UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Guilherme Henrique Resende de Andrade

On the Bias of Toxicity and Sentiment Analysis Methods on the African American English

Belo Horizonte
2023

Guilherme Henrique Resende de Andrade

On the Bias of Toxicity and Sentiment Analysis Methods on the African American English

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Flavio Vinicius Diniz de Figueiredo

Belo Horizonte
2023

Andrade, Guilherme Henrique Resende de

A554o On the bias of toxicity and sentiment analysis methods on the african american english [recurso eletrônico] / Guilherme Henrique Resende de Andrade — 2023.
1 recurso online (78 f. il, color.)

Orientador: Flavio Vinicius Diniz de Figueiredo.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Departamento de Ciência da Computação, Instituto de Ciências Exatas.
Referências: f. 73-78.

1. Computação – Teses. 2. Computação Social – Teses 3. Análise de Sentimentos – Teses. 4. Aprendizado de máquina – Teses. I. Figueiredo, Flavio Vinicius Diniz de. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III. Título.

CDU 519.6*04(043)

Ficha Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

On the Bias of Toxicity and Sentiment Analysis Methods on the African American English

GUILHERME HENRIQUE RESENDE DE ANDRADE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Flavio Vinicius Diniz de Figueiredo - Orientador
Departamento de Ciência da Computação - UFMG

Jussara Marques de Almeida Gonçalves

Profa. Jussara Marques de Almeida Gonçalves
Departamento de Ciência da Computação - UFMG

Fabrício Benevenuto

Prof. Fabrício Benevenuto de Souza
Departamento de Ciência da Computação - UFMG

Lesandro Ponciano

Prof. Lesandro Ponciano dos Santos
Departamento de Ciência da Computação - PUC Minas

Prof. Evandro Landulfo Teixeira Paradela Cunha
Faculdade de Letras - UFMG

Belo Horizonte, 23 de junho de 2023.

To Guilherme from the past, and to my family.

Acknowledgments

It has been a long path since I began my studies at Universidade Federal de Minas Gerais (UFMG). Indeed, I was impacted by so many people, in so many ways, that it would be an impossible task to name each one of them. However, some people definitely could not be left unnoticed.

First of all, I would like to thank my entire family for their support and motivation throughout this entire process. I am grateful to have believed it would be a possible and life-changing endeavor.

Secondly, I would like to thank my advisor Flavio for all the knowledge, patience, kindness, and friendship. Your situational comprehension made this a more enjoyable experience that helped me combine work and study. I also want to thank all the friends I have made throughout the way. In special, my friends from “Mestres dos Métodos”, and the Social Computing Laboratory (LoCuS).

Finally, I thank UFMG and every person who directly or indirectly was part of this learning process.

“*Gradually men will come to realise that a world whose institutions are based upon hatred and injustice is not the one most likely to produce happiness*”

(Bertrand Russell)

Resumo

A linguagem é um componente dinâmico da nossa cultura que evoluir quando em contato com diferentes tecnologias e partes da sociedade. Com o aumento crescente do acesso a internet, particularmente de grupos historicamente marginalizados, a Web propiciou a difusão e evolução de diferentes dialetos, por exemplo, o Inglês Afro-Americano (AAE). Contudo, a difusão de dialetos também trás barreiras de adoção quando os mesmos são aplicados online. Ainda no nosso estudo de caso acerca do AAE, salientamos que embora o número de frases com termos AAE online tenha aumentado ao longo dos anos, o dialeto também tem encontrado diferentes formas de censura online, sendo que no âmbito desta dissertação, a censura surge no forma de pontuações de modelos de análise de toxicidade e sentimento.

É essencial observar que a moderação do AAE por meio de modelos de análise de toxicidade e sentimento não surge deliberadamente. O número cada vez maior de postagens online torna difícil moderar a mídia online. Esse aumento no conteúdo levou as empresas a desenvolver ferramentas automáticas (por exemplo, modelos de análise de toxicidade e sentimento) para ajudar a filtrar conteúdo nocivo (tóxico, racista, agressivo e assim por diante). Nesse sentido, ferramentas de moderação foram criadas para fomentar debates online razoáveis (não tóxicos). No entanto, como discutimos nesta dissertação, essas soluções podem sair pela culatra e, em última análise, perpetuar o tratamento díspar (por exemplo, negligenciando o conteúdo discriminatório ou censurando o discurso da minoria) dos grupos sociais que pretendiam empoderar/impulsionar. Por que isso acontece? O uso de gírias e termos reapropriados (como n*gger) online é visto por tais modelos como conteúdo nocivo. Em sua forma atual, a IA não consegue diferenciar um enunciado tóxico de um não tóxico, dependendo da presença de termos-chave. De acordo com a ferramenta Perspective do Google, um enunciado como "Todos os negros merecem morrer com respeito. A polícia nos mata." é considerado tóxico. Visto que, "Os afro-americanos merecem morrer com respeito. A polícia nos mata." não é. Essa diferença de pontuação provavelmente surge porque a ferramenta é incapaz de entender a reapropriação do termo "n*gger". Para ser justo, a maioria dos modelos de IA são treinados em conjuntos de dados limitados e é mais provável que o uso de tal termo em dados de treinamento apareça em um enunciado tóxico. Embora essa possa ser uma explicação plausível, a ferramenta (se empregada em um site/fórum) cometerá erros independentemente da explicação.

