender of the deceased, and ethnicity of the law enforcement agent.

After completion of the statistical analysis of each variable, it was determined that the selected comments from YouTube, in general, had a higher frequency of the ethnic slur “nigger” in comparison to Reddit, as well as a higher frequency of animal referents for African Americans and derogatory language aimed at law enforcement. It also must be noted that according to the results of the statistical analysis, the ethnicity of the deceased, the gender of the deceased, and the ethnicity of the law enforcement agent played a significant role in the frequency of the ethnic slur directed at African Americans, animal referents for African Americans, and derogatory language aimed at law enforcement agents (with YouTube, deceased African American, female gender, and Hispanic law enforcement ethnicity being determining factors).

In conjunction with a statistical analysis of frequency, a sentiment analysis was done on the corpus (divided into sections based on events and forum) to determine the sentiment (positive, neutral, or negative) of each comment section. Sentiment is a fairly new field of linguistics that uses specialized software to analyze the semantic orientation of lexical items and to discern the global semantic orientation of discourse. Surprisingly, the results of this analysis showed that some comments generated by ethnically charged events on YouTube appear to create a global semantic orientation of neutral, despite the significant frequency of ethnic slurs found within the global YouTube text. This phenomena will be addressed fully in the discussion chapter of this study. To fully understand how this study contributes to the field of linguistics, it is important to detail the purposes of corpora, previous studies related to corpora, user commentary supported online social networks, and sentiment analysis; as well as a thorough detailing of the present study which includes methodology, results, discussion, and limitations.

CHAPTER 2

PURPOSES OF CORPORA

In the past, the act of written discourse analysis was associated with devoting great lengths of time to analyzing and categorizing individual words and speech segments; however, decades of advancements in technology have culminated in efficient and readily accessible language analysis tools in the form of the online corpus. An online corpus is a collection of multiple sets of language events, in the form of texts, which can be studied for linguistic (e.g. collocations, frequency, etc.) and non-linguistic (e.g. fillers, pauses, etc.) similarities, differences, and anomalies amongst its collection of texts (Ghadessy, Henry, & Roseberry, 2001; Hunston, 2002). However, simply collecting a group of texts should not be sufficient justification for pronounce said texts to be a corpus. Ghadessy et al. (2001) states that a corpus must be assembled on the basis of external criteria that link the texts by their socio-cultural or discourse functions. For example, a corpus driven analysis of the word "Muslim" in British press between 1998 and 2009 (socio-cultural) and a corpus study of adjectives in 16th century prayers and letters (discourse functions) can yield relevant data about societal language use under specific conditions (Baker, Gabrielatos, & McEnery, 2012; Valente, 2009). Hunston (2002) also adds that a corpus should be comprised of a collection of naturally occurring written or spoken texts. However, Rey (2014) and similar researchers have contributed to expanding the definition of a corpus by positing that genre specific scripted dialogue can be analyzed to determine such occurrences as how language use is linked to possible shifts in gender roles over periods of time in popular television series (e.g. Star Trek). As technology advancements lead to the recording and storage of varying degrees of language data, this type of specialized corpus appears to be gaining popularity; however, to fully understand the nature and adaptability of specialized and other online corpus types, a brief explanation of online corpus types should be explored.

COMPARABLE/PARALLEL

Hunston (2002) lists the basic types of corpora as comparable, parallel, learner, pedagogic, historical, monitor and specialized. Each of these corpus types portrays actual language use as it relates to their specific parameters, in the form of written, spoken, scripted and non-scripted discourse. For example, a comparable corpus compares similar texts across a variety of languages (e.g. newspapers, books, etc.) in an effort to note differences and equivalences across languages. In their 2008 study, Moreno and Suarez engaged in comparable corpus analysis to show that Spanish language literary academic reviews appear to be less critical (and negative in general) than their Anglo-American counterparts. Likewise, research in parallel corpus compares texts across a variety of languages; however, parallel corpus analysis has a key component not shared by comparable corpus analysis. Parallel corpus studies are dissimilar to comparable because they compare the same document translated across various languages. Yang (2010) uses parallel corpus analysis to explore the issue of domestication of documents (translation plus removing foreignness) and foreignization of documents (translation while maintaining foreignness).

LEARNER/PEDAGOGICAL

While the goal of analyzing the aforementioned corpus types seems to be more socio-cultural, some categories of corpus analysis are designed more for academic purposes. For example, a learner corpus exists as a collection of written texts authored by learners of a second language. This corpus can then be compared and contrasted to native speaker texts, and other non-native speaker texts, to determine differences and similarities. Cotos (2014) showed that learners who engaged in instructor led learner corpus analysis of their own writings displayed an increased learning drive and greater awareness of their patterns in relation to native speaker texts. In addition to learner corpora, instructors may also create a pedagogic corpus, which is a collection of all the text the learner has been exposed to in an academic setting (e.g. course textbooks, readers, etc.). Willis and Willis (1996), who coined the phrase pedagogic corpus, suggests that this type of corpus is important because it can raise the learner’s awareness of academic words and phrases that they may have encountered in different contexts.

HISTORICAL/MONITOR

The ability to analyze academic words and phrases in different contexts is an important tool for second language learning; however, researchers need to be able to see language use and language change over various periods of time throughout history. Historical corpora provide such an opportunity since they are comprised of documents from specific periods in history that allow for noting the development of aspects of language over time. For example, Rissanen (2000) states that a well-focused historical corpus could lead to determining what factors cause language change over time (e.g. extra-linguistics, sociolinguistics, regional, etc.). In a similar fashion, monitor corpora track historical changes in language; however, unlike historical corpora, monitor corpora are comprised of contemporary language and tracks language change from the contemporary to the future. Language data is added daily, monthly, and yearly to a monitor corpus; therefore, real-time language change tracking can occur. For example, Davies (2010) used The Corpus of Contemporary American English (a 400 million word corpus) to track real-time usage and change of key words such as web and green over a twenty year timespan. Likewise, Wennemark (2011) employed the help of Brigham Young University's 100 million word corpus of Time Magazine articles to track the semantic evolution of the word gay because this particular term has undergone a significant semantic shift between the 1920's and current times. Though historical/monitor corpora appear to be the most dynamic form of corpora as it relates to language data collection over a given period of time, another type of corpora referred to as a specialized corpus, by its nature, is more dynamic and diverse.

SPECIALIZED

According to Hunston (2002), a corpus should not be judged as either good or bad in itself; instead, a corpus should be judged by its ability to suit a particular purpose. In this regard, the specialized corpus appears to be the most dynamic and diverse for engaging in language analysis. Specialized corpora are unique because technically they can encompass the parameters of many other categories of corpora while still leaving room for inventive attempts at language analysis. For example, these types of corpora can be made of language data including such diverse texts as lectures, textbooks, television scripts, newspapers, interviews, speeches, online comments, web sites, etc.) Even though specialized corpora are dynamic and diverse, it must be noted that attempting to represent a language or even a portion of a language can be a problematic process. Biber, Conrad and Reppen (1998) list a set of parameters to consider when designing a specialized corpus; for example, approaches to sampling, diversity, size, and consistency of the resource used to gather language data, are all important aspects to consider when designing a specialized corpus. Though studies utilizing specialized corpora have been used for a myriad of reasons in modern times, recent ethnically charged events (e.g. 9/11, immigration debates, killings of unarmed minorities, etc.) have steered some researchers towards analyzing language as it is used to represent the whole of certain ethnic groups.

CHAPTER 3

PREVIOUS STUDIES

The usage of specific language to categorize certain ethnic groups is not a new phenomenon; however, the ability to use specialized corpora to detect and describe this language usage (e.g. frequency, collocations, etc.) is a fairly new phenomenon that emerged out of advancements in technology (e.g. CPU processing power, computer accessibility, etc.) and language analysis tools (e.g. Wordsmith, AntConc, etc.). For example, Baker et al. (2012) used a corpus of 143 million words collected from British newspaper articles (ranging from 1998 to 2009) to ascertain collocates of the noun “Muslim”. This frequency study classified each collocate into categories such as conflict, religion, ethnic/national identity, etc. and also found that within this specific data set, Muslims are represented as easily offended, in conflict with non-Muslims, and seen as belonging to a homogenized Muslim community, despite their obvious ethnic and national differences. Previously, Akbarzadeh and Smith (2005) argued in the defense of the media by stating that texts are incapable of controlling the meaning extracted from them; however, it is fairly clear from Baker et al. (2012) that socio-cultural attitudes can have an influence on media that is often seen as non-biased. Although the aforementioned studies focus on the existence of language that categorizes certain ethnic groups within traditional media, the conversation should be expanded to include an analysis of online language usage since an enormous amount of discourse has now shifted to online formats.

When the internet was still in its infancy, many theorized that the culture of fantasy built around virtual environments would allow humanity to escape the boundaries of race and the experience of racism in our society (Daniels, 2013). It was thought that as more people were connected online, the democratization of information would occur and that online anonymity would lead to discourse divorced from the prejudice that can accompany face to face interactions. However, in reality, these theorists, at the very least, appear to have been mistaken about their predictions of online discourse and its divorce from prejudice. Contrary to theorists' predictions, it seems that the democratization of information, due to the creation of online networks, has led to an increase of racism expressed by negative language used to categorize ethnic groups online. In fact, Klein (2012) reports that modern search engines and social networks unintentionally enable peddlers of bigotry to invade online spaces of public discourse. In addition, Omernick and Sood (2013) posit that anonymity online leads to a greater usage of words and discourse that denote anger, negative ethnic categorization/ethnic slur, and words categorized as profanity in English. Of the three aforementioned types of discourse, ethnic categorization/ethnic slur type words and phrases appear in many forms, and as such, necessitates a thorough detailing.

Currently, using words and phrases that denote ethnic categorization/ethnic slur appears to be a legitimate form of discourse within online commentary forums (Klein, 2012). This discourse can appears in various forms ranging from coded language (implicit) to ethnic slurs (explicit). To explain, coded language can be used to deliver insidious, yet powerful, subtle ethnic categorizations (Hughey & Daniels, 2013). For example:

a) "Not all Islamic people are terrorists, but all terrorists are Islamic!" (Hughey, Daniels, 2013)

This type of coded language is insidious because it seeks to usher the reader toward a subconscious negative semantic entailment resulting in a correlation in the readers mind between terrorism and Islam. In addition to negative semantic entailments, Ronkin and Karn (1999) posit that parodying of stereotypical features of a target ethnic group can also be classified as coded ethnic categorization. For example, African American parodying includes:

b) "Ah am certain dat dis be right fo us"
-Voiced alveolar stop substitution for interdental fricatives – "that" to "dat", "this" to "dis"
-Habitual "be" – "be right"
-Deletion - [r] dropping – "for" to "fo" (Ronkin & Karn, 1999)

Ronkin and Karn (1999) classifies some of the core features of African American parody such as the hyper use of "be" and phonetic changes; however, more explicit and disturbing language in the form of ethnic slurs pervades a great deal of online comment forums.

A subsection of ethnic categorization is the ethnic slur. According to Jeshion (2013), these word types are meant to derogate or dehumanize a particular ethnic group and carry with them the speaker's intent to do so. While this definition is appropriate for some researchers, others have gone a step further and linguistically analyzed ethnic slurs to highlight their features and difference from words not meant to derogate or dehumanize. To explain, previous research appears to show that ethnic slurs are expressive word types as opposed to descriptive word types that usually carry no derogatory force or intent (Hedger, 2012; Hom, 2008). For example (adapted from Croom, 2011):

c) X is a *teacher*. (Descriptive, no derogatory force or intent)

d) X is a *nigger*. (Expressive, derogatory force and intent)

In example (c), the word *teacher* is purely used as a descriptor that carries no inherent negative meaning; however, the word *nigger* in example (d) carries derogatory force and intent beyond that of pure description. Furthermore, researchers posit that ethnic slurs as expressive words differ from descriptive words in that an ethnic slur's behavior in a sentence clause when denial is employed, carries derogatory meaning by default (Hedger, 2012; Hom, 2008; McCready, 2010). For example (adapted from Croom, 2011):

e) I went out with a teacher, but I deny expressing any feelings about him.

f) I went out with a nigger, but I deny expressing any feelings about him.

In example (e), the presence of the descriptive word *teacher* in the first clause does not affect the claim in the second clause; however, in example (f), the presence of the ethnic slur *nigger* in the first clause, definitely distorts the validity of the claim made in the second claim. Researchers claim that the aforementioned examples show that there is a difference between descriptive words and expressive words (ethnic slurs, etc.) and some researchers have even posited that such language has a negative effect on the psyche of the recipient (Bailey, 2006; Hedger, 2012; Hom, 2008; McCready, 2010). Furthermore, it has been speculated that a greater use of the aforementioned types of language is a direct result of democratization of information dispersal methods and user allowed feedback (comments) on online social networks (Lee, Kim, Cho, & Woo, 2013).

CHAPTER 4

USER COMMENTARY SUPPORTED ONLINE SOCIAL NETWORKS

According to Lee et al. (2013) participants in online social networks readily discuss and dispute current events and opinions in comment sections of social media. These comments can range from a few comments to thousands of comments, and can give voice to individuals or communities often overlooked in traditional offline discourse. Mishra and Rastogi (2012) suggests that these social media comment sections are highly biased and contain words that denote hate speech; however, these comment sections should not be used to represent a real-time public opinion poll of current events. Instead, online language data gathered can be assessed for frequency of racism expressed by negative language used to categorize ethnic groups in comment sections of social networking sites such as Reddit and YouTube.

In recent times, Reddit and YouTube have become the standard online forums for discussing current events amongst an increasing diverse online population. For example, YouTube currently has one billion registered users (unregistered viewership not known) worldwide that have the ability comment on videos being uploaded at a rate of three hundred hours of video per minute (Cheng, Dale, & Liu, 2008; Freeman & Chapman, 2007). YouTube is also localized in seventy five countries and available in sixty one languages by way of non-mobile or mobile device (50% of YouTube’s viewership is on mobile devices). On the other hand, Reddit is comprised of over seven billion pages comment threads with an average of one hundred sixty nine million unique visitors per month (Reddit, 2016; Steinbaur, 2012). These users hail from over two hundred nine different countries and at any given period approximately three million users are logged on to Reddit. Given the overwhelming amount of language data generated by Reddit and YouTube comments, it seems logical that this data should be collected into various types of specialized corpora in an effect to study online language phenomena. For example, Mills (2011) sought to uncover possible links between Reddit's placement of majority opinions on major news events and the concept of "hive mind" mentality. Mills posits that Reddit's system of "up voting" comments (placing comments on the front page) that the majority of users agree with or simply find entertaining, creates an environment where the minority (and possibly dissenting) voice is blocked out by the majority. This interesting to note within online communities because researchers argue that when a particular offline, like-minded, discourse community engages in discussion, the unconscious goal of that discussion is to strengthen the groups prevailing attitudes and opinions, as opposed to meaningful resolution aimed discussion (Milgram, 1974). While Mills (2011) highlights noteworthy issues with online discourse communities, currently, there are very few published studies that compile Reddit and YouTube language data into specialized corpora to analyze online language phenomena across platforms. Likewise, there are very few published studies which use the aforementioned collected language data to analyze the frequency of words that denote negative ethnic categorizations.

CHAPTER 5

THE CURRENT SOCIAL CLIMATE

Mills (2011) and other previous research attempted to analyze the methodology of online comment forums to uncover a lack of egalitarianism amongst commenters; however, the events that inspire said comments may be equally as interesting to note. To explain, previous research shows that advancements in technology allow for online forums to monitor the types of events users are viewing, for how long they view them, and whether or not users comment on the certain types of events more than others (Boczkowski, 2010; Boczkowski & Mitchelstein, 2012; MacGregor, 2007). In addition, some previous research appears to show that political events often receive the most attention from online commenters during a period of election in the United States, and that there appears to be a connection between what events are broadcasted by offline news agencies and the number of comments an online post receives (Boczkowski, 2010; Boczkowski & Mitchelstein, 2012; Tenenboim & Cohen, 2015). However, the majority of research shows that events related to controversy, negativity, and personalization, remain amongst the most highly commented on, despite the presence of political activity (Anderson, 2015; Ziegele, Breiner, & Quiring, 2014). For example, a survey of the Pew Research Center (2012) showed that offline and online consumers of news in the United States were drawn to controversial news events such as the killing of the unarmed African American teenager Treyvon Martin by German/Peruvian neighborhood watch coordinator George Zimmerman in 2012. According to the study, 35% of Americans closely followed events related to the shooting with the highest number of followers from the African American community (70% African American, 30% various Caucasian ethnic groups) (Pew Research Center, 2012). With the advent of democratizing technologies, users from any societal group can engage in custom online news reporting and news consumption, despite the absence of safe guards, fact checking, or responsible journalism; and these parameters have led to an increase in online comment production when that event is deemed controversial. For example, controversial law enforcement interactions with suspects has garnered an enormous amount of attention from online commenters.