Este trabalho investiga amplamente os vieses em outros modelos de análise de toxicidade (Perspectiva do Google e modelos do Detoxify de código aberto) e de sentimento (Vader, TextBlob e Flair). Nossos experimentos são realizados em dois conjuntos de dados baseados na Web (YouTube e Twitter) e um conjunto de dados baseado em entrevistas. Cada conjunto de dados tem expressões em inglês de afro-americanos e não-afro-americanos. Nossa análise mostra inicialmente como a maioria dos modelos apresenta vieses em relação à AAE na maioria dos conjuntos de dados, e os vieses são mais proeminentes no Twitter e menos proeminentes em entrevistas pessoais. Além disso, explicamos nossos resultados por meio de recursos linguísticos com o auxílio do software Linguistic Inquiry and Word Count (LIWC), Part of Speech (POS) de modelos de processamento de linguagem natural (NLP) e Word Mover's Distance (WMD). Apresentamos resultados consistentes sobre como o uso mais frequente de termos de AAE pode fazer com que o falante seja considerado substancialmente mais tóxico do que os não-falantes de AAE, mesmo quando falam quase sobre o mesmo assunto.

Palavras-chave: Computação Social, Toxicidade, Análise de Sentimentos, Aprendizagem de Máquinas.

Abstract

Language is a dynamic aspect of our culture that evolves when in touch with different technologies and parts of society. With the ever-growing access to the Internet, particularly for marginalized social groups, the Web has enabled the diffusion and evolution of different dialects, for example, African American English (AAE). However, this diffusion of dialects also finds barriers in the adoption of the same dialect online. Still in our case study of AAE, we point out that while the number of sentences with AAE terms online has risen over the years, the dialect has also found different forms of online censorship, wherein in the scope of this dissertation, censorship arises in the form of scores from toxicity and sentiment analysis models.

It is essential to note that AAE's moderation via toxicity and sentiment analysis models does not arise deliberately. The ever-increasing number of online posts makes it hard to moderate online media. This increase in content led companies to develop automatic tools (e.g., toxicity and sentiment analysis models) to help filter out harmful (toxic, racist, aggressive, and so forth) content. In this sense, moderation tools were created to foster reasonable (non-toxic) online debates. Nevertheless, as we discuss in this dissertation, these solutions may backfire and ultimately perpetuate disparate treatment (e.g., by neglecting discrimination content or censoring the minority's discourse) of the social groups they were intended to empower. Why does this happen? The usage of slang and re-appropriated terms (such as n*gger) online is viewed by such models as harmful content. In its current form, AI cannot differentiate a toxic from a non-toxic utterance depending on the presence of key terms. If you go online to Google's Perspective tool, an utterance such as "All n*ggers deserve to die respectfully. The police murders us." reaches a higher toxicity score than "African-Americans deserve to die respectfully. The police murders us.". This difference scores likely arises because the tool is unable to understand the re-appropriation of the term "n*gger". To be fair, most AI models are trained on limited datasets and the usage of such a term in training data is more likely to appear in a toxic utterance. While this may be a plausible explanation, the tool (if employed on a website) will make mistakes regardless of the explanation.

This work broadly investigates the biases in other toxicity (Google's Perspective and models from the open-source Detoxify) and sentiment (Vader, TextBlob, and Flair) analysis models. Our experiments are performed on two Web-based (YouTube and Twitter) datasets and an interview-based dataset. Every dataset has English utterances from both African-Americans and Non-African Americans. Our analysis initially shows how most models present biases towards AAE in most datasets, and biases are more prominent on Twitter and less prominent in in-person interviews. Also, we explain our results via linguistic features with the aid of the Linguistic Inquiry and Word Count (LIWC) software, Part-of-Speech (POS) tagging from Natural Language Processing (NLP) models, and the Word Movers Distance (WMD). We present consistent results on how a heavy usage of African-American English terms may cause the speaker to be considered substantially more toxic than Non-African American English speakers, even when speaking about nearly the same subject.

Keywords: Social Computing, Bias, Toxicity, Sentiment Analysis, Machine Learning.

List of Figures

List of Tables

Contents

Chapter 1

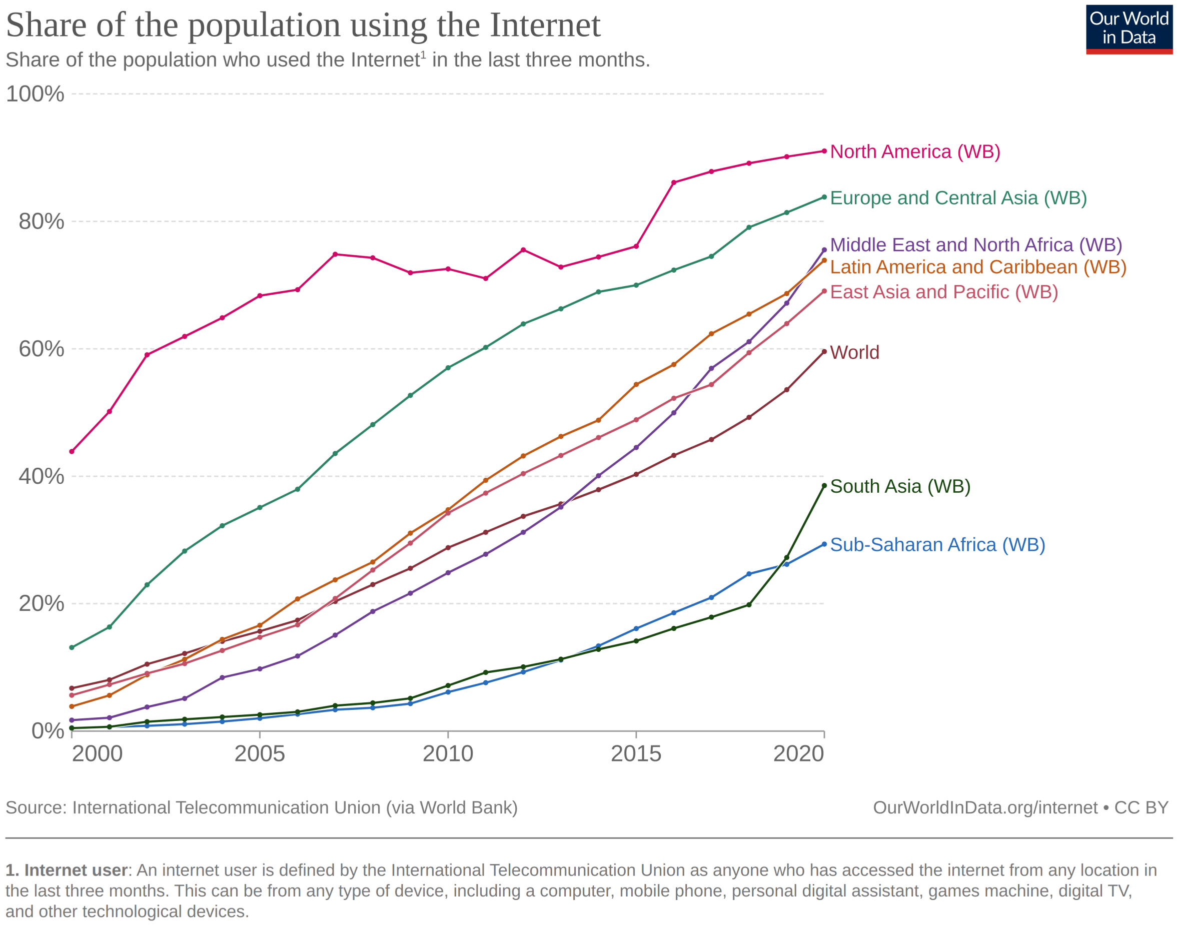
Introduction

In the last few decades, we have witnessed a substantial rise in Internet usage. According to [53], the number of internet users increased from approximately 400 million in 2000 to 4.7 billion in 2020. With this increase in usage, it is only natural that the Web enables a wide diversity of social groups to interact among themselves and with other groups. Nonetheless, the insertion into online environments was not equal throughout the years across different regions of the globe. For example, still according to [53], by the year 2000 nearly 40% of the North American population was using the Internet. In contrast, the remaining regions were only capable of reaching similar proportions after a decade, with some areas such as South Asia and Sub-Saharan Africa still figuring under such threshold up to the time of writing (See Figure 1.1). Despite enabling social interactions via rich media content (e.g., YouTube or Tik Tok videos), the exchange of written language is still one of the most common forms of interaction on the Web. For instance, approximately 500 million tweets are published each day according to Internet Live Stats[1].

The diversity in Web environments has been increasing over the years due to the insertion of different peoples and demographics. Since such applications foster a more open and dynamic form of speech, we began to see the emergence of the written form of dialects that previously were predominantly seen in the spoken form [9]. However, such massive amounts of textual data can make manual content moderation impracticable. In other words, the heavy usage of social media evidenced the urge for automatic moderation tools that measure and moderate improper behavior online. One of the main concerns is the public display of negative/toxic sentiments against a person or specific group, more drastically when the target is a minority group historically marked with discrimination and stereotypes, e.g. black community, LGBT, Muslims, disabled people, and so on. The visible necessity to deal with the increasing number of deviating content has led many researchers and companies to develop alternatives to identify such events [52].