The controversy surrounding, and resulting increase in online commenter discourse related to these events, appears to stem from a debatably high number of current law enforcement interactions with suspects that have resulted in homicide. According to a year-long study done by Kindy and Fisher (2016), 991 law enforcement homicides occurred in 2015 and out of those homicides, 101 involved an unarmed suspect. Officially, the Centers for Disease and Control and Prevention (CDC), the Federal Bureau of Investigation (FBI), the American Civil Liberties Union (ACLU), and state law enforcement agencies, do not have databases which centralize data related to law enforcement related deaths; therefore, Kindy and Fisher (2016) collected and centralized data from the aforementioned sources as well as various activist groups, news agencies, etc. from across the United States. Interesting, a 2016 update to their study shows that after only two months into 2016, the rate of law enforcement related homicides is at 161.

Kindy and Fisher (2016) and other likeminded research also highlights the disproportionate amount of unarmed African Americans killed in homicides related to law enforcement interactions between 2014 and 2016, and the availability of this information may further contribute to an increase in online commenter discourse (Bradner, 2014; Gabrielson, Grochowski Jones, & Sagura, 2014; Ghitis, 2014). For example, Yvette Smith, an unarmed, forty seven year old, African American mother, was fatally shot by Bastrop County Sheriff's Deputy Daniel Willis after Willis responded to a 911 emergency dispatch stating that several males were involved in a physical altercation at a residence (Abbey-Lambertz, 2015). According to a law enforcement spokesperson, Smith was shot twice as she was ordered to exit the residence in question and despites Willis' claims of Smith having a weapon, no weapon was found by investigators (Abbey-Lambertz, 2015). After scrutinizing Deputy Willis' previous and current record, a grand jury indicted Willis for murder (Abbey-Lambertz, 2015). While the aforementioned event meets the criteria of being deemed controversial, other law enforcement involved deaths of African Americans may even be noted as more controversial. For example, Sandra Bland, a twenty eight year old African American female was found hanged in her cell after an interaction with Texas State Trooper Brian T. Encinia involving a minor traffic violation, which cumulated in her arrest and proclaimed suicide by the Harris County Institute of Forensic Science (Montgomery, 2016). The parameters that qualify this law enforcement related homicide as more controversial is the fact that Sandra Bland had no history of suicide attempts or mental illness and Texas State Trooper Brian T. Encinia was indicted by a grand jury for perjury related to the paperwork detailing Sandra Bland's arrest (Montgomery, 2016). Henceforth, interactions between law enforcement agent(s) and a suspect(s), when the suspect(s) and law enforcement agent(s) do not share the same ethnicity, will be known as ethnically charged events. This labeling is done as part of the process of discerning differences in events types, while acknowledging that possible differences may exist.

While it appears clear that controversial law enforcement related homicides alone can generate a large amount of online comments, often the level of commenter discourse is increased due to perceptions and declarations of unfair treatment of minorities and ethnic bias by law enforcement during offline encounters (Brandl, Frank, Worden, & Bynum, 1994; Weitzer, 1999). To explain, research has shown that in-group bias (bias based on the ethnic group to which a person claims allegiance) and split second decisions are both linked to law enforcement related homicides where the deceased was a member of the African American community (Correll et al., 2007; James, Klinger, & Vila, 2014; Kenworthy, Barden, Diamond, & del Carmen, 2011). According to James et al. (2014), African Americans are viewed as dangerous by some ethnic groups and law enforcement members due to the media portrayal and media coverage of African American communities. In addition, Kenworthy et al. (2011) notes that in their study of shoot/no shoot role playing scenarios, ethnically Caucasian participants were more likely to choose to shoot African Americans suspects after being primed with news reports featuring predominately African American suspects. Payne (2006) supports this conclusion and posits that the choice to see African Americans as dangerous or possessing a weapon in a role playing shoot/no shoot scenario is not simply a case of bias; instead, timing plays a key element as we as in-group bias. Payne (2006) appears to show a link between decision making times allowed for shooters in the shoot/no shoot scenario and reduction in their ethnic bias. For example, if the shooter is allowed more time to think before action, a choice based less on in-group bias will occur. From the previous studies, it clear to see how controversial events and the biases that birth said events can also lead to an increase in online commenter discourse; however, very few studies have sought to classify or give commentary about the specific sentiment of comments found in online commentary forums.

CHAPTER 6

SENTIMENT IN ONLINE DISCOURSE FORUMS

Sentiment Analysis is the process of determining the attitude of a target language producer, with respect to any given topic, by extracting and analyzing lexico-grammatical features in corpora (Mukherjee & Bhattacharyya, 2013). Previous research has accomplished the task of analysis by mostly relying on the semantic orientation (positive/negative) of adjectives, nouns, verbs, adverbs, intensifiers, and negation found in the text and their degree of orientation (Mukherjee & Bhattacharyya, 2013; Pang, Lee, & Vaithyanathan, 2002; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Turney, 2002). For example (adapted from Taboada et al., 2011):

Adjective -5 to +5, Intensifier -100% to + 100%

g) Adjective = sleazy (negative degree -3)
h) Intensifier = somewhat (negative degree - 30%)
i) Adjective + Intensifier = somewhat sleazy = -3 x (100% -30%) = negative degree -2.1

Noun -5 to +5, Adverb -100% to +100%, verbs -5 to +5

j) Noun = books (positive degree +4)
k) Adverb = foolishly (negative degree -40%)
l) Verb = hate (negative degree -5)
m) Adverb + Verb + Noun = foolishly hate books = -5 + 4 (100% - 40) = -2.6

This process can be done offline by human agency or with online and software help from language toolkits.

Language toolkits are programs designed to process natural language and can extract such things as an author’s opinion, expressed as positive or negative polarity, or even militancy, when analyzing document level discourse, sentence level discourse, and aspect levels of discourse. (Loper & Bird, 2002; Mukherjee & Bhattacharyya, 2013; Pang et al., 2002; Taboada et al., 2011; Turney, 2002). To explain, analysis at the document level seeks to classify the totality of the document, while sentence level analysis requires subjectivity analysis of individual sentences, and aspect analysis seeks to identify the recipient of the opinion (Pang et al., 2002; Turney, 2002). An exemplar of document level analysis by language toolkit usage is Gawron et al. (2012) which developed a process for automatically ranking documents by degree of militancy. Gawron et al. (2012) combined elements of language toolkits with behavioral science to produce a language classifier that can be used to analyze any organized group's discourse to determine the degree of group militancy. Language toolkits are built on supervised machine learning tasks which infer a function from training data (Taboada et al., 2011). In supervised machine learning, the training data is composed of a set of training examples that familiarizes the program with n-grams (a contiguous sequence of [n] item in a given sequence of text) such as phonemes, syllables, letters, words, collocates, etc. (Pang et al, 2002; Turney, 2002). Each example in the training data consists of an input object (vector) and the output value desired (supervisory signal); and a supervised learning algorithm analyzes this pair in an effort to correctly determine labels for seen patterns and unseen future occurrences of said patterns (Gawron et al., 2012; Pang et al., 2002; Turney, 2002). Language toolkits are usually suites of open source programs built from programming languages (e.g. Python, Linux, etc.) and are trained on large corpora such as the 450 million word Corpus of Contemporary American English (COCA) and the 22 million word American National Corpus (ANC). Furthermore, these toolkits afford computational linguist the opportunity to engage with corpora in various ways to determine lexico-grammatical features of a text (Loper and Bird, 2002; Martínez-Cámara, Martín-Valdivia, Urena-López, & Montejo-Ráez, 2014; Mukherjee & Bhattacharyya, 2013). Reliability of the language toolkit is dependent of various factors such as: 1) ease of use – ease of use equals less time spent learning to operate the toolkit and more time analyzing the data; 2) consistency – the toolkit should produce consistent results after it is trained; and 3) extensibility – new modules must be able to be easily integrated into the core language toolkit (Loper & Bird, 2002). When engaging in sentiment analysis, it is also important to choose mostly subjective discourse as opposed to mostly objective discourse because subjectivity is based solely on opinion, while objectivity is based on consideration of facts and is expected to be devoid of personal feelings (Martínez-Cámara et al., 2014). Online comment forums may be a perfect candidate for sentiment analysis; especially, when such discourse is often generated by controversial and ethnically involved events, and very few studies exist which attempt to analyze online comments. Likewise, there are very few published studies which use language toolkit data to analyze the frequency of words that denote negative ethnic categorization and or ethnic slurs in online comment forums. Henceforth, events classified as law enforcement homicides involving a suspect of a different ethnicity than the law enforcement agent will be labeled as an ethnically charged events.

CHAPTER 7

THE CURRENT STUDY

The present study is motivated by the current wave of police involved deaths of unarmed citizens in the United States, and the sparse language data analysis done on Reddit and YouTube comments generated by these events. In addition, another motivating factor is the need to create a specialized corpus representative of user comments related to instances of law enforcement involved deaths of unarmed citizens in the United States. These ethnically charged events, especially those involving unarmed African Americans, have sparked multitudes of debate in online forums and should be carefully documented for their linguistics features (e.g. frequency of ethnic slurs, sentiment, etc.). Furthermore, it is important to determine what kind of language students are exposed to in online forums related to ethnically charged events that may skew their perception of ethnic groups (type of ethnic slur) and law enforcement (negative law enforcement referent). Given the fact that any of these events may become the focal point of university discussions inside and outside of the classroom, an awareness and analysis of the language related to ethnically charged event seems appropriate, and may provide instructors with quantitative data that is useful for mediating discussions centered on these ethically charged events. To address the aforementioned needs and goals, the following research questions were created:

1. Which online comment platform has a higher degree of semantically expressive language (e.g. ethnic slurs), when the inspiring event is ethnically charged, the same across platforms, and the ethnicity and gender of the deceased are considered?
2. Which online comment platform has a higher degree of animal referents for African Americans, when the inspiring event is ethnically charged, the same across platforms, and the ethnicity and gender of the deceased are considered?
3. Which online comment platform has a higher degree of derogatory language aimed at law enforcement, when the inspiring event is ethnically charged, the

same across platforms, and the ethnicity and gender of the deceased are considered?

4. Does the presence of semantically expressive language in online comment sections related to ethnically charged events change the overall semantic orientation the sections?

From these research questions, hypotheses were established to determine the effect (if any) of online forum, gender, and ethnicity on user comments, as well as semantic orientation of the discourse across platforms that stems from the same ethnically charged event. The hypotheses are as follows:

H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of semantically expressive language, when online comments are inspired by ethnically charged events.

H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of African American animal referents, when online comments are inspired by ethnically charged events.

H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of derogatory language aimed at law enforcement, when online comments are inspired by ethnically charged events.

H_0 - The frequency of semantically expressive language is not an accurate predictor of the semantic orientation of a text.

CHAPTER 8

METHODOLOGY

To address internal validity concerns, certain parameters of the study were clearly defined. To explain, as previously mentioned, some instances of unarmed suspect deaths resulting from law enforcement interactions were labeled as ethnically charged since the law enforcement agent(s) and the suspect(s) do not share the same ethnicity. However, these events were also selected and labeled ethnically charged based on four global parameters:

n) The event had a presence on Reddit and YouTube.

o) The event generated large amounts of user comments on both Reddit and YouTube (over 100,000 words per event).

p) The event received national attention from American news institutions (both online and offline).

q) At least one event contained an element of gender in relation to the deceased.

As a control and for exploration of unique variables, one incident of a police related shooting where the deceased and the officer share the same ethnicity was included (labeled Non-Ethnically Charged), as well as an incident where the deceased was in possession of a weapon (Ethnically Charged w/Weapon). The latter incident is particularly of note because while the deceased was armed, police dash camera video showed the deceased moving away from the officers involved at the time of death. During this interaction, a law enforcement agent fired sixteen bullets into the suspect within a thirteen second period of time. In addition, the latter incident was the only incident analyzed in which the officer faces indictment for murder. Finally, to account for possible effects related to the gender of the deceased, data from comments generated by a police involved death of an unarmed female was included (labeled Ethnically Charged w/Gender). This incident also has also garnered public outrage because the deceased female was in law enforcement custody at the time of death. Initially, a corpus of over 600,000 words was created by manually copying and pasting pages of user comments related to the aforementioned events on Reddit and YouTube;

however, this process proved to be exceedingly time consuming and ineffectual because the time spent collecting language data greatly outweighed the benefits of the small amount that was collected manually.

To facilitate the creation of a more precise and less time consuming corpus of online comments generated by select events, two data mining tools (tools for manipulating large amounts of data) were used to access the Application Program Interface (API) of Reddit and YouTube (Google owned). To explain, an API is a set of routines, protocols, and web tools that dictate how users access a specific online environment; and in this case, for communication with Reddit API and Google API (YouTube), Enthought Canopy and YouTube Comment Scrapper were used. To explain, Enthought Canopy is a scientific and analytic program that allows users to interact with data in a Python (open source) computer language environment. What adds to Enthought Canopy's uniqueness is its use as an interface for a fairly comprehensive set of pre-built, pre-tested, and diverse set of Python packages designed to provide iterative data analysis, visualization of that data analysis, and the ability to accept expanded tools for various needs through user/developer creation. Enthought Canopy's user/developer created Python Reddit API Wrapper (PRAW) toolkit was used to effectively gather large amounts of language data related to the select events and threads on Reddit. Enthought Canopy provides a Graphical User Interface (GUI) for Python access within Microsoft Windows 7 and a Python windowed environment for PRAW, where the user can insert a specific preconfigured string of Python language (script) meant to produce a desired effect. A specific Python script created by Dr. Robert Malouf of San Diego State University was executed within Enthought Canopy's Python window and a 1,004,875 word corpus of Reddit user comments related to select events was compiled. The specific Python script used is as follows:

r) PRAW Script for Reddit Comment Collection

```
from __future__ import print_function
import praw

THREAD_ID = '2nslre'
user_agent = "Hasan's thesis"
r = praw.Reddit(user_agent=user_agent)
```

```
7.
8.  submission = r.get_submission(submission_id=THREAD_ID)
9.  submission.replace_more_comments(limit=None, threshold=0)
10. flat_comments = praw.helpers.flatten_tree(submission.comments)
11.
12. with open('c:/thesis/'+THREAD_ID, 'w') as f:
13.     for comment in flat_comments:
14.         print(comment.body.encode('utf-8'), file=f)
15.
```

It is important to note that the Reddit thread identification must be entered exactly how it is displayed in the Uniform Resource Locator (URL) (i.e. web address) related to the target event if data generated by that event is to be successfully gathered. For Example, the aforementioned script states in command four that the URL Reddit identification number is '2nslre'. Upon completion of the extraction process, the data from the target Reddit URL's was saved in a .TXT file that includes the commenters' online identification (the real world identification remains anonymous), date of the comment, the timestamp of the comment, the unedited comment text, and any of the aforementioned data concerning replies to the original comment. The data was then opened and viewed in Microsoft Windows 7 notepad.

Similar to the usage of Enthought Canopy and PRAW, YouTube Comment Scrapper was used to gather language data from selected events found on YouTube that mirrored events on Reddit, generated the appropriate amount of user comments, and were suitable for addressing the research questions. YouTube Comment Scrapper is YouTube API interaction module written in the open source programming language Node.js, which allows users to request all comments associated with a specific YouTube video URL. Unlike Enthought Canopy and PRAW, YouTube Comment Scrapper employed a completely web based GUI and allowed extraction of the commenters' online identification (again, the real world identification remains anonymous), date of the comment, timestamp of the comment, the unedited comment text, and any of the aforementioned data concerning replies to the original comment. Also, unlike the Python toolkits, YouTube Comment Scrapper saved the extracted information in a file with a .JSON extension (opened with Windows 7 Notepad), with the option of saving in a .CSV extension (able to be accessed through Microsoft Excel). Once information was extracted and saved into a .JSON file, Microsoft Windows 7 Notepad was used to view the data.

The aforementioned tools allowed for the compiling of a 1,883,703 word corpus comprised of Reddit and YouTube user comments generated by the following mirrored events suitable for addressing the research questions:

1. Dillon Taylor (age 20) was an unarmed Caucasian male shot and killed by a Hispanic Salt Lake City, Utah Police Officer after refusing the officers verbal command to take his hands out of his pockets. Instead of complying with the officers' command, Dillon backed away with his hands in his pockets. The officer claims he thought Dillon was armed and the shooting was ruled justified by the Salt Lake City District Attorney.
   Date of Suspect's Death: 11-8-2014
   Reddit Word Count: 119,084
   YouTube Word Count: 242,484
   Total Word Count: 361568
   Designation: Ethnically Charged

2. Eric Garner (age 43) was an unarmed African American male choked to death by Caucasian New York City Police Officers during his arrest for illegally selling cigarettes on a New York street corner. The officers in question used a banned chokehold as Eric had his hands raised in surrender; despite this recorded fact, a grand jury decided not to prosecute the officers involved.
   Date of Suspect's Death: 7-17-2014
   Reddit Word Count: 398,053
   YouTube Word Count: 115,360
   Total Word Count: 513413
   Designation: Ethnically Charged

3. Laquan McDonald (age 17) was an African American male who was shot and killed by Caucasian police officer Jason Van Dyke in Chicago, Illinois after McDonald brandished a knife while police attempted to take him into custody. The police dash camera video that captured the event was withheld by the Chicago Police Department for thirteen months until The Freedom of Information Act caused its release. The police dash camera confirmed that Van Dyke shot McDonald sixteen times within thirteen seconds as McDonald walked away; and as a result, Van Dyke was indicted for first degree murder.
   Date of Suspect's Death: 10-20-2015
   Reddit Word Count: 124,088
   YouTube Word Count: 141,746
   Total Word Count: 265,834
   Designation: Ethnically Charged w/Weapon

4. Sandra Bland (age 28) was an African American female who was found dead in her cell after being taken into police custody by Hispanic State Trooper Brian Encinia in Waller County, Texas after a minor traffic stop that escalated into an

event that allegedly necessitated her arrest. Bland was found deceased in her cell three days after the initial arrest by what the county coroner claims to have been self-asphyxiation. Bland had no history of mental illness, and no reason has been given for why Bland would have chosen suicide over facing a relatively minor charge. After the release of dash camera footage, investigators determined that Encinia did not follow proper procedure during the arrest of Bland, though no charges have been filed in relation to her mysterious death.
Date of Suspect's Death: 7-13-2015
Reddit Word Count: 212,823
YouTube Word Count: 173,401
Total Word Count: 386,224
Designation: Ethnically Charged w/Gender

5. Zackary Hammond (age 19) was an unarmed Caucasian male shot and killed by Caucasian undercover narcotics agent Mark Tiller in Seneca, South Carolina. Hammond was shot twice by the undercover law enforcement agent after being pulled over on suspicion of having a known drug dealer within his vehicle. Though no evidence appears on the officer's dash camera, Tiller claims that Hammond attempted to strike him with his vehicle. Hammond's passenger was cited for having 10 grams of marijuana and the shooter, officer Tiller, was found to have followed proper police protocol.
Date of Suspect's Death: 7-16-2015
Reddit Word Count: 150,827
YouTube Word Count: 205,842
Total Word Count: 365,669
Designation: Non-Ethnically Charged)

A corpus of well over 1.8 million words yielded many unwieldy and time consuming opportunities to collect data; therefore, to add uniformity and to limit the scope of the study, certain parameters were created that this research adhered to. For example, the frequency of a select word that denoted an ethnic slur primarily aimed at African Americans was accounted for and compared within and across the target online forums. As previously stated, words of this type function to derogate or dehumanize and are closely related to the commenter's intent to do so (Croom, 2011; Hedger, 2012; Jeshion, 201). This word was chosen due to its relatively frequent use as terms to dehumanize or derogate African Americans within and across certain discourse communities. The target word accounted for was as follows:

s) Nigger

In addition to the aforementioned ethnic slur, animal referents that are also used to dehumanize or derogate African Americans were also accounted for and compared within and across the discourse. These words, and their context, were deemed important to note since some commenters may have relied on animal referents as a tactic to qualify ethnic groups during online forum speech acts. These words were also chosen due their relatively frequent use as terms to dehumanize or derogate African Americans within and across certain discourse communities. The target words accounted for were as follows:

t) African American Animal Referents - Monkey, Chimpanzee, Gorilla, Ape.

Furthermore, in response to the research question regarding the level of derogatory language aimed at law enforcement agents, animal referents for law enforcement and their collocations were accounted for, as well as profanity aimed at law enforcement agents. These words were also chosen due their relatively frequent use as terms to dehumanize or derogate law enforcement agents within and across certain discourse communities. The target words accounted for were as follows:

u) Law Enforcement Animal Referent – Pig

v) Derogatory Law Enforcement Collocations - Fuck Cops, Fucking Cop, Fucking Cops, Pussy Cop, Pussy Cops.

To discern the frequency of the aforementioned lexical items and their collocations, Antconc was used. Antconc is freeware concordance toolkit that has various other built in features that allow for language analysis. As per the functionality of a concordance tool, Antconc allowed for the production of sorted lists comprised of the target (tokens) words and their frequency. In addition, Antconc also provided collocations of the target lexical items and contextual examples of their use in authentic unedited discourse. Moreover, Antconc provided the raw numbers needed for another program Statistical Package for Social Science (SPSS), which was used to show the statistical significance of the data. To explain, after a frequency count of the specified tokens, all relevant data and variables were entered into SPSS and a Poisson regression was initiated to discern the significance of any given variables in relation to the proposed hypotheses. The variables entered into SPSS were labeled as follows:

w) Online Forum: Reddit = 1, YouTube = 2

x) Ethnicity of Deceased: African American = 5, Caucasian = 6, Hispanic = 7

y) Ethnicity of Law Enforcement: African American = 5, Caucasian = 6, Hispanic = 7

z) Gender: Male = 1, Female = 2

Finally, to account for semantic orientation, an online document-level sentiment classifier was used to discern a real number measurement of positive or negative sentiment. Sentiment Analyzer is a general purpose tool designed provide a holistic classification of text polarization (positive or negative sentiment); therefore, Sentiment Analyzer considers all the text in the sample before committing to an overall score. To ensure an accurate assessment, Sentiment Analyzer was trained with English language samples from the American National Corpus (ANC) which were comprised of over 8,000 transcripts of spoken word and writing samples from various genres. Sentiment Analyzer operates in a similar fashion as a search engine in that it uses complex data calculations, learned vocabulary, and learned syntax rules to categorize and assign atomic lexical items a polarity, as well as a classification of polarization based on how said atomic lexical items form constituents. Once the texts generated by the target events were entered into Sentiment Analyzer, a sentiment score was computed that reflects the holistic polarity of the text based on a score that ranged from -100 to 0 to +100; whereas, -100 indicated a very negative tone, 0 indicated a neutral tone, and +100 indicated a very positive.

CHAPTER 9

RESULTS

Results from the Antconc frequency count indicated that instances within the 1, 883,703 word corpus equals less than 1% for each target lexical item (i.e. ethnic slur and animal referents for African American, derogatory language aimed at law enforcement agents). However, instances of these target lexical items, in all cases, appeared to be more frequent on YouTube comment sections in comparison to Reddit, when the inspiring event is ethnically charged. In addition, the results of the Poisson regression test indicated that when these lexical items occurred, the online platform (YouTube), the ethnicity of the deceased (African American), and the gender of the deceased (female) were predictors and had a significant effect on the frequency of the expressive ethnic slur and animal referents for African Americans. Furthermore, expressive derogatory language aimed at law enforcement agents was only shown to be significant when the deceased and the agent shared the same ethnicity (Caucasian).

The following sections contain a restatement of the relevant hypothesis, the Antconc frequency results, and the results of the Poisson regression tests done in SPSS for all target lexical items; which is followed by an acceptance or rejection of the null hypothesis. The portion of this section will present the results of a sentiment analysis preformed on specific portions of the corpora.

Ethnic Slur for African Americans

Restatement of the Hypothesis:

> H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of semantically expressive language, when online comments are inspired by ethnically charged events.

Table 1 contains the relative frequency results for the ethnic slur for African Americans inspired by ethnically charged events on Reddit and YouTube. The target lexical item was nigger and the table appears to show that instances of this ethnic slur appear more frequently on YouTube in comparison to Reddit.

Table 1. Antconc Relative Frequency Results – Ethnic Slur for African Americans

Online Forum	*Dillion Taylor*	*Eric Garner*	*Laquan McDonald*	*Sandra Bland*	*Zack Hammond*
Reddit	0	.00000054	.0000265	.00000371	0
YouTube	.00000054	.00002760	.00000436	.00010470	0

In addition to the frequency analysis done with Antconc, Table 2 shows that a Poisson probability regression was run in SPSS to determine if the online platform, the ethnicity of the deceased, or the gender of the deceased affected the probability of the occurrence of the ethnic slur for African Americans.

Table 2. Model Information

Dependent Variable	*Frequency of Slur*
Probability Distribution	*Poisson*
Link Function	*Log*
Offset Variable	*Log_Total_Words*

Furthermore, Table 3 shows the total number of cases (e.g. subjects), as well as the total number that were included in the Poisson distribution analysis and the number of cases excluded from the analysis. In this case, there were 12 total cases (100%); however, 10 were included (83.3%) and 2 were not included (16.7%) due to 2 cases having one or more missing values. Similarly, Table 4 shows the number and percentages of the independent categorical variables in this Poisson distribution analysis. Online chat forum and ethnicity of the deceased were fairly balanced; however, gender of the deceased was unbalanced at male 8 (80%) and female 2 (20%). Also, it is important to note that data in Table 3 is the same for all instances of analysis of the target lexical items; henceforth, it will no longer be included in the following results. Likewise, data in Table 4 is repeated information and will not be displayed in the following results until the addition of the variable labeled as *ethnicity of the law enforcement agent* to the data set in section 3.

Table 3. Case Processing Summary

	N	*Percent*
Included	*10*	*83.3%*
Excluded	*2*	*16.7%*
Total	*12*	*100.0%*

Table 4. Categorical Variable Information

			N	*Percent*
Factor	*Online Chat Forum*	*Reddit*	*5*	*50.0%*
		YouTube	*5*	*50.0%*
		Total	*10*	*100.0%*
	Ethnicity of the Deceased	*African American*	*6*	*60.0%*
		Caucasian	*4*	*40.0%*
		Total	*10*	*100.0%*
	Gender of the Deceased	*Male*	*8*	*80.0%*
		Female	*2*	*20.0%*
		Total	*10*	*100.0%*

Table 5 displays the minimum (in this case 0) and the maximum (in this case 192) usages of the ethnic slur for African Americans (M = 26.60, SD = 60.215) and Figure 1, Figure 2, and Figure 3 show the mean relative frequency of the ethnic slur with online chat forum (Reddit; YouTube), ethnicity of the deceased (African American, Caucasian, Hispanic), and the gender of the deceased (male, female) as predictors.

Table 5. Continuous Variable Information

		N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Dependent Variable	*Frequency of Slur*	*10*	*0*	*192*	*26.60*	*60.215*
Offset	*Log_Total_Words*	*10*	*11.66*	*12.89*	*12.0717*	*.38773*

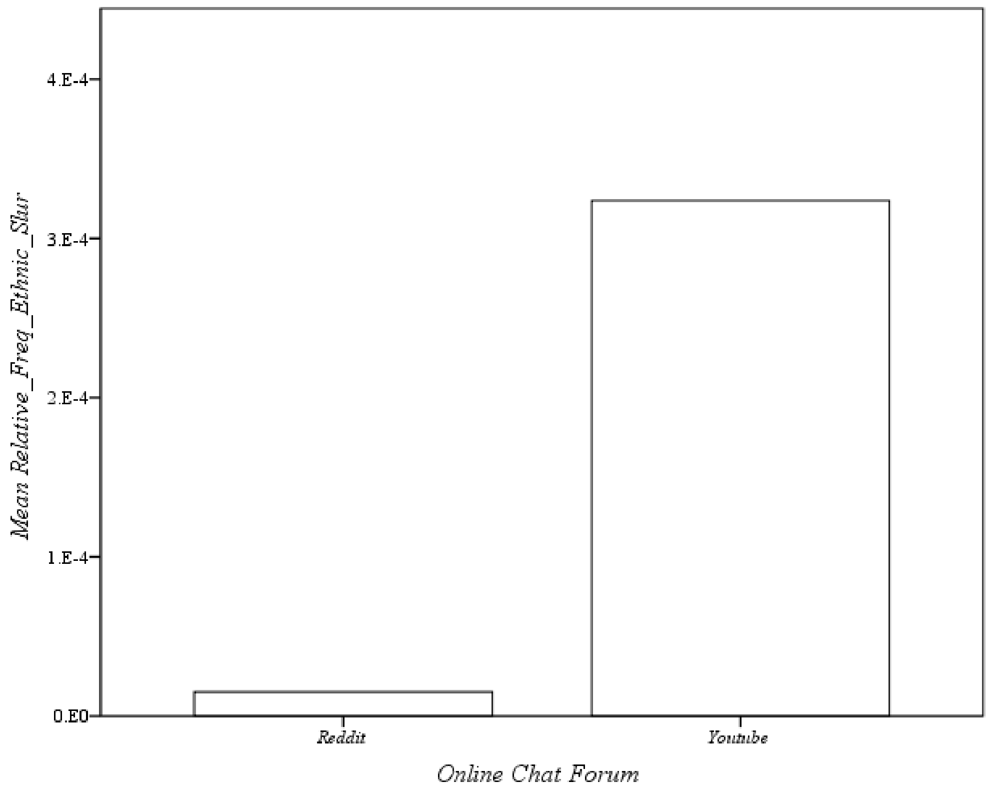

Figure 1. Mean Relative Frequency of Ethnic Slur with Online Chat Forum as a Predictor.

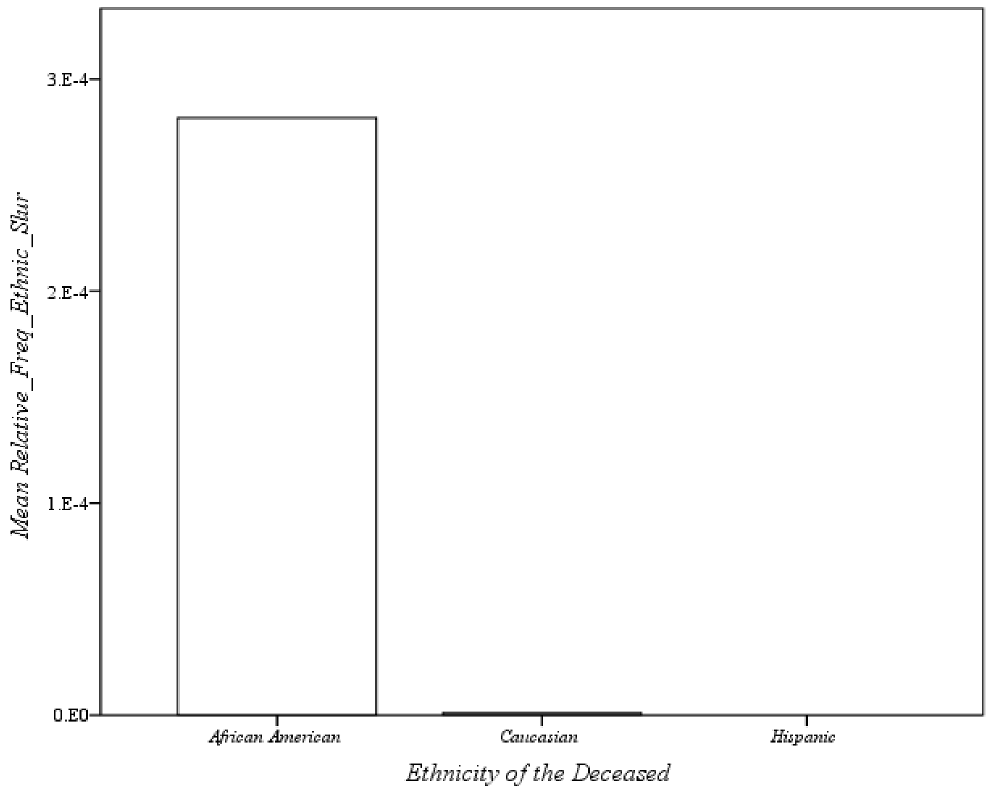

Figure 2. Mean Relative Frequency of Ethnic Slur with Ethnicity of the Deceased as a Predictor.

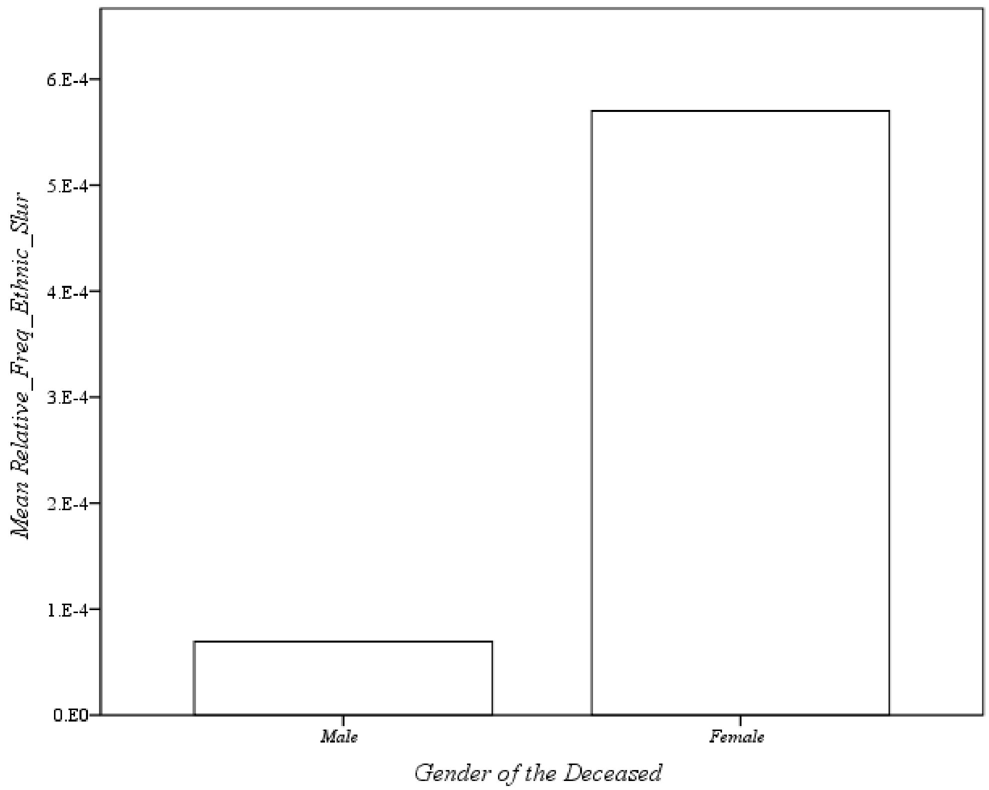

Figure 3. Mean Relative Frequency of Ethnic Slur with Gender of the Deceased as a Predictor.