Concurrently to the increase in Web usage, African-American English (AAE) has gone from being seen as a marginalized dialect of English to a consolidated vernacular of the language [25]. Despite starting to be officially taught in schools in the 1990s, there were some attempts towards mapping and understanding the social and linguistic

[1] https://www.internetlivestats.com/twitter-statistics/



Share of the population using the Internet

Share of the population who used the Internet[1] in the last three months.

Source: International Telecommunication Union (via World Bank)

OurWorldInData.org/internet • CC BY

[1] **Internet user**: An internet user is defined by the International Telecommunication Union as anyone who has accessed the internet from any location in the last three months. This can be from any type of device, including a computer, mobile phone, personal digital assistant, games machine, digital TV, and other technological devices.

Figure 1.1: Share of population using the internet according to each region of the globe.

features of African-American English back in the 1960s [39, 16]. Like most dialects, the African-American English initially was heavily used in spoken form and had the Web as a crucial influence on its emergence in the written form [9]. However, and as we have discussed, the Web is not only a disseminator of cultural aspects of our society but also a vehicle where toxicity campaigns against African Americans are prone to occur[2]. Even though most of the online forums and social networks have well-defined community guidelines, the partial anonymity and unaccountability leaves a lot of room to misbehavior. Consequently, a major concern online is the public display of negative/toxic sentiments against minorities.

The urge to deal with the increasing number of deviating content has led many researchers and companies to develop toxicity/sentiment analysis tools, such as Google's Perspective [33] and others [52, 20, 67]. These are the tools that help to determine what is proper and improper behavior online. Nevertheless, as previous research has discussed, automatic content moderation is likely to backfire and present biases towards

[2] https://theconversation.com/the-rings-of-power-is-suffering-a-racist-backlash-for-casting-actors-of-colour-but-tolkiens-work-has-always-attracted-white-supremacists-189963

Models	P_{SE_1}	P_{AAE_1}	P_{SE_2}	P_{AAE_2}	P_{SE_3}	P_{AAE_3}
Perspective	0.2396	0.7886	0.2546	0.4256	0.0406	0.2359
Textblob	0	0	0	-0.1666	0	0
Vader	0	0	0	0	0	0
Detoxify	0.0012	0.6145	0.8766	0.9718	0.0257	0.7601
Detoxify Unbiased	0.0013	0.6842	0.2549	0.8903	0.0162	0.5332
Flair	0.9830	-0.7296	0.9994	0.9992	0.9957	0.9111

Table 1.1: In the table above, Perspective, Vader, and Detoxify range from [0, 1] with a score equal to 1 being the worst-case scenario for each model. For the remaining ones, the scores range from [-1, 1], with -1 representing the most negative scenario. Here, P_{SE_1} stands for "All my friends on the porch and never in the house", whereas P_{AAE_1} stands for "All my n*ggas on the porch and neva ina house". Similarly, P_{SE_2} is to "You're white", as P_{AAE_2} to "You're black". Finally, P_{SE_3} is to "I can't forget you", as P_{AAE_3} to "Cant fuhgit you". As we can see, there is disparate treatment despite the sentences not being of any harm.

minorities [55, 30, 63, 11]. For instance, a tool for toxicity analysis may present high scores for non-toxic African-American English sentences for no apparent reason. As motivating examples, when we employ toxicity and sentiment analysis models to online text, it is quite easy to find problematic phrases when we employ slang terms such as *"n*ggas"*. In Table 1, we contrast three pairs of sentences that should reach similar levels of toxicity/negative sentiment.

Why does the problem arise? From a linguistic perspective, dialects may inherently manifest behaviors and cultural aspects of the groups in which they were created [5]. Terms such as "n*gger" are problematic since both the term and its variations have a historical pejorative usage[3]. Nevertheless, this very same term was re-appropriated by the black community in a way that its use ceased to be considered problematic when used by people inside the black community. If such a fine line between causal speaking and offensive discourse is problematic from a human perspective, from a computational perspective these interpretations serve as confounding factors to automatic content moderation tools. In other words, toxicity/sentiment analysis tools are usually developed using either manual rules or supervised machine learning techniques that employ human-labeled data to extract patterns. The disparate treatment embodied by machine learning models is usually a replication of discrimination patterns historically practiced by humans when interacting with processes in the real world. Due to biases in this process, a lack of context leads both rule-based and machine learning-based models to a concerning scenario where minorities do not receive equal treatments [23, 59, 1, 15].

In order to better understand this issue, we here present a broad-scale analysis of biases from both toxicity and sentiment analysis models against the African-American

[3] https://en.wikipedia.org/wiki/Nigga

English dialect. Our work explores the performance of six different models which are widely known and used by practitioners and companies. Also, our analysis focuses on four datasets where some demographic information is publicly available or indicative of a group. The models we study can be divided into toxicity (Google's Perspective [33], Detoxify, and Detoxify Unbiased [26]), and sentiment analysis (Flair [3], TextBlob [41], and Vader [31]) models, but also can be segmented into transformer-based (Google's Perspective, Flair, Detoxify, and Detoxify Unbiased), and lexical-based (Vader, and TextBlob).

Our discussion so far leads to the driving research question behind this dissertation, which is: *Is there a systematic bias on toxicity/sentiment analysis models towards deeming AAE speakers more toxic than non-AAE speakers in usual utterances?* The results show that biases are more prominent on online datasets, such as Twitter and YouTube, and less strongly but still present in spoken English interviews. Overall, our research shows AAE will likely suffer discrimination from moderation tools online routinely.

Even though there are dozens of tools for both sentiment and toxicity analysis [52, 20, 67], we point out that our goal in this dissertation is not to pin-point models with the best accuracy. Our focus is on showing a systematic tendency of AAE to be deemed as more toxic or negative by several approaches. The datasets where we show this issue range from online texts from Twitter [9, 10], spoken English datasets gathered by linguists [51, 35], and online single speaker English from YouTube movie reviews. The YouTube dataset (see Section 3.2) was a manual effort toward gathering and labeling data from different demographic groups developed throughout this research. This dataset was made available to the community as a means to improve the current, and yet to come, models.

The rest of this dissertation is organized as follows: Chapter 2 presents an overview of the research landscape on topics such as Natural Language Processing when applied to representation learning, toxicity/sentiment analysis techniques, as well as the biases that emerge from such techniques. In Chapter 3, we describe the datasets used in this dissertation. Chapter 4 outlines the pre-processing steps applied to the data before proceeding to further experimentation and analysis. Chapter 6 comprises the description of the experiments along with the following results. Finally, Chapter 7 presents an overview about the limitations, further investigations, and final conclusions.

Chapter 2

Background and Related Work

In this chapter, we present an overview of the literature on language models, focusing on the evolution of representation learning solutions. Representation Learning techniques are of utmost importance when developing text classification solutions. This importance derives from their ability to bring deeper contextual and semantic information into the representations. Subsequently, we also discuss the motivation behind sentiment analysis tools, available alternatives, their major strengths and shortcomings, and how toxicity relates to sentiment analysis. Finally, we discuss bias in machine learning methods, and how they are capable of negatively influencing individuals online and suppressing the discourse of minority groups.

2.1 Language Models

Humankind has evolved to be able to reason and communicate in terms of words and numbers. However, these communication abilities derive directly from a system based on the transmission of electric impulses among many chunks of neurons in our brains. Similarly, computers may be able to communicate their calculations and conclusions about some problem with natural language, but under the hood, their representations of the very problem are nothing more than a binary sequence. Consequently, when developing Natural Language Processing (NLP) models, we need to find ways to represent the data numerically, i.e. to embed the ingrained meaning and the word usage scenarios into the representation.

One of the first used approaches was to represent the words in terms of an N-dimensional vector, with N being the size of the whole vocabulary in the data. Each vector would be one-valued at the same position assigned to the word when considering the sorted vocabulary. Nonetheless, vocabularies usually assume considerable sizes, e.g. Oxford English Dictionary[1] has accounted (until 2020) for more than 600k different terms. This

[1] https://www.oed.com/

approach, named One-Hot Encoding, besides suffering from the curse of dimensionality, is also incapable of aggregating meaning and context to the words' representations. The Curse of Dimensionality is the problem of having so many dimensions in the data that sparsity becomes a problem and the differences among data points become insignificant to the point of being hard to differentiate them.

Efforts were applied towards building more complex representations [13, 45, 61], even though at higher computational costs. Two approaches with prominent performances were Bag of Words and Skip-Gram [43]. The main difference between these two methods lies in the fact that Continuous Bag of Words tries to predict a word given a surrounding context, and Skip-Gram tries to predict the surrounding context given a word. Mikolov et al. [43] proposed to use these resources in the training process of shallow neural networks so we can benefit from cheap computational costs, and at the same time add some kind of non-linearity in the words' relationships. This solution was named Word2Vec. Further improvements were made by Mikolov et al. [44] about computational costs and embedding quality, allowing the approach to also embed composed terms such as New York, Machine-Learning, and so on.