Table 6 provides a summary of the significance of each independent variable and Table 7 shows the final results of the Poisson regression test. To explain, the final results of the Poisson regression test indicated that that when the probability of an ethnic slur for African Americans (i.e. nigger) was considered, the online forum Reddit was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = -3.319$), the African American ethnicity of the deceased was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = 4.696$), and the gender of the deceased (male) was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = -1.525$)

Table 6. Tests of Model Effects

Source	*Type III Wald Chi-Square*	*df*	*Sig.*
(Intercept)	*498.778*	*1*	*.000*
Online_C_Forum	*135.849*	*1*	*.000*
Ethnicity_Deceased	*21.722*	*1*	*.000*
Gender	*114.951*	*1*	*.000*

Dependent Variable: Frequency of Slur
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, offset = Log_Total_Words

Table 7. Parameter Estimates

			95% Wald Confidence Interval		*Hypothesis Test*				*95% Wald Confidence Interval for Exp(B)*	
Parameter	*B*	*Std. Error*	*Lower*	*Upper*	*Wald Chi-Square*	*df*	*Sig.*	*Exp(B)*	*Lower*	*Upper*
(Intercept)	*-11.510*	*1.0100*	*-13.489*	*-9.530*	*129.856*	*1*	*.000*	*1.003E-5*	*1.385E-6*	*7.263E-5*
[Online_C_Forum=1]	*-3.319*	*.2848*	*-3.877*	*-2.761*	*135.849*	*1*	*.000*	*.036*	*.021*	*.063*
[Online_C_Forum=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Ethnicity_Deceased=5]	*4.696*	*1.0076*	*2.721*	*6.671*	*21.722*	*1*	*.000*	*109.545*	*15.202*	*789.393*
[Ethnicity_Deceased=6]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Gender=1]	*-1.525*	*.1422*	*-1.804*	*-1.246*	*114.951*	*1*	*.000*	*.218*	*.165*	*.288*
[Gender=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
(Scale)	*1*[b]									

Dependent Variable: Frequency of Slur
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, offset = Log_Total_Words
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.
Null Hypothesis: Rejected

Animal Referents for African Americans

Restatement of the Hypothesis:

> H_0 **-** There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of African American animal referents, when online comments are inspired by ethnically charged events.

Table 8 contains the relative frequency results for African American animal referents inspired by ethnically charged events on Reddit and YouTube. The target lexical items were ape, monkey, gorilla, and chimpanzee; and the table appears to show that instances of these referents appear more frequently on YouTube in comparison to Reddit.

Table 8. Antconc Relative Frequency Results – Animal Referent for African Americans

Online Forum	*Dillion Taylor*	*Eric Garner*	*Laquan McDonald*	*Sandra Bland*	*Zack Hammond*
Reddit	0	.00000318	.00000054	.00000106	0
YouTube	.00000265	.00000477	.00000371	.00001061	.00000159

In addition to the frequency analysis done with Antconc, Table 9 shows that a Poisson probability regression was run in SPSS to determine if the online platform, the ethnicity of the deceased, or the gender of the deceased affected the probability of the occurrence of the animal referents for African Americans.

Table 9. Model Information 1

Dependent Variable	*Frequency Derogatory Ethnic Classifiers -Ape-Monkey-Gorilla-Chimpanzee*
Probability Distribution	*Poisson*
Link Function	*Log*
Offset Variable	*Log_Total_Words*

Table 10 displays the minimum (in this case 0) and the maximum (in this case 20) usages of the ethnic slur for African Americans (M = 5.30, SD = 6.001) and Figure 4, Figure 5, and Figure 6 show the mean relative frequency of the ethnic slur with online chat forum (Reddit, YouTube), ethnicity of the deceased (African American, Caucasian, Hispanic), and the gender of the deceased (male, female) as predictors.

Table 10. Continuous Variable Information 1

		N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Dependent Variable	*Frequency Derogatory Ethnic Classifiers -Ape-Monkey-Gorilla-Chimpanzee*	*10*	*0*	*20*	*5.30*	*6.001*
Offset	*Log_Total_Words*	*10*	*11.66*	*12.89*	*12.0717*	*.38773*

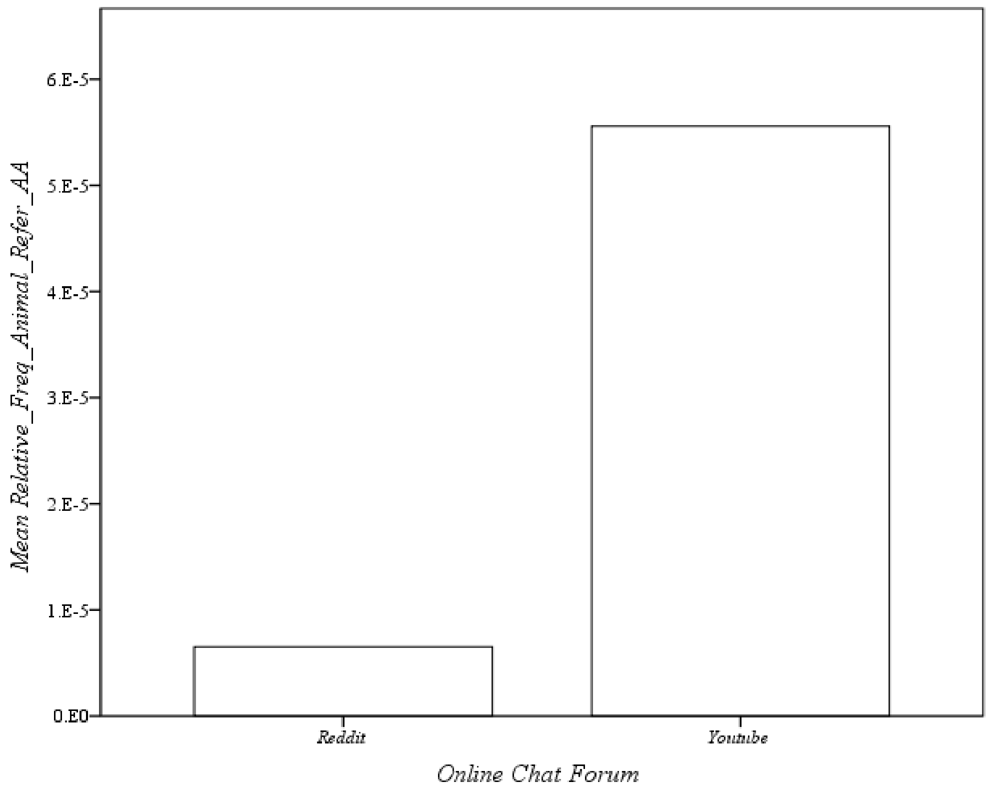

Figure 4. Mean Relative Frequency of African American Animal Referent with Online Chat Forum as a Predictor.

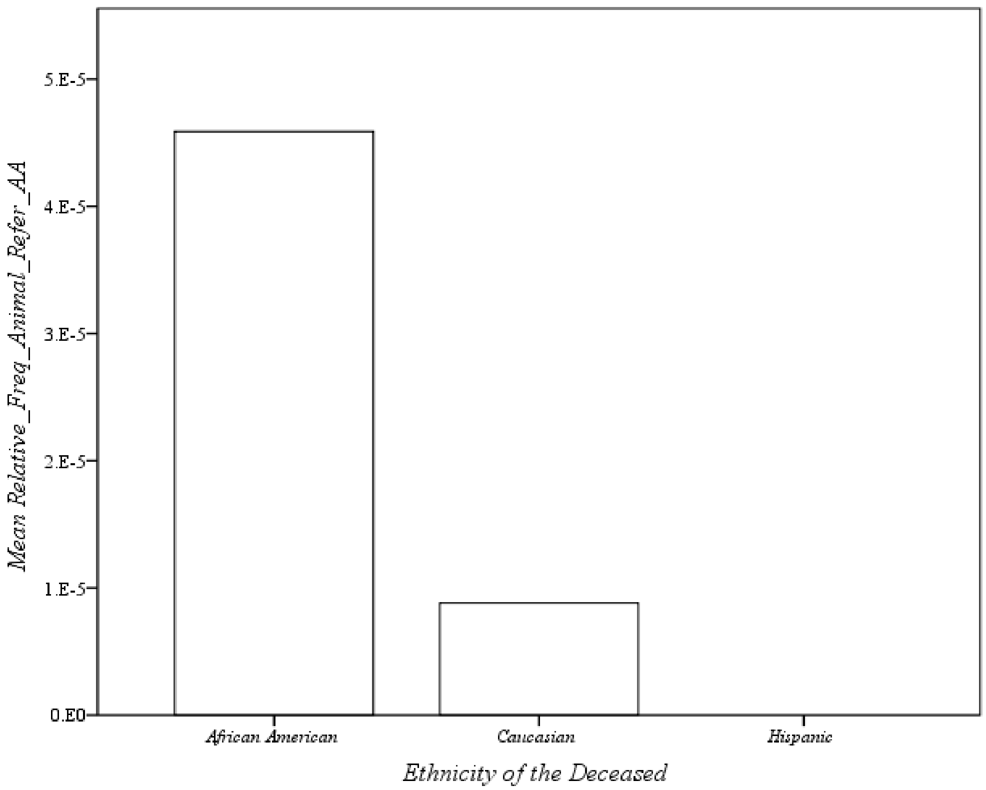

Figure 5. Mean Relative Frequency of African American Animal Referent with Ethnicity of the Deceased as a Predictor.

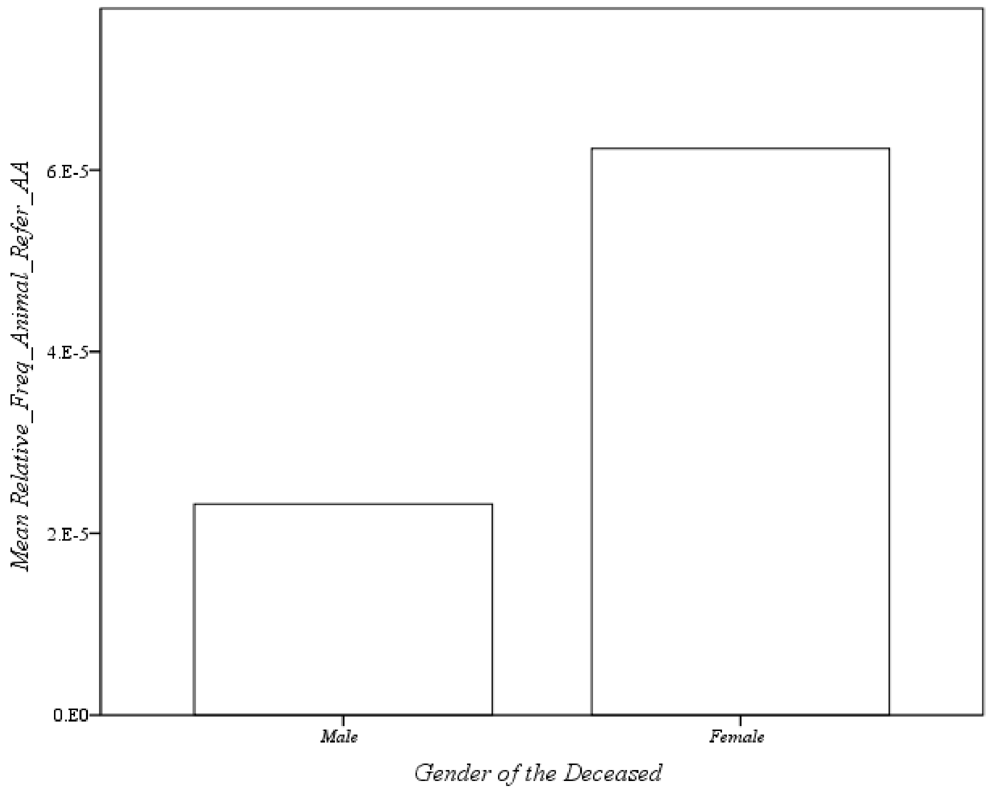

Figure 6. Mean Relative Frequency of African American Animal Referent with Gender of the Deceased as a Predictor.

Table 11 provides a summary of the significance of each independent variable and Table 12 shows the final results of the Poisson regression test. To explain, the final results of the Poisson regression test indicated that that when the probability of African American animal referents was considered, the online forum Reddit was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = -1.958$), the African American ethnicity of the deceased was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = 1.442$), and the gender of the deceased was not significant ($p = .140$; therefore, $p > .05$ and the coefficient estimate $B = -.442$)

Table 11. Tests of Model Effects 1

Source	*Type III Wald Chi-Square*	*df*	*Sig.*
(Intercept)	*1937.606*	*1*	*.000*
Online_C_Forum	*28.094*	*1*	*.000*
Ethnicity_Deceased	*12.130*	*1*	*.000*
Gender	*2.175*	*1*	*.140*

Dependent Variable: Frequency Derogatory Ethnic Classifiers -Ape-Monkey-Gorilla-Chimpanzee
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, offset = Log_Total_Words

Table 12. Parameter Estimates 1

			95% Wald Confidence Interval		*Hypothesis Test*				*95% Wald Confidence Interval for Exp(B)*	
Parameter	*B*	*Std. Error*	*Lower*	*Upper*	*Wald Chi-Square*	*df*	*Sig.*	*Exp(B)*	*Lower*	*Upper*
(Intercept)	*-10.574*	*.4625*	*-11.480*	*-9.667*	*522.604*	*1*	*.000*	*2.558E-5*	*1.033E-5*	*6.333E-5*
[Online_C_Forum=1]	*-1.958*	*.3694*	*-2.682*	*-1.234*	*28.094*	*1*	*.000*	*.141*	*.068*	*.291*
[Online_C_Forum=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Ethnicity_Deceased=5]	*1.442*	*.4139*	*.630*	*2.253*	*12.130*	*1*	*.000*	*4.227*	*1.878*	*9.515*
[Ethnicity_Deceased=6]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Gender=1]	*-.442*	*.2995*	*-1.029*	*.145*	*2.175*	*1*	*.140*	*.643*	*.357*	*1.156*
[Gender=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
(Scale)	*1*[b]									

Dependent Variable: Frequency Derogatory Ethnic Classifiers -Ape-Monkey-Gorilla-Chimpanzee
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, offset = Log_Total_Words
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.
Null Hypothesis: Rejected

DEROGATORY LANGUAGE AIMED AT LAW ENFORCEMENT

Restatment of the Hypothesis:

H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of derogatory language aimed at law enforcement, when online comments are inspired by ethnically charged events.

Table 13 contains the relative frequency results for animal referents to law enforcement inspired by ethnically charged events on Reddit and YouTube. The target lexical item were pig and pigs; and the table appears to show that instances of this type of language used to describe law enforcements is more frequently on YouTube in comparison to Reddit.

Table 13. Antconc Relative Frequency Results - Derogatory Animal Referents to Law Enforcement

Online Forum	*Dillion Taylor*	*Eric Garner*	*Laquan McDonald*	*Sandra Bland*	*Zack Hammond*
Reddit	.00004140	.00001964	.00000530	.00001486	.00001167
YouTube	.00005521	.00005361	.00001539	.00002335	.00003238

In addition to the frequency analysis done with Antconc, Table 14 shows that a Poisson probability regression was run in SPSS to determine if the online platform, the ethnicity of the deceased, the gender of the deceased, or the ethnicity of the law enforcement agent affected the probability of the occurrence of target animal referents for law enforcement.

Table 14. Model Information 2

Dependent Variable	*Frequency of Pig or Pigs Used to Refer to Law Enforcement*
Probability Distribution	*Poisson*
Link Function	*Log*
Offset Variable	*Log_Total_Words*

Table 15 shows the number and percentages of the independent categorical variables in this Poisson distribution analysis. Online chat forum and ethnicity of the deceased were fairly balanced; however, gender of the deceased was unbalanced at male 8 (80%) and female 2 (20%) and ethnicity of the law enforcement was unbalanced at Caucasian 8 (80%), Hispanic 2 (20%), and African American 0 (0%). Also, it is important to note that data in Table 15 is the same data used for this section's analysis of target lexical items (e.g. derogatory collocation of law enforcement agents); henceforth, it will not be included in that section.