According to Pennington et al., [49], we can cluster representation learning methods into two families, namely, global matrix factorization, and local context window methods. The main drawback of the first family is the lack of semantic properties on the produced representations, whereas the latter falls short considering more general statistics from the data other than the local contexts. Taking this into consideration, the authors propose Global Vector Representations (GloVe), an approach that tries to instill global word importance into the training process along with a cheap computational cost that enables the model to be trained on bigger corpora and with less time. Such improvements allow GloVe to consistently outperform Word2Vec under the same conditions.

In a different direction, [50] proposes to consider the word embeddings as a function over the layers of a language model instead of a simple output from their last layers. This approach uses Bidirectional Long Short Term Memory models as an alternative to better grasp the gist of the nearby context. The designed language model has many layers that usually learn different aspects from the data, e.g. syntactic information is more present at lower layers whereas the semantics are better represented at higher layers. Hence, the final embedding is defined as different combinations - depending on the problem at hand - of all layers' outputs. One of the core achievements of this work was being able to reach considerable performances in downstream tasks, while also learning important syntactic, semantic, and polysemic characteristics.

The Transformer architecture was initially proposed at [62], and is an improvement of previous models about parallelization, long-range dependencies within the sentences, and performance. The vast majority of the previous representative solutions employed recurrent networks to engender meaningful context information into the representations, however, models such as [28] present a time-dependency that cannot be decoupled and parallelized, which causes them to incur in longer execution times. A Transformer, on the other hand, is highly parallelizable since it solely uses attention mechanisms distributed over many attention heads - along with self-attention - which are capable to relate and combine many characteristics from the sentence, including a representation of the whole sentence itself. Further improvements were made by [14] who proposed a bidirectional attention-based model pre-trained over different tasks and datasets. This new approach caused the subsequent tasks to perform considerably better in various contexts. Hyperparameters and optimization issues were also investigated at [40] causing the proposed architecture to better use the resources available in the training and fine-tuning phases.

Even though the last developments were capable of providing incredibly powerful representations of the words, they have also grown a great shortcoming: their sizes and huge computational costs during the training phase. Recently, [54] proposed a distilled version of BERT which happens to be 40% smaller, preserving around 97% of the knowledge and capabilities from its larger version, while being approximately 60% faster.

It is worth noticing that, despite not being necessarily a classification task itself, building better data representations is crucial to reaching higher performances in downstream tasks such as text classification due to the extra amount of information aggregated into the representation.

2.2 Toxicity and Sentiment Analysis

Sentiment Analysis is the task of identifying sentiments and quantifying their intensity in sentences. Given the increase in the amount of data in recent years, manual identification of deviating content has become impracticable. One way to tackle such a problem is through automated Sentiment Analysis tools which can be classified into two major categories, namely, machine learning-based and lexical-based.

The machine learning-based (ML) approaches [21, 64, 68, 57, 3] are built over a sample of data points comprising as many examples as possible from the groups to be learned. Usually, the learning procedure targets data drawn from a context of interest (e.g. Twitter, Facebook, Marketplaces, etc.) which can either have been previously labeled, or not. This family of methods often benefits from complex word representations and is capable of grasping deeper relationships implied in daily conversations.

Lexical-based methods [48, 29, 58, 31], on the other hand, begin by listing seed words considered to be representatives of groups of emotions. Once the seed list is complete, it is incremented with similar words and synonyms. Such approaches must actively deal with normal word usage that may change the intent/intensity of the sentence, such as negations, punctuations, capitalization, etc. Since this approach is based on the usual understanding/application of terms they may perform well in different scenarios other than the one they were first aimed to [48].

When compared to lexical-based methods, the machine learning approaches are usually capable of grasping deeper meaning, and relationships between words, due to the improvements in vocabulary representations such as that proposed in [44]. Nonetheless, the majority of the proposed solutions are based on supervised methods, which has the problem of heavily relying on considerable amounts of quality labeled data, a sometimes difficult-to-reach pre-requisite. On the other hand, lexical-based approaches need to explicitly address negations, punctuations, out-of-vocabulary occurrences, and more complex relations between terms [52]. To address the gaps left by each family of methods, there are also hybrid solutions [46, 60, 66]. For example, [60] gathers complementary lexicons with established sentiments while also implementing a manually developed set of rules, and, to improve the downstream analysis, it can also apply a machine learning procedure to learn better weights to the available terms.

In a wider sense if compared to sentiment analysis, there is also toxicity classification models [36, 33, 26]. Toxic speech is usually considered to be an umbrella term that comprises hate speech, abusive language, racism, and so on [27, 38]. Despite the efforts to address toxic speech, there is not a clear agreement about what means for a sentence to be toxic. One of the most established definitions is that presented by [18] defining toxicity as a rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion.

Due to the lack of consensus on the toxicity definition, there can be a lot of room for ambiguity when labeling sentences. The vast majority of datasets use human labelers which are influenced by their previous experiences and, in most of the time, do not have access to the underlying context where the respective sentence was drawn from. This subjectivity and lack of context may cause considerable labeling issues. For example, Kumar et al. [38] state that people who have suffered harassment in the past are more prone to label random sentences from some social networks as toxic than those that did not face such problems.

Maybe due to its less restrictive definition, and to the capacity of encompassing many types of harassment, toxicity models are actively used in practice to moderate discourse in many platforms[42, 47, 2], however, with some known bias problems.

2.3 Bias

As stated, our work is focused on assessing the differential treatment disparities across African-American English and Non-African-American English for toxicity and sentiment analysis models. Moreover, as we have also mentioned, such models are usually created either via manually labeled sentences or via different human-curated rules. This manual labor will significantly bias underrepresented groups. Consider, for instance, Natural Language Processing (NLP) overall. In the last few years, we have witnessed many considerable breakthroughs in the discrimination power of language models due to the improvements in word representation [43, 49, 62, 14]. However, the evolution path from simple shallow neural networks to deep attention models has come partially at the cost of large amounts of training data. Due to this direct dependence, whenever there is no inherent quality and group stratification in the available data, the resulting model will perpetrate previously practiced prejudices. This is the subject of many previous endeavors as we now summarize.

Starting from Jia et al. [32], the authors investigated the proportions in which men and women appeared in news articles' images. The authors found that men are considerably more frequently represented than women. Garcia et al. [19] also described a consistent bias towards men in Twitter content. That is, on Twitter, female users tend to describe more events in which men play important roles. Babaeianjelodar et al. [4] explored the nuances of gender biases over ML models. In all datasets considered, models perform disparately against unprivileged subgroups. Similar findings were raised by several other authors considering countries [24], age [15], religion [1], and sexual orientation [18]. Regarding dialects, Blodgett et al. [9] studied how language characteristics can change considerably within the same country. The work focuses on learning distinguishable features between Standard and Black English with a geographic context. In [22] the authors also present another clear differentiation between English focusing on Drag queens. Here, the authors find that drags have a speaking characteristic that is consistently seen as more toxic by ML models.

As stated by Bamman et al. [7], language is always situated within a context. Neglecting this surrounding context leads to disparate treatments. For example, the drags' way of speaking might be seen as toxic if used by someone outside the LGBT community. However, this may be a defense mechanism to cope with tough situations imposed by society [22]. Similar language signals are passive to be found within the black community and their dialect, the African-American English. Studies were already performed as an attempt to comprehend, and measure the extent to which ML models are biased against AAE speakers [9, 6, 55].

Overall, we can state that nowadays it is not hard to find discrimination episodes involving artificial intelligence systems[15, 1, 18, 30]. For example, Abid et al. [1] interacted with a conversational artificial intelligence model touching religion-related subjects and noting the inner associations with the topic. Finally, they found a consistent bias associating Muslims with terrorists (in 23% of the test cases), and Jews to money (in 5% of the test cases). In the opposite direction, efforts to comprehend and mitigate such biases [11, 18, 9, 4, 27]. Nevertheless, as studied by Gonen et al. [23], persistent bias may stick with the model even after active effort has been applied to remove it. Since ML model complexity has increased in the last few years, we could also expect the bias to be more elaborated and hard to fight against. This leads us to the problem of using biased models for sensible tasks that may perpetrate harmful behavior. Currently, sentiment analysis models are deliberately being used to moderate forum discussions of relevant news media, and magazines [2, 47, 34, 42].

Our research differs from previous works by investigating biases in models of different families (Transformer-based, and lexical-based methods) and throughout many datasets representing different contexts (in-person conversations, single-speaker movie reviews, and personal social media posts). We here focus on easy-to-use methods and those already being applied in real-world forums. With our investigation, we present some visualizations that help to understand the degree of disparate treatment between Standard English and African-American English speakers.

In the next section, we present broader descriptions of the datasets used to investigate biases in ready-to-use toxicity/sentiment analysis solutions.

Chapter 3

Datasets

In order to understand the extent of biases in toxicity/sentiment analysis models and when they present themselves more strongly, we explore three datasets of different natures. Initially, we make use of the TwitterAAE dataset [9, 10]. This dataset is interesting as it was manually labeled as to create POS tagging models for the African-American English (AAE) dialect. Thus, it enabled us to explore POS features to explain our results. Moreover, Twitter is one of the major platforms where one would expect that toxicity and sentiment analysis models could mitigate unwanted behavior. On the negative side, as the dataset contains general Tweets, it does not control for confounding factors such as dialogues, debates, and potentially controversial topics. Thus, we complement this research with two other datasets described next.

Our YouTube dataset is comprised of subtitles extracted from YouTube movie reviews with a single speaker, discoursing about a unique topic per video. The topics are movies from Rotten Tomatoes 100 Best Movies of All Time. We targeted single-speaker videos to control for any confounding variables that may appear with dialogues. Also, we focus on the top 50 most acclaimed films ever produced[1] to control for the possible negative influence of bad content. Each of these movies is highly rated and highly popular. Our goal with the YouTube dataset is to control both content and dialogue. It thus enabled us to present a more general view of biases from models.