Table 15. Categorical Variable Information 2

			N	*Percent*
Factor	*Online Chat Forum*	*Reddit*	*5*	*50.0%*
		YouTube	*5*	*50.0%*
		Total	*10*	*100.0%*
	Ethnicity of the Deceased	*African American*	*6*	*60.0%*
		Caucasian	*4*	*40.0%*
		Total	*10*	*100.0%*
	Gender of the Deceased	*Male*	*8*	*80.0%*
		Female	*2*	*20.0%*
		Total	*10*	*100.0%*
	Ethnicity of the Law Enforcement Officer	*Caucasian*	*8*	*80.0%*
		Hispanic	*2*	*20.0%*
		Total	*10*	*100.0%*

Table 16 displays the minimum (in this case 10) and the maximum (in this case 104) usages of the law enforcement animal referents (M = 51.40, SD = 33.194) and Figure 7, Figure 8, Figure 9, and Figure 10 show the mean relative frequency of law enforcement animal referents with online chat forum (Reddit; YouTube), ethnicity of the deceased (African American, Caucasian, Hispanic), the gender of the deceased (male, female), and ethnicity of the law enforcement agent (African American, Caucasian, Hispanic) as predictors.

Table 16. Continuous Variable Information 2

		N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Dependent Variable	*Frequency of Pig or Pigs Used to Refer to Law Enforcement*	*10*	*10*	*104*	*51.40*	*33.194*
Offset	*Log_Total_Words*	*10*	*11.66*	*12.89*	*12.0717*	*.38773*

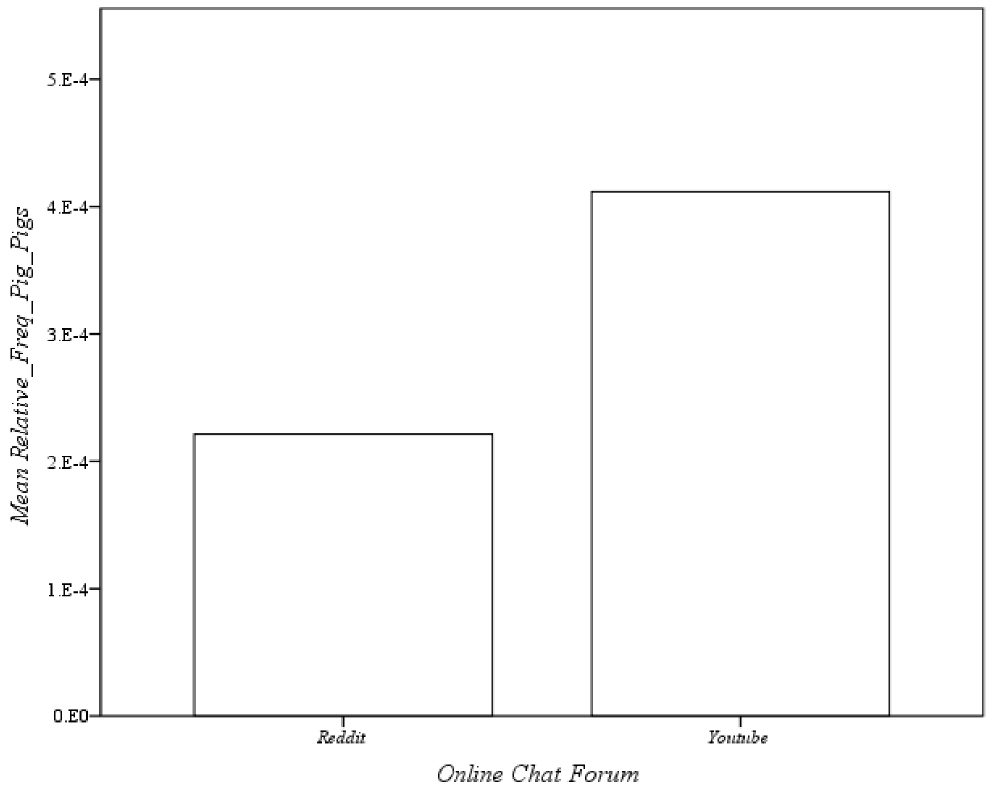

Figure 7. Mean Relative Frequency of Law Enforcement Animal Referent with Online Chat Forum as a Predictor.

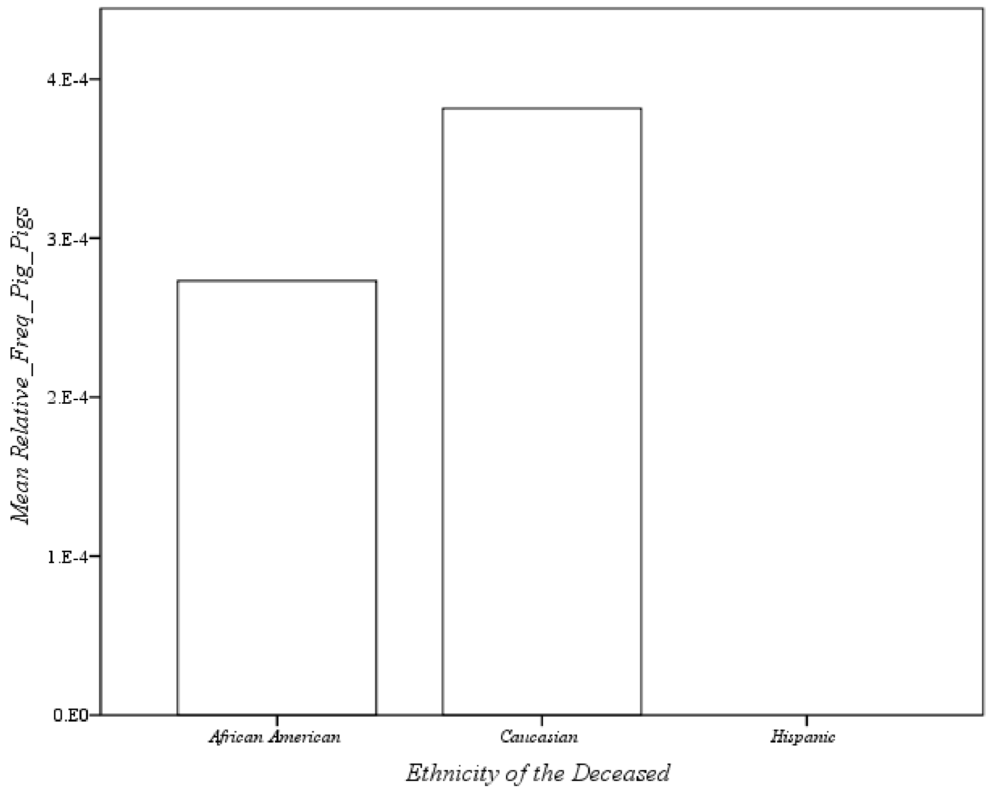

Figure 8. Mean Relative Frequency of Law Enforcement Animal Referent with Ethnicity of the Deceased as a Predictor.

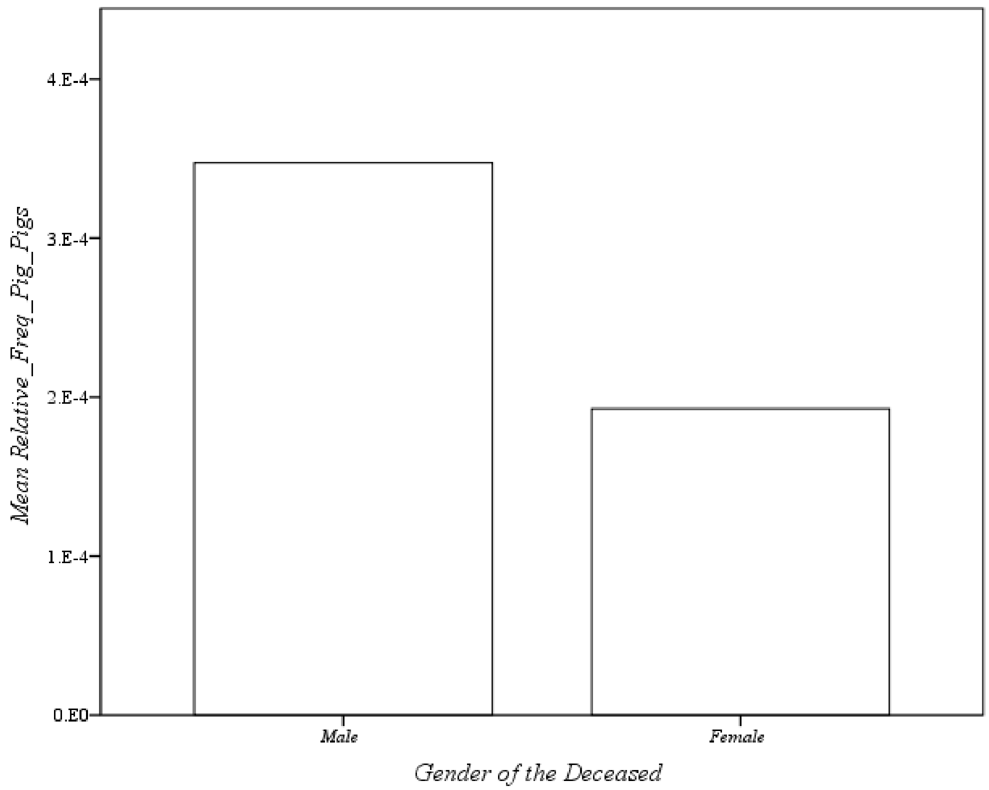

Figure 9. Mean Relative Frequency of Law Enforcement Animal Referent with Gender of the Deceased as a Predictor.

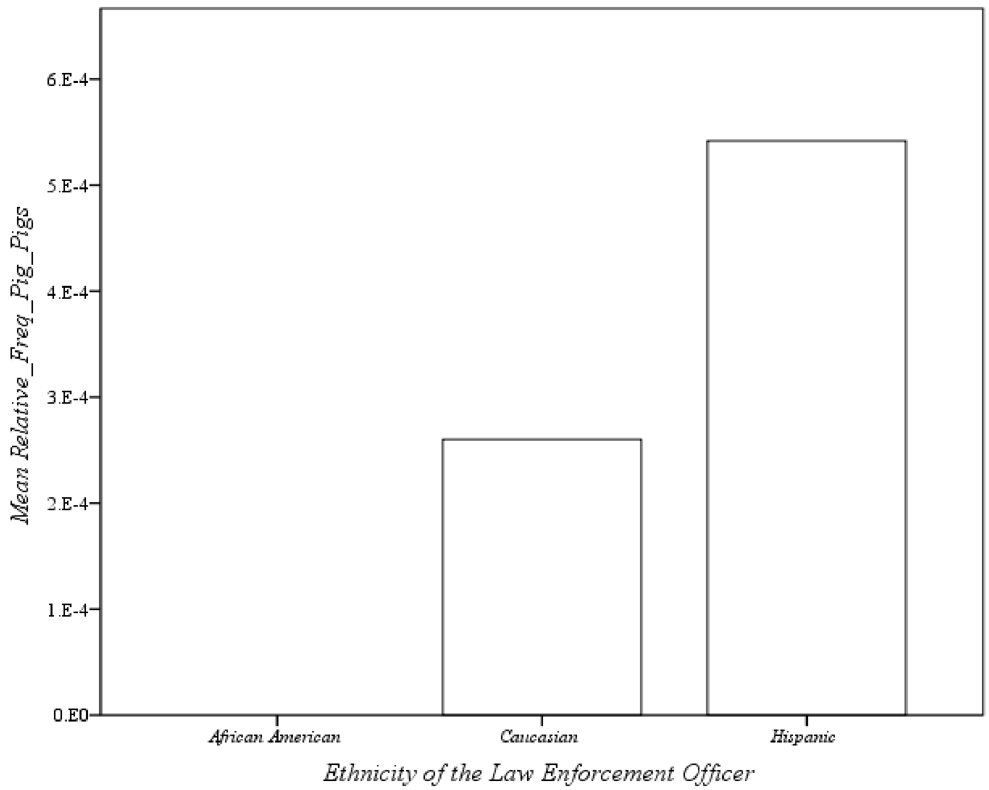

Figure 10. Mean Relative Frequency of Law Enforcement Animal Referent with Ethnicity of the Law Enforcement Agent as a Predictor.

Table 17 provides a summary of the significance of each independent variable and Table 18 shows the final results of the Poisson regression test. To explain, the final results of the Poisson regression test indicated that that when the probability of an ethnic slur for law enforcement animal referents (i.e. pig, pigs) were considered, the online forum Reddit was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = -.690$), the African American ethnicity of the deceased was not significant ($p = .281$; therefore, $p > .05$ and the coefficient estimate $B = .146$), the gender of the deceased (male) was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = .283$), and the Caucasian law enforcement agent variable was significant ($p = .000$; therefore, $p < .05$ and the coefficient estimate $B = -.714$).

Table 17. Tests of Model Effects 2

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	12731.054	1	.000
Online_C_Forum	50.692	1	.000
Ethnicity_Deceased	1.163	1	.281
Gender	4.067	1	.044
Ethnicity_LawEnfor	28.992	1	.000

Dependent Variable: Frequency of Pig or Pigs Used to Refer to Law Enforcement
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, Ethnicity_LawEnfor, offset = Log_Total_Words

Table 18. Parameter Estimates 2

			95% Wald Confidence Interval		Hypothesis Test				95% Wald Confidence Interval for Exp(B)	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Exp(B)	Lower	Upper
(Intercept)	-7.698	.1584	-8.008	-7.387	2362.226	1	.000	.000	.000	.001
[Online_C_Forum=1]	-.690	.0969	-.880	-.500	50.692	1	.000	.501	.415	.606
[Online_C_Forum=2]	0[a]	.	.	.	.	.	.	1	.	.
[Ethnicity_Deceased=5]	.146	.1350	-.119	.410	1.163	1	.281	1.157	.888	1.507
[Ethnicity_Deceased=6]	0[a]	.	.	.	.	.	.	1	.	.
[Gender=1]	.283	.1403	.008	.558	4.067	1	.044	1.327	1.008	1.747
[Gender=2]	0[a]	.	.	.	.	.	.	1	.	.
[Ethnicity_LawEnfor=6]	-.714	.1326	-.974	-.454	28.992	1	.000	.490	.378	.635
[Ethnicity_LawEnfor=7]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1[b]									

Dependent Variable: Frequency of Pig or Pigs Used to Refer to Law Enforcement
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, Ethnicity_LawEnfor, offset = Log_Total_Words
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.
Null Hypothesis: Rejected

DEROGATORY COLLOCATIONS OF LAW ENFORCEMENT AGENTS

Restatement of the Hypothesis:

> H_0 - There is no relationship between the online forum, gender, ethnicity of the deceased, and frequency of derogatory language aimed at law enforcement, when online comments are inspired by ethnically charged events.

Table 19 contains the relative frequency results for the collocations of law enforcement agents inspired by ethnically charged events on Reddit and YouTube. The target lexical item were fuck cops, fucking cop, fucking cops, pussy cop, and pussy cops; and the table appears to show that instances of this type of language used to describe law enforcements is more frequently on YouTube in comparison to Reddit.

Table 19. Antconc Relative Frequency Results - Collocations of Law Enforcement Agents

Online Forum	*Dillion Taylor*	*Eric Garner*	*Laquan McDonald*	*Sandra Bland*	*Zack Hammond*
Reddit	.00000477	.00000690	.00000106	.00000106	.00000265
YouTube	.00001220	.00000583	0	.00000637	.00000436

In addition to the frequency analysis done with Antconc, Table 20 shows that a Poisson probability regression was run in SPSS to determine if the online platform, the ethnicity of the deceased, the gender of the deceased, or the ethnicity of the law enforcement agent affected the probability of the occurrence of target collocations of law enforcement agents.

Table 20. Model Information 3

Dependent Variable	*Frequency of Negative collocates to Describe Law Enforcement*
Probability Distribution	*Poisson*
Link Function	*Log*
Offset Variable	*Log_Total_Words*

Table 21 displays the minimum (in this case 0) and the maximum (in this case 23) usages of collocations of law enforcement agents (M = 8.30, SD. = 6.819) and Figure 11, Figure 12, Figure 13, and Figure 14 show the mean relative frequency of law enforcement animal collocations of law enforcement agents with online chat forum (Reddit, YouTube), ethnicity of the deceased (African American, Caucasian, Hispanic), gender of the deceased (male, female), and ethnicity of the law enforcement agent (African American, Caucasian, Hispanic) as predictors.

Table 21. Continuous Variable Information 3

		N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Dependent Variable	*Frequency of Negative Collocates to Describe Law Enforcement*	*10*	*0*	*23*	*8.50*	*6.819*
Offset	*Log_Total_Words*	*10*	*11.66*	*12.89*	*12.0717*	*.38773*

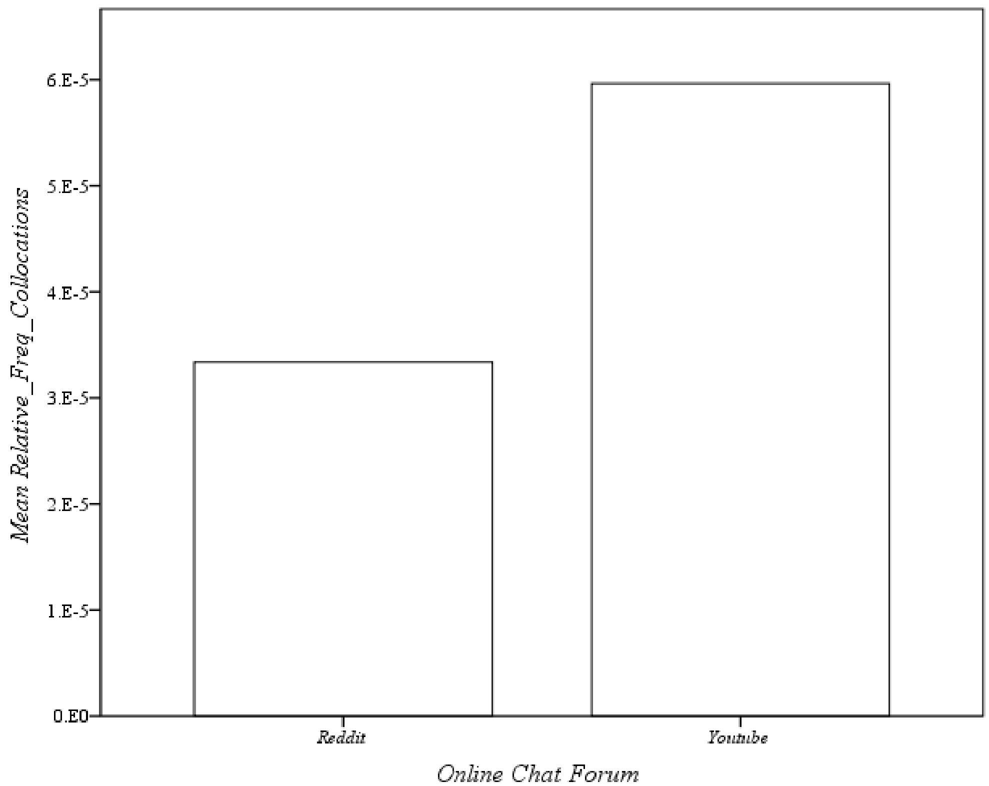

Figure 11. Mean Relative Frequency of Collocations of Law Enforcement Agents with Online Chat Forum as a Predictor.

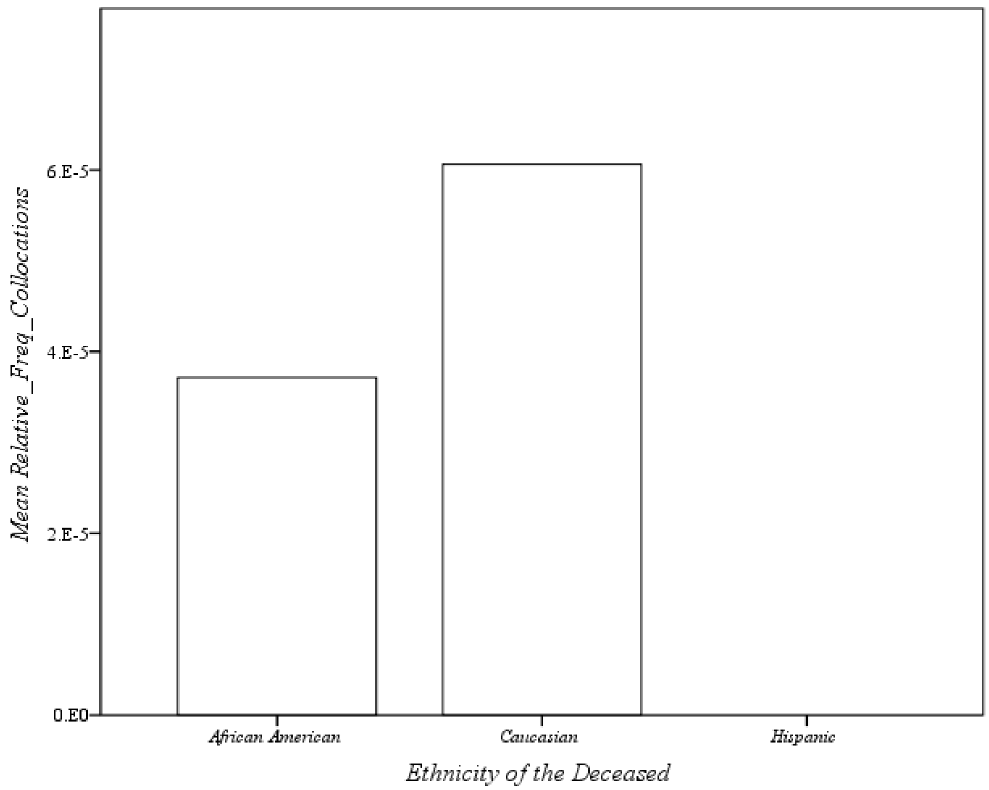

Figure 12. Mean Relative Frequency of Collocations of Law Enforcement Agents with Ethnicity of the Deceased as a Predictor.

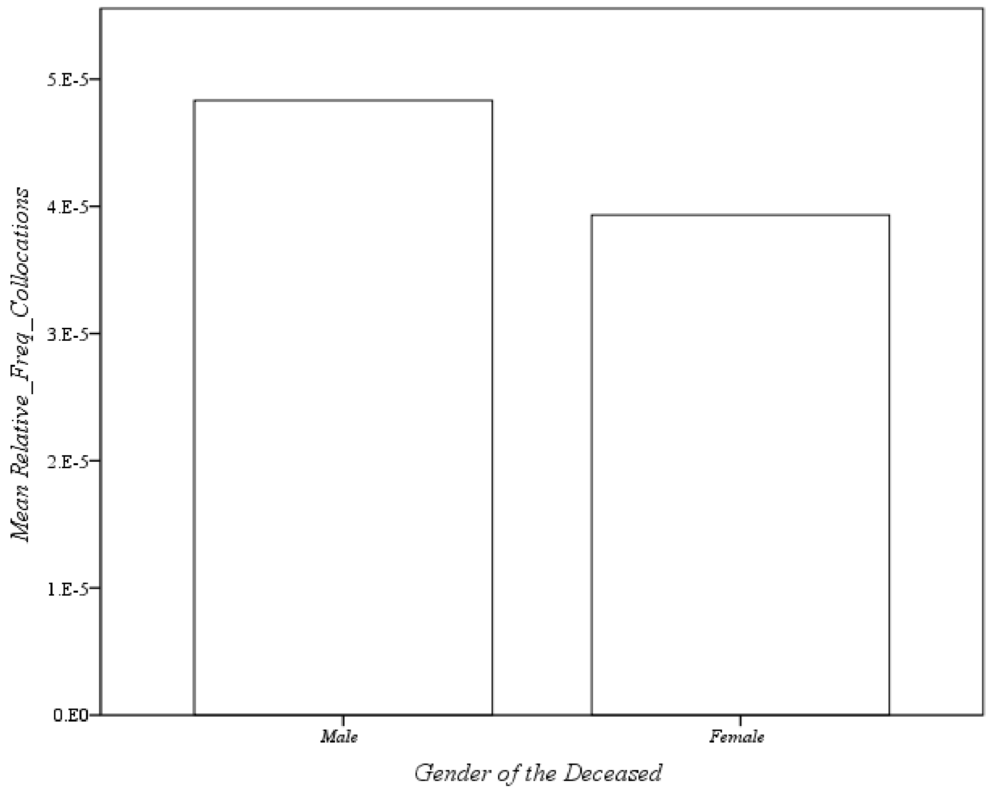

Figure 13. Mean Relative Frequency of Collocations of Law Enforcement Agents with Gender of the Deceased as a Predictor.

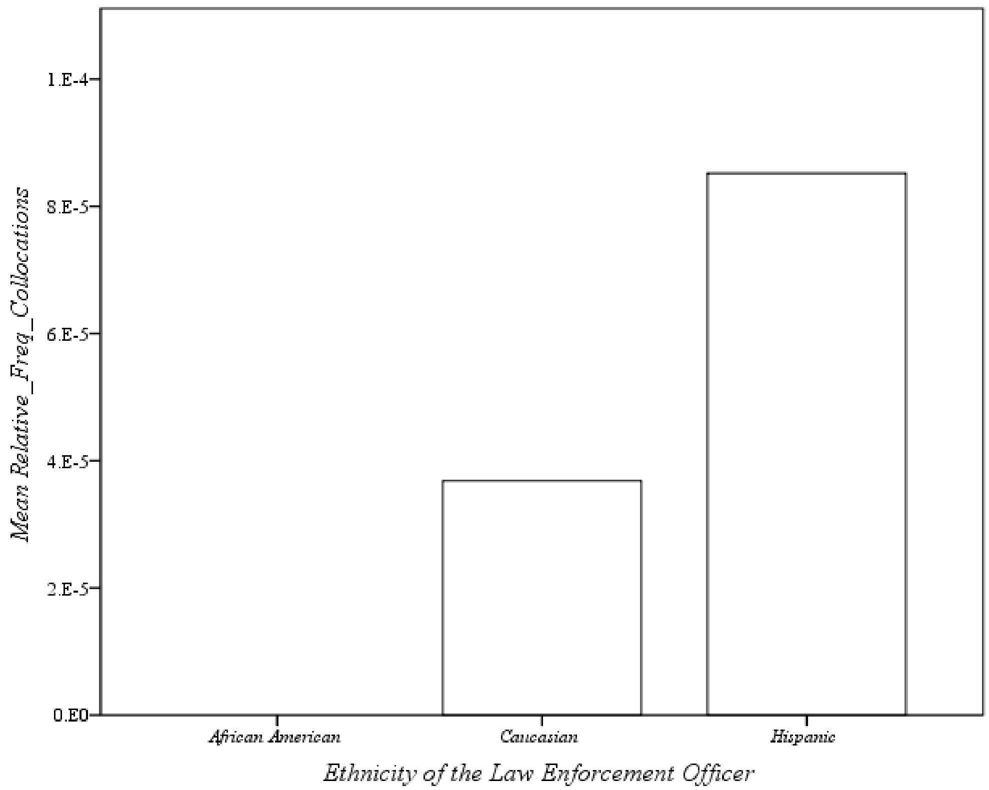

Figure 14. Mean Relative Frequency of Collocations of Law Enforcement Agents with Ethnicity of the Law Enforcement Agents as a Predictor.

Table 22 provides a summary of the significance of each independent variable and Table 23 shows the final results of the Poisson regression test. To explain, the final results of the Poisson regression test indicated that that when the probability of collocations of law enforcement agents (i.e. fuck cops, fucking cop, fucking cops, pussy cop, and pussy cops) were considered, the online forum Reddit was significant ($p = .022$; therefore, $p < .05$ and the coefficient estimate $B = -.537$), the African American ethnicity of the deceased was not significant ($p = .897$; therefore, $p > .05$ and the coefficient estimate $B = .3444$), the gender of the deceased was not significant ($p = .960$; therefore, $p > .05$ and the coefficient estimate $B = -.017$), and the Caucasian law enforcement agent variable was significant ($p = .011$; therefore, $p < .05$ and the coefficient estimate $B = -.841$).

Table 22. Tests of Model Effects 3

Source	*Wald Chi-Square Type III*	*df*	*Sig.*
(Intercept)	*3530.245*	*1*	*.000*
Online_C_Forum	*5.224*	*1*	*.022*
Ethnicity_Deceased	*.017*	*1*	*.897*
Gender	*.002*	*1*	*.960*
Ethnicity_LawEnfor	*6.520*	*1*	*.011*

Dependent Variable: Frequency of Negative collocates to Describe Law Enforcement
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, Ethnicity_LawEnfor, offset = Log_Total_Words

Table 23. Parameter Estimates 3

			95% Wald Confidence Interval		*Hypothesis Test*				*95% Wald Confidence Interval for Exp(B)*	
Parameter	*B*	*Std. Error*	*Lower*	*Upper*	*Wald Chi-Square*	*df*	*Sig.*	*Exp(B)*	*Lower*	*Upper*
(Intercept)	*-9.169*	*.3764*	*-9.906*	*-8.431*	*593.393*	*1*	*.000*	*.000*	*4.985E-5*	*.000*
[Online_C_Forum=1]	*-.537*	*.2348*	*-.997*	*-.076*	*5.224*	*1*	*.022*	*.585*	*.369*	*.926*
[Online_C_Forum=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Ethnicity_Deceased=5]	*.045*	*.3444*	*-.630*	*.720*	*.017*	*1*	*.897*	*1.046*	*.532*	*2.054*
[Ethnicity_Deceased=6]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Gender=1]	*-.017*	*.3328*	*-.669*	*.636*	*.002*	*1*	*.960*	*.984*	*.512*	*1.888*
[Gender=2]	*0*[a]	.	.	.	.	.	.	*1*	.	.
[Ethnicity_LawEnfor=6]	*-.841*	*.3294*	*-1.487*	*-.196*	*6.520*	*1*	*.011*	*.431*	*.226*	*.822*
[Ethnicity_LawEnfor=7]	*0*[a]	.	.	.	.	.	.	*1*	.	.
(Scale)	*1*[b]									

Dependent Variable: Frequency of Negative collocates to Describe Law Enforcement
Model: (Intercept), Online_C_Forum, Ethnicity_Deceased, Gender, Ethnicity_LawEnfor, offset = Log_Total_Words
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.
Null Hypothesis: Rejected

Sentiment Analysis Results

Restatement of the Hypothesis:

> H_0 - The frequency of semantically expressive language is not an accurate predictor of the semantic orientation of a text.

Table 24 shows that despite the significant frequency of the target derogatory language, individual YouTube comment sections were classified as neutral in semantic orientation; while individual Reddit comment sections, despite having fewer instances of the target derogatory language, were deemed semantically negative.

Table 24. Sentiment Analysis Results – Global Positive or Negative Sentiment

Online Forum	*Dillion Taylor*	*Eric Garner*	*Laquan McDonald*	*Sandra Bland*	*Zack Hammond*
Reddit	*Negative*	*Neutral*	*Negative*	*Negative*	*Negative*
YouTube	*Neutral*	*Negative*	*Neutral*	*Neutral*	*Neutral*

Null Hypothesis: Accepted

CHAPTER 10

DISCUSSION

This study was designed to determine the frequency of ethnic slurs and animal referents for African Americans, as well as derogatory language aimed at law enforcement agents found in online comment sections of Reddit and YouTube, when the inspiring event is designated as ethnically charged. Specifically, this study sought to determine whether the frequency of the aforementioned lexical items was significant with regard to the online platform, ethnicity of the deceased, gender of the deceased, and ethnicity of the law enforcement agent. As the results show, both YouTube and Reddit appear to have multiple instances of the ethnic slur and animal referents used to categorize African Americans, as well as language meant to derogate or dehumanize members of law enforcement agencies. However, the results show that within this sample of ethnically charged events, YouTube stands as a positive predictor for each target lexical item (e.g. frequency of ethnic slur, African American animal referents, and derogatory language aimed at law enforcement). In addition, it must be noted that when the target lexical items are present, they collectively appear to factor as less than 1% percent of the entire 1,883,703 word corpus. These results are surprising considering that the current American political and social climate resembles that of 1960's America and all of that era's negative aspects; for example, ethnic fear mongering politicians and civil unrest caused by the death of unarmed African Americans after interactions with law enforcement agents. Furthermore, it is also important to try to discern the possible contributing parameters that affect the occurrences of the target lexical items. To explain, this researcher believes that there are four possible factors that may contribute to the relatively sparse occurrences of the target items and those parameters are as follows: The rules of the online forums regarding posts; community member moderation of online forums; a greater level of tolerance within the online community; and new and more clever ways to derogate and dehumanize participants. Each of these parameters will be thoroughly analyzed in the following sections to determine their merit.

Possible Reasons for the Low Frequency of the Target Lexical Items

Online Forum Rules Regarding Posts

It is easy to imagine that both Reddit and YouTube might want to be the mediators of language generated in their online forums due to both companies desire to have worldwide appeal; however, this is not necessarily the strict policy of either online forum. To explain, Reddit has a policy of removal of posts only if the post contains personal information (e.g. full names, person contacts, and addresses, etc.) or threats of violence towards others (Reddit, 2015). These types of violations will results in the offender being banned from the Reddit community; however, according to the Reddit's regulations, posting language classified as hate speech or the personal contact information of a CEO or congressperson is acceptable (Reddit, 2015). Therefore, it appears that Reddit's rules of the online forum do not contribute to the relatively small amount of the target lexical items present.

With regard to YouTube, the sparse occurrences of the target lexical items may be influenced by rules YouTube adopted in 2008. To explain, in 2008, YouTube adopted new rules and regulations aimed at preventing what YouTube classifies as hate speech (Sarno, 2008; YouTube, 2008). For example, any language that seeks to derogate or dehumanize a YouTube community member based on their ethnic group, religion, disability, gender, age, veteran status, and sexual orientation is actively reported by members, which results in the removal of the offending language. YouTube's policy also dictates that threats of violence towards community members will results in a ban of the offender (Sarno, 2008; YouTube, 2008). Though samples of the YouTube comment sections have more of the target lexical items in comparison to Reddit, YouTube's current rules and regulations may be prevent frequent occurrences of this language.

Community Member Moderation

Community moderation plays a large role in Reddit's online comment platform. For example, members have the ability to up or down vote a comment they disagree with. Since down voting causes a comment to be buried within a Reddit thread, this tactic effectively prevents the comment from being seen due to the sheer number of posts on a daily basis. However, the tools used in this study were able to retrieve those comments and yet, very few of the target lexical items appeared in buried Reddit threads. A reason for this may be because Reddit allows the user(s) responsible for the creation of the comment space to delete threads that they find offensive or unnecessary (Reddit, 2015). This fact could account for the sparse amount of the target items in the comments inspired by ethnically charged events on Reddit.

As previously stated, YouTube initiated new rules and regulations regarding hate speech and tasked users to report such speech in an effort to facilitate its removal (Sarno, 2008; YouTube, 2008). Though it can be assumed that the aforementioned process has some effect, it is unclear how effective it is at preventing occurrences of the target lexical items. Furthermore, it must be once again noted that while the comments analyzed from YouTube and Reddit both have a surprisingly small amount of the target lexical items, YouTube appears to have noticeably more.

A Greater Level of Tolerance Online

A possible explanation for the lack of derogatory and dehumanizing language in the data set could also be attributed to a greater level of tolerance in online communities due to forced diversity created by online technology being affordable and available to virtually all ethnic groups and societal groups. During the 1990's and even throughout portions of the 2000's, the internet usage mostly was relegated to those who could afford it; therefore, it can be assumed that as technology becomes more affordable, various socioeconomic groups and societal groups will be counted as contributors to online comment forums. Given that possibility, a more diverse community would arise, resulting in less of a need for language meant to derogate and dehumanize others.

New Ways to Derogate and Dehumanize

While the sentiment of the previous argument is comforting, it may not necessarily be the correct. To explain, users of online comment forums may have new ways to derogate and dehumanize others within their community. In fact, it can be safely assumed that online commenters, out of fear of being banned, may engage in creative ways to insult when members are in conflict. Though they were not relevant to this study, words such as devil (meant to derogate Caucasians) and expressive language (such as profanity) heavily inhabited the data set and may not be specifically classified as hate speech by YouTube or Reddit. Moreover, all the previously mentioned parameters appear to play a role in the sparse frequency of the target language found within the data set; however, possible reasons for the occurrences of these items must also be addressed.

Possible Reasons for the Frequency of the Target Lexical Items

Anonymity

Anonymity may play a role in the appearance of the target lexical items because online forum participants appear to be more likely to use controversial, derogatory, and dehumanizing language when their identity cannot be reveled (Omernick & Sood, 2013). Both Reddit and YouTube allow for anonymous access to their sites; however, many users of YouTube may be logged in as their GMAIL identities since Google owns both GMAIL and YouTube. This possibility is worth mentioning because if users chose a real world name for their GMAIL account, then they will most likely be less inclined to use language in violation with YouTube’s doctrine. Despite the possibility of usage of an account connected to a potential personal email account, YouTube members in the data set used more of the target lexical items in comparison to Reddit users. Furthermore, it is clear that YouTube members are comprised of the anonymous and non-anonymous users; however, it is unclear to what extent anonymity plays in the usage of the target lexical items; although, it can be assumed based on previous research that it plays some role (Omernick & Sood, 2013).

The Video Aspect

Another possible reason for the disparity between the occurrences of the target lexical items within the Reddit and YouTube data set could be due to a feature of YouTube that Reddit does not possess. To explain, YouTube users have the advantage (or disadvantage) of viewing a video of an ethnically charged event before they comment; and it is possible that the video could influence their comment and inspire them to derogate and dehumanize ethnic or societal groups represented in the video. For example, the video viewed could be a biased representation of the ethnically charged event and it could contain user audio commentary that could potential lead the viewer in a negative direction. During this study, it was noticed that videos inspired by the death of unarmed suspects during interactions with law enforcement agents were often prefaced with user commentary containing language that derogated and dehumanized law enforcement agents. Likewise, similar commentary was observed as epilogues to the ethnically charged events. Though the accompaniment of a video and even anonymity may be factors in the overall presence of the target lexical items, certain parameters were overserved within the sample revealed themselves as predictors. These parameters are ethnicity of the deceased, gender of the deceased, and ethnicity of the law enforcement agent mirroring the ethnicity of the deceased. Each will be discussed thoroughly in the following sections.

ETHNICITY OF THE DECEASED

A relevant predictor for the high occurrence of the ethnic slur nigger was the ethnicity of the deceased (African American). Likewise, the predictor for the high occurrence of the target lexical items monkey, chimpanzee, gorilla, and ape (when used as African American animal referents) was also the ethnicity of the deceased (African American). This finding is not surprising when the deceased was of the African American ethnic group, since that target lexical item is often used by hate groups to derogate African Americans. However, 1 instance of the word nigger and 5 African American animal referents were found in the YouTube data set inspired by the death of Dillon Taylor (Caucasian), as well as 3 African American animal referents in the data set inspired by the death of Zack Hammond (Caucasian). To show the usage of the target lexical item nigger, an excerpt from YouTube discourse inspired by the death of Eric Garner is provided in Table 25.

As Table 25 shows, not only does the target ethnic slur appear 2 times in the discourse (used by Detective Hawk and Neko Neko), it is followed by an ethnically inflammatory prediction and classification of the victim’s motive for being involved in suspected criminal activity. Many instances of these types of inflammatory rhetoric appeared within the data sets inspired by the deaths of African Americans; however, this instance is particularly significant because of the comment made by the user “James Baledrokadroka”.

To explain, this user, despite the previous user's inflammatory word choice, attempts to show sympathy for the possible issues facing African Americans in American society. However, user Neko Neko immediately rebuffs this show of sympathy by posting an ethnically inflammatory comment outlining Eric Garner's possible actions that directly led to his death.

Table 25. Excerpt of Unedited YouTube Comments Inspired by the Death of Eric Garner 1

Relevant Aspects: Ethnic Slur for African Americans
Detective Hawk: *Nigger probably deserved it, if he didn't commit a crime then he was bound to commit one in the future. It is in his blood….*
James Baledrokadroka: *Im not an American but an Australian. I must say that I feel for African American people in the U.S. I feel their pain. Discrimination, Racism, Hate.*
Neko Neko: *Eric Garner was basically choked to death because he was just another stupid nigger who thought it was a good idea to resist an arrest attempt.*

Similar to Table 25, Table 26 is also representative of the discourse found in the YouTube data sets. Table 26 shows language that is representative of an increase in the usage of the ethnic slur nigger and African American animal referents. When the ethnicity of the deceased is African American.

Furthermore, Table 26 is particularly of interest because it contains lexical items such as "cracker" and "devil", which are meant to derogate and dehumanize the Caucasian ethnic group. While not exactly relevant for the goal of this study, it must be noted that many instances of language meant to derogate and dehumanize Caucasians were found within many of the data sets. In fact, Table 26 shows how this language used with hyperbolic language and the target lexical items (e.g., ethnic slur for African Americans, and derogatory language aimed at law enforcement), is used to initiate and respond to posts related to ethnically charged events. For example, phrases such as "some worthless rapist nigger from a hundred years ago" and "fucking cops savagely murdering and dismembering an innocent human" are clear markers of ethnically charged and hyperbolic language used to qualify the ethnically charged event. To explain, Eric Garner was not the ethnic slur used to describe him, his life was not worthless, and his recent death (not one hundred years ago) was not due to dismemberment by law enforcement. Furthermore, the ethnicity of the deceased (African

American) was found to be a predictor for the appearance of African American animal reference, as seen in the posited by “IHaveEbola”.

Table 26. Excerpt of Unedited YouTube Comments Inspired by the Death of Eric Garner 2

Relevant Aspects: Ethnic Slur For African Americans - African American Animal Referent - Derogatory Language Aimed Law Enforcement
Ethnologizer: *Learn to read, cracker.*
IHaveEbola: *"conclusions" are not evidence, nigger.*
Ethnologizer: *No, devil, conclusions are based on evidence. I like how you try to use Le Vin's report against me only to immediately try to cast aspersions on it when I point out that it doesn't support you. You really are just dumb, aren't you? You're also kinda starting to bore me.*
IHaveEbola: *i've been bored of arguing about what happened to some worthless rapist nigger from a hundred years ago for a while.*
Ethnologizer: *Wouldn't really call what you've been doing "arguing" - more like obvious lies and pathetic squirming in the face of fucking cops savagely murdering and dismembering an innocent human. Entertaining at first, but now repetitive.*
IHaveEbola: *whatever, rape ape.*

Gender of the Deceased

The most relevant predictor for the high frequency of the ethnic slur nigger within the data sets was the combination of the gender of the deceased (African American) and the gender of the deceased (female). Likewise, the most relevant predictor for the high occurrence of the target lexical items monkey, chimpanzee, gorilla, and ape (when used as African American animal referents) was also the ethnicity of the deceased (African American) and the gender of the deceased (female). This finding is surprising because previous to this study, it could have been assumed that gender would play a smaller role in the appearance of the target lexical items than ethnicity, and that the combination of the two would have minimal influence on the data set. Table 27 and Table 28 both show examples of how the combination of the ethnicity of the deceased (African American) and gender of the deceased (female) inspire significant occurrences of the target lexical items and contradicts the aforementioned assumption.

Table 27. Excerpt of Unedited YouTube Comments Inspired by the Death of Sandra Bland 1

Relevant Aspects: Ethnic Slur for African Americans - African American Animal Referent - Gender Female - Derogatory Language Aimed at Law Enforcement
Jeremy YouTube: *What a pig. He basically let his ego escalate things into an arrest.*
Americarulez55: *Im getting sick of cuck nigger lovers like you. Stop defending apes you prick.*
Themrajester2: *Bitch can u spell yo dumbass.*
RicFortin: *Slightly better than you can obviously…go back to your PS4 kid, your brain must be exhausted after that six word sentence….*

Table 27 is interesting because it represents typical comments observed when the gender of the deceased is female. Specifically, it represents the higher frequency of nigger, monkey, gorilla, chimpanzee, and ape, in the data set when the deceased is an African American female (Sandra Bland). Furthermore, Table 27 is even more noteworthy because it contains an instance of a YouTube member choosing to attack another member for spelling errors, while disregarding a member who uses language meant to derogate and dehumanize African Americans. However, this type of occurrence was not isolated to gender related comments since it was found to occur throughout many of the data sets. In contrast, Table 28 shows examples of a higher frequency of the target lexical items coupled with gender specific attacks.

As Table 28 shows, user "Prophett Mohamed" uses an ethnic slur and an animal referent to refer to Sandra Bland and another member of the YouTube community. What makes this selection of discourse interesting is "Prophett Mohamed's" gender specific attacks such as "This cunt should have stayed in the kitchen where she belonged" and "You fucking nigger cunt. Why aren't you in the kitchen prepping me a meal?? Do I need to beat your ape ass again? The last time I did that you shat out a little boy (extra slave for me)". While portions of this language is not designated as the target lexical items of this study, it cannot be ignored because it shows a connection between the gender of the deceased (female) and an increase in derogatory language in general. Furthermore, even when it assumed that "Prophett Mohamed" is speaking to a male member ("Peepowman"), the user still attacks with rhetoric designed to derogate and dehumanize females. This behavior found in other discourse within the corpus, could be the reason an ethnic slur for African Americans and African American animal referents were more frequent when the ethnicity of the deceased was African American, and the gender was female.

Table 28. Excerpt of Unedited YouTube Comments Inspired by the Death of Sandra Bland

Relevant Aspects: Ethnic Slur for African Americans - African American Animal Referent - Gender Female
Thobias Marandu: *What was His intention to move her to the side of the street where the camera cannot capture what he was doing to her?*
Prophett Mohamad: *+Thobias Marandu Because NIGGERS aren't allowed on the roads. This cunt should have stayed in the kitchen where she belonged.*
Peepowman: *+Prophett Mohamad there's a special place in hell for ignorant cunts like you. You have the words "prophet Mohammed" in your profile, you stupid, uneducated, racist cunt-bag. I suggest you go dig yourself a hole and jump in it. You fucking idiot twat.*
Prophett Mohamad: *You fucking nigger cunt. Why aren't you in the kitchen prepping me a meal?? Do I need to beat your ape ass again? The last time I did that you shat out a little boy (extra slave for me).*

Ethnicity of the Law Enforcement Agent/Deceased

Finally, it was observed that within the corpus, the ethnicity of the law enforcement agent, when it matched the ethnicity of the deceased, was a predictor for the high frequency of derogatory language aimed at law enforcement agents; however, another interesting phenomenon must be addressed first. To explain, the results indicated that as a whole, law enforcement agents were derogated and dehumanized twice as much as ethnic groups or genders within the data sets. Furthermore, users observed within the YouTube comment samples used twice as many terms to derogate and dehumanize law enforcement agents than Reddit users. However, as previously stated, the ethnicity of the law enforcement agent, when it mirrored that of the ethnicity of the deceased, was a predictor of the higher frequency of the target lexical items. To explain, when the results indicated that when the law enforcement agent was Caucasian and the deceased was Caucasian, the overall frequency of the derogatory language aimed at law enforcement is higher. This could possible imply that users (at least within this corpus) are less likely to use derogatory language aimed at law enforcement when the deceased is African American as opposed to Caucasian. It also must be noted that according to the analysis of the data, Hispanic law enforcement agents are likely to be the target of the most frequent usages of derogatory language, when the deceased is Caucasian. Table 29 and Table 30 show language that represents the aforementioned phenomena.

Table 29. Excerpt of Unedited Reddit Comments Inspired by the Death of Zack Hammond

Relevant Aspects: Derogatory Language Aimed Law Enforcement
Eli5: *why do cops not get automatically fired for refusing to release footage? Can't the footage just be taken away?* *#everylifematters #exceptdirtypigs #fuckthepolice* ***Thisbymaster:*** *This is disgusting... I don't want to be on this planet anymore. Fuck cops. Every last one.* ***treehuggerguy:*** *K bye* ***Thisbymaster:*** *Just a few bad apples spoil the bunch.* ***Eli5:*** *God you have to feel bad for the family. After going through the death of their child, they have to put up with the police screwing around with their kids dead body? It's plain criminal. You've got to give them a hand for being so brave.....*

Table 29 is interesting because it shows the level of sympathy for the deceased and the level of hatred for law enforcement found within the data set when the law enforcement agent is Caucasian and the deceased is Caucasian. For example, Table 29 is representative of how users in the data sets used hash tags (e.g. #fuckthepolice) in conjunction with target the lexical items pig, pigs, and fuck cops to derogate and dehumanize law enforcement agents at a higher frequency when the deceased was Caucasian. It is also to note, that while not particularly relevant to the frequency of the target lexical items designated by this study, this excerpt displays strong sympathy for the deceased and their family, and sentiment of this type was noticed to be more frequent in the data sets where the deceased was Caucasian. In contrast, Table 30 represents user's strong negative sentiment towards law enforcement agents, when the law enforcement agents is Hispanic and the deceased is Caucasian.

Table 30. Excerpt of Unedited YouTube Comments Inspired by the Death of Dillon Taylor

Relevant Aspects: Ethnicity of the Deceased - Ethnicity of the Law Enforcement Agent - Derogatory Language Aimed Law Enforcement
BrandWatkins: *That dumbfuck cop knows he's dead. him saying stay with me stay with me pisses me off.*
ICECREAMKING1: *Im sorry but this pig needs to be executed.*
Sexy Kaleja: *I hate pigs fuk all pigs…die rot in hell*
JoeMurray: *Kill an innocent man…if you see a cop before he sees you…shot him*
lola: *in this case the faults is justice corruption…i dont know how the jury for it justify. must pay jury. jury corruption…judge corruption..pig corruption.*

The language in Table 30 is typical of the discourse found in the data set where the law enforcement agent was Hispanic and the deceased was Caucasian. As Table 30 shows, language meant to derogate and dehumanize law enforcement agents is of a noticeably higher frequency than the other data sets that do not meet the same parameters. It also important to note, that while not specifically relevant to the goal of this study, inflammatory language such as the kind in Table 30 clearly violates YouTube's previously mentioned rules and regulations because it contains threats of violence towards a YouTube members (i.e. law enforcement).

SENTIMENT ANALYSIS

The results from the sentiment analysis done on each data set was interesting because it may shed light on a possible flaw in the sentiment analysis procedure. To explain, despite the higher frequency of the target lexical items, individual YouTube comment sections were classified as neutral in semantic orientation; while individual Reddit comment sections, despite having fewer instances of the target lexical items, were deemed semantically negative. The two possible explanations for this phenomena in this researcher's opinion are as follows:

1. Some semantically expressive language, specifically the ethnic slur for African Americans (i.e. nigger), is semantically subject to the same judgements as descriptive language and semantically expressive profanity (e.g. shithead, asshole, fucker, etc.), with no special allowances given for the effect this lexical item can have on the African American community. Therefore, certain YouTube data sets with a higher frequency of the ethnic slur remained neutral in their semantic orientation.
2. Since the frequency of the semantically expressive ethnic slur nigger makes up less than 1% of the 1,883,703 word corpus, its semantic orientation is overshadowed by the lexical items that are not an ethnic slur for African Americans. Therefore, certain YouTube data sets with a higher frequency of the ethnic slur remained neutral in their semantic orientation.

Either of these is a probable reason for the semantic orientation of certain YouTube data sets; however, it is possible to assume both parameters may have played a role in determining the semantic orientation of certain YouTube data sets. To explain, semantically expressive language, specifically ethnic slurs, may be problematic when assigning value. It is easy to assume that most non-ethnically biased or non-misguided individuals would agree that ethnic slurs deserve a negative semantic orientation; however, the level of negative assignment may be in dispute. For example, a Hispanic coder may not assign the same negative rating as an African American coder to an ethnic slur for African Americans because Hispanics have a different history with the lexical item. Likewise, a Caucasian coder might assign a more negative rating to an ethnic slur for African Americans than an African American coder might, for fear of being labeled a racist by users of the sentiment analysis software. Moreover, coders of all ethnic groups may opt out of the problematic nature of assigning a negative semantic orientation to ethnic slurs and instead, leave coding that addresses these lexical items out of the program altogether. In conjunction with issues related to coder preference, the size of the corpus in relation to the amount of semantically expressive language most likely had an effect on the semantic orientation of the YouTube data sets despite the number of occurrences of the target lexical items. This answer is quantifiable and most likely contributed more to the results of the semantic analysis of the corpus.

CHAPTER 11

CONCLUSIONS AND LIMITATIONS

This study sought to create a specialized corpus to determine the frequency of an ethnic slur (i.e. nigger) and animal referents for African Americans (e.g. monkey), as well as derogatory language aimed at law enforcement agents (e.g. pig) found on Reddit and YouTube; specifically, when the language inspiring event is ethnically charged and is mirrored on both platforms. While the results indicated that YouTube stands as a positive predictor for an overall high frequency of each of the target lexical items, there were other predictors as well. For example, when the deceased was African American, there was a high rate of ethnic slurs and African American Animal referent; however, when the deceased was African American and female, the frequency of the target lexical items was at its highest. Furthermore, when the ethnicity of the law enforcement agent was Hispanic and the deceased Caucasian, a relatively large amount of derogatory language aimed at law enforcement was present in the data sets of both YouTube and Reddit. To discern this information, a 1,883,703 word corpus was created using Enthought Canopy, Python Reddit API Wrapper, and YouTube Comment Scrapper; however, there were severe limitations inherently built into the study due to scope. To explain, online comment forums, specifically Reddit and YouTube, exist as depositories of enormous amounts of potentially anonymous data which is likely from received from all around the world. Given this fact, it is virtually impossible to create a study which seeks to obtain an acceptable sample size to show whether or not phenomena exists within Reddit and YouTube, or one that seeks to classify potentially anonymous user's ethnic backgrounds. Therefore, this study only sought to collect data from and analyze a small sample of what this researcher classifies as language inspired by ethnically charged events found on Reddit and YouTube, and to be descriptive of those events. This study in no way means to classify the whole of the Reddit and YouTube communities based on the findings. Furthermore, this researcher acknowledges that even the small chosen sample size may be suspect due to its lack or certain parameters. For example, this study only analyzed events from events designated as ethnically charged or events related to the death of armed/unarmed suspects after interactions with law enforcement agents. However, a neutral unrelated event (e.g. a sporting event) should have been added as a better control group than the one used (Caucasian deceased with Caucasian law enforcement agent). Likewise, an analysis of a female other than Sandra Bland (African American) might have yielded results that showed whether the gender of the deceased was a predictor, without it being coupled with the ethnicity of the deceased parameter. Any future studies from this researcher with take the aforementioned limitations into account with an eye towards designing a more comprehensive study.

As previously mentioned, the anonymous aspect of Reddit and YouTube makes the data potentially unreliable for various reasons. Simple stated, there is no way to verify the ethnic makeup or age of any of the users who comment on the both platforms. This aspect severely limited the research because it would have been useful to provide some kind of quantitative analysis of what ethnic groups are contributing the frequency of the target lexical items on Reddit and YouTube. Likewise, there is no reasonable way to verify that each user or comment is unique, and not the product of a few individuals seeking to engage in trolling or the spamming of online comment forums with ethnically inflammatory language.

In addition to the aforementioned inherent and design limitations, the lexical items chosen for analysis by this study may also count as a limitation. To explain, though thoroughly justified and initially noticed in a test run of the Antconc software, many of the lexical items were chosen purely out of convenience or selfish reasons. For example, the ethnic slur nigger for African Americans was chosen as a target lexical item solely because this researcher is a member of the African American community and it was noticed in the initial test of the Antconc software. Similarly, the lexical items for African American animal referents and the derogatory language aimed at law enforcement was chosen due to their presence during the initial test of the Antconc software. For the aforementioned reasons, this study failed to account for the presence of other words meant to derogate and dehumanize African American and other ethnic groups. For example, as mentioned in the Discussion portion of this study, words such as “cracker” and “devil’ were used to derogate and dehumanize Caucasians; however, the frequency of these words were not accounted for.

Future research should include a comprehensive list of justifiable ethnic slurs that are representative of the major ethnic groups in America.

Finally, Sentiment Analyzer was fairly problematic when it came to discerning if semantically expressive language, specifically an ethnic slur for African Americans, affected the semantic orientation of date sets. To explain, as previously mentioned in the Discussion section of this study, there is no way to determine if the coding for sentiment analysis or the frequency of the target lexical item is responsible for the neutral assignment given to some of the YouTube data sets (despite their overall higher frequencies of the target lexical items). The freeware tool Sentiment Analyzer, created by Dr. Daniel Soper, was used to assigned sentiment; however, the website does not contain an explanation on how the software classifies ethnic slurs.

While the scope of this study remains relatively small, it represents a step forward in using self-made corpora to analyze language inspired by real world events that affect the general public. These events can be of an ethnically charged nature and often serve as topics for university classrooms discussions and papers; therefore, instructors should be aware of the kind of language found in online forums related to these events that may influence student's analysis and output of tasks based on the event. The study was also completed at a minimal cost due to the various freeware and open source tools available for the creation and analysis of corpora. Hopefully, these tools will democratize and diversify research in the same way that affordability and technological advancements have democratized and diversified online spaces.

REFERENCES

Abbey-Lambertz, K. (2015, February 13). *These 15 black women were killed during police encounters. Their lives matter, too* [Newspaper article]. Retrieved from http://www.huffingtonpost.com/2015/02/13/black-womens-lives-matter-police-shootings_n_6644276.html

Akbarzadeh, S., & Smith, B. (2005). *The representation of Islam and Muslims in the media: (The Age and Herald Sun newspapers)* [Scholarly project]. Retrieved from http://asiainstitute.unimelb.edu.au/__data/assets/pdf_file/0017/571112/Islam-in-the-Media.pdf

Anderson, M. (2015). *5 facts about online video, for YouTube's 10th birthday*. Retrieved from http://www.pewresearch.org/fact-tank/2015/02/12/5-facts-about-online-video-for-youtubes-10th-birthday/

Bailey, J. (2006). Offensive language afflicts black youth's psyche. *Western Journal of Black Studies*, *30*(3), 142.

Baker, P., Gabrielatos, C., & McEnery, T. (2012). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics*, *34*(3), 255-278.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, England: Cambridge University Press.

Boczkowski, P. J. (2010). *News at work: Imitation in an age of information abundance*. Chicago, IL: University of Chicago Press.

Boczkowski, P. J., & Mitchelstein, E. (2012). How users take advantage of different forms of interactivity on online news sites: Clicking, E-mailing, and commenting. *Human Communication Research*, *38*(1), 1-22.

Bradner, E. (2014). *Factcheck: Grim statistics on race and police killings*. Retrieved from http://www.cnn.com/2014/12/02/politics/kristoff-oreilly-police-shooting-numbers-fact-check/

Brandl, S. G., Frank, J., Worden, R. E., & Bynum, T. S. (1994). Global and specific attitudes toward the police: Disentangling the relationship. *Justice quarterly*, *11*(1), 119-134.

Brunson, R. K. (2007). "Police don't like black people": African-American young men's accumulated police experiences. *Criminology & Public Policy*, *6*(1), 71-101.

Cheng, X., Dale, C., & Liu, J. (2008). Statistics and social network of YouTube videos. In *16th International Workshop on Quality of Service* (pp. 229-238). doi: 10.1109/IWQOS.2008.32

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*(02), 202-224.

Croom, A. M. (2011). Slurs. *Language Sciences*, *33*(3), 343-358.

Daniels, J. (2013). Race and racism in Internet studies: A review and critique. *New Media & Society*, *15*(5), 695-719.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447-464.

Feagin, J. (2013). *Systemic racism: A theory of oppression*. New York, NY: Routledge.

Freeman, B., & Chapman, S. (2007). Is "YouTube" telling or selling you something? Tobacco content on the YouTube video-sharing website. *Tobacco Control*, *16*(3), 207-210.

Gabrielson, R., Grochowski Jones, R., & Sagara, E. (2014). *Deadly force, in black and white.* Retrieved from https://www.propublica.org/article/deadly-force-in-black-and-white?utm_source=et

Gawron, J. M., Gupta, D., Stephens, K., Tsou, M. H., Spitzberg, B., & An, L. (2012, June). *Using group membership markers for group identification in web logs*. Poster session presented at Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland.

Ghadessy, M., Henry, A., & Roseberry, R. L. (Eds.). (2001). *Small corpus studies and ELT: Theory and practice* (Vol. 5). Amsterdam, Netherlands: John Benjamins Publishing.

Ghitis, F. (2014). *Police shootings in U.S. out of hand*. Retrieved from http://www.cnn.com/2014/12/01/opinion/ghitis-police-shootings/index.html

Hedger, J. A. (2012). The semantics of racial slurs: Using Kaplan's framework to provide a theory of the meaning of derogatory epithets. *Linguistic and Philosophical Investigations*, *11*, 74.

Hom, C. (2008). The semantics of racial epithets. *The Journal of Philosophy*, *105*(8), 416-440.

Hughey, M. W., & Daniels, J. (2013). Racist comments at online news sites: A methodological dilemma for discourse analysis. *Media, Culture & Society*, *35*(3), 332-347.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, England: Cambridge University Press.

James, L., Klinger, D., & Vila, B. (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: experimental results from high-fidelity laboratory simulations. *Journal of Experimental Criminology*, *10*, 323-340.

Jeshion, R. (2013). Expressivism and the offensiveness of slurs. *Philosophical Perspectives*, *27*(1), 231-259.

Kenworthy, J. B., Barden, M. A., Diamond, S., & del Carmen, A. (2011). Ingroup identification as a moderator of racial bias in a shoot–no shoot decision task. *Group Processes & Intergroup Relations*, *14*(3), 311-318.

Kindy, K., & Fisher, M. (2016). *Investigation: People shot and killed by police this year*. Retrieved from https://www.washingtonpost.com/graphics/national/police-shootings/

King, M. L., Jr. (1967). *Where do we go from here: Chaos or community?* (Vol. 1). Boston, MA: Beacon Press.

Klein, A. (2012). Slipping racism into the mainstream: a theory of information laundering. *Communication Theory*, *22*(4), 427-448.

Lee, Y. J., Kim, E. K., Cho, H. G., & Woo, G. (2013). Detecting and visualizing online dispute dynamics in replying comments. *Software: Practice and Experience*, *43*(12), 1395-1413.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and methodologies for Teaching Natural Language Processing and Computational Linguistics: Volume 1* (pp. 63-70). Stroudsburg, PA: Association for Computational Linguistics.

MacGregor, P. (2007). Tracking the online audience: Metric data start a subtle revolution. *Journalism Studies*, *8*(2), 280-298.

Maddow, R. (2016). *Trump campaign style harkens back to George Wallace in 1968*. Retrieved from http://www.msnbc.com/rachel-maddow-show/watch/trump-campaign-style-recalls-wallace-in-1968-596500547585

Martindale, C. (1990). Coverage of black Americans in four major newspapers, 1950-1989. *Newspaper Research Journal*, *11*(3), 96.

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), 1-28.

McCready, E. S. (2010). Varieties of conventional implicature. *Semantics and Pragmatics*, *3*(8), 8-1.

Meneses, E. H. (1994). Race and ethnicity: An anthropological perspective. *Race, Sex & Class*, *1*(2), 137–146

Milgram, S. (1974). Compliant subjects (Book reviews: Obedience to authority. An experimental view). *Science*, *184*, 667-669.

Mills, R. (2011). *Researching social news – is reddit.com a mouthpiece for the 'hive mind' or a collective intelligence approach to information overload?* Paper presented at the meeting of ETHICOMP 2011 the Social Impact of Social Computing, Sheffield, UK. Retrieved from http://eprints.lancs.ac.uk/id/eprint/
61646

Mishra, A., & Rastogi, R. (2012). Semi-supervised correction of biased comment ratings. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 181-190). doi:10.1145/2187836.2187862

Montgomery, D. (2016). *Texas trooper who arrested Sandra Bland is charged with perjury*. Retrieved from http://www.nytimes.com/2016/01/07/us/texas-grand-jury-sandra-bland.html?_r=0

Moreno, A. I., & Suárez, L. (2008). A study of critical attitude across English and Spanish academic book reviews. *Journal of English for Academic Purposes*, *7*(1), 15-26.

Mukherjee, S., & Bhattacharyya, P. (2013). *Sentiment analysis: A literature survey*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1304/1304.4520.pdf

Omernick, E., & Sood, S. O. (2013). The impact of anonymity in online communities. In *2013 International Conference on Social Computing* (pp. 526-535). doi: 10.1109/SocialCom.2013.80

O'Neil, D. (2006, July 5). *Ethnicity and race: Overview.* Retrieved from http://anthro. palomar.edu/ethnicity/ethnic_1.htm

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing: Vol. 10. Association for Computational Linguistics* (pp. 79-86). doi:10.3115/1118693.1118704

Payne, B. K. (2006). Weapon bias split-second decisions and unintended stereotyping. *Current Directions in Psychological Science*, *15*(6), 287-291.

Pew Research Center. (2012). *Trayvon Martin killing is public's top news story*. Retrieved from http://www.people-press.org/2012/03/27/trayvon-martin-killing-publics-top-news-story/

Reddit. (2015). *Frequently asked questions*. Retrieved from https://www.reddit.com/wiki/ faq

Reddit. (2016). *Traffic stats*. Retrieved from https://www.reddit.com/r/AskReddit/about/ traffic

Rey, J. M. (2014). Changing gender roles in popular culture: Dialogue in *Star Trek* episodes from 1966 to 1993. In G. Leech & J. Thomas (Eds.), *Variation in English: Multi-dimensional studies* (pp. 138-156). New York, NY: Routledge.

Rissanen, M. (2000). The world of English historical corpora. *Journal of English Linguistics*, *8*(1), 7-20.

Ronkin, M., & Karn, H. E. (1999). Mock Ebonics: Linguistic racism in parodies of Ebonics on the Internet. *Journal of Sociolinguistics*, *3*(3), 360-380.

Sarno, D. (2008, December 3). Slipping down the slope of YouTube's new sex and language guidelines [Web log post]. Retrieved from http://latimesblogs.latimes.com/technolo gy/2008/12/slipping-down-t.html

Steinbaur, T. (2012). *Information and social analysis of Reddit*. Retrieved from http://snap.stanford.edu/class/cs224w-2011/proj/tbower_Finalwriteup_v1.pdf

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267-307.

Tenenboim, O., & Cohen, A. A. (2015). What prompts users to click and comment: A longitudinal study of online news. *Journalism*, *16*(2), 198-217.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417-424). doi: 10.3115/1073083.1073153

Valente, A. (2009). *A corpus-based study on adjective use in 16th century prayers and letters*. Munich, Germany: GRIN Verlag.

Weitzer, R. (1999). Citizens' perceptions of police misconduct: Race and neighborhood context. *Justice Quarterly*, *16*(4), 819-846.

Wennemark, I. (2011). *A diachronic study of the semantics of the adjectives mean and gay* (Unpublished master's thesis), Sweden/Linnaeus University: School of Language and Literature, Smaland, Sweden. Retrieved from http://www.diva-portal.org/smash/get /diva2:391541/FULLTEXT02

Willis, D., & Willis, J. (Eds.). (1996). *Challenge and change in language teaching (Handbooks for the English classroom)* (2nd ed., Vol. 1). Norwood, Australia: Macmillan Education Australia.

Yang, W. (2010). Brief study on domestication and foreignization in translation. *Journal of Language Teaching and Research*, *1*(1), 77-80.

YouTube. (2008). *Community guidelines*. Retrieved from https://www.youtube.com/yt/ policyandsafety/communityguidelines.html#communityguidelines-line-crossing

Ziegele, M., Breiner, T., & Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, *64*(6), 1111-1138.