Finally, we explore the Corpus of Regional African American Language (CORAAL) [35] and Buckeye [51] dataset as representations of spoken African-American English and

[1] https://www.rottentomatoes.com/top/bestofrt/

Dataset	Group	# Documents	# Sentences	# AAE Terms	AAE Terms Ratio
Youtube	Black	150	17828	18308	122.05
	White	484	41464	42729	85.67
Twitter	AAE	250	250	372	1.49
	SE	250	250	259	1.04
CORAAL	AAE	142	64493	61651	434.16
Buckeye	SE	39	19304	18712	479.79

Table 3.1: Datasets statistics. The *AAE Terms Ratio* represents the average number of African-American English terms per document in the corpus. Note that the number of documents may be slightly smaller than the one described in each dataset subsection. Such differences arise due to reading problems related to the documents.

Non-African American English, respectively. As the name states, CORAAL is focused on African-American English, whereas Buckey focuses on middle-class Caucasian speakers, stratified by sex and age (under thirty and over forty) from central Ohio. Using these two datasets, we were able to explore model behavior in spoken English.

In Table 3.1 we present a summary of our datasets in the number of sentences (or utterances), number of words (non-unique), and number of African-American English terms present. A lexical characteristic of African-American English is the usage of specific words and slang. We created a list of these terms based on the Black Talk Dictionary (see Section 4.4) to calculate the occurrences of terms regarding the dialect in the dataset (see Table 3.1).

Over the next few subsections, we dive into the details of each dataset, and how they were gathered.

3.1 Twitter

The Twitter dataset comes from the TwitterAAE[2] website. In order to create the dataset, the authors [9, 10] developed a Latent Dirichlet Allocation (LDA) based topic model that took into account both the frequency of common terms used in AAE, as well as Census data. That is, based on the location where the account tweeted from, an initial estimate of race would be possible. This information is combined with the presence of key terms in order to derive different latent topics for the corpus. These topics were then explored in order to label AAE and Non-AAE tweets.

Although the authors label over 80,000 tweets, in our analysis we focus on a smaller sample of 500 tweets that were manually inspected by the authors. These tweets were manually labeled with POS tags in order to derive an African-American English POS model. According to the authors, more than 18% of the terms used within the African-American tweets are not in the standard English dictionary. It is also very common to find words written in their phonological style in AAE - e.g. tha (the), iont (I don't), ova (over), and so on - while the contrary was found to never happen in the Non-AAE tweets. Overall, from this corpus we can expect to find posts spread over different demographic groups and geographic locations, and within a wide age spectrum.

[2] http://slanglab.cs.umass.edu/TwitterAAE/

3.2 YouTube

This dataset is a collection of subtitles from YouTube movie review videos in which there is a single speaker talking to the audience about a movie production listed among the most relevant movies of all time. We considered Rotten Tomatoes' top 100 best movies of all time ranking due to their prestige among the audience, and because they have a higher probability of being well-spoken in a review. For each of the top 50 movies from the ranking, we manually searched and cataloged as many videos as possible from four demographic groups, namely, Black Women, Black Men, White Women, and White Men in order of appearance when querying the movie name on YouTube. Since YouTube doesn't naturally disclose demographic information about its users, we had to restrict our search only to producers who happened to appear on the screen at least once throughout the entire video. The list of movies and the respective YouTube channels used in the construction of this dataset is available at this projects' GitHub [3].

When publishing videos on YouTube, the creators are free to either explicitly inform the subtitles to their videos or to let them be automatically captioned by the YouTube transcription model. Nonetheless, differently from manually informed subtitles, the captioning mechanism, by default[4], censors any potentially offensive terms with the special token [__], besides that, it also does not apply any kind of punctuation to the sentences. Even after addressing this limitation, for fair comparisons, we only considered automatically generated subtitles, even when manual transcriptions were available for a given video. Finally, it is important to state that transcriptions are not punctuated, an issue corrected using machine learning models for punctuation as described in Section 4.

Considering the observational nature of our study, an extensive effort was applied to control the confounding variables' effect on the conclusions. The selection of the most prestiged movies of all time was an attempt to reduce the chance of having negative reviews which in turn would comprise higher scores in the toxicity analysis. We also tried to find at least one single-speaker video review for every movie, to reduce any sampling bias impact. More importantly, the first-person nature of the reviews helps to eliminate the possibility of other people's opinions influencing the argumentative paths. We also believe that a bigger variety of content producers within a given demographic group reduces the influence of a single person on the conclusions. It is worth noticing that since the identification of race and gender is subjective, we have no guarantees on the demographic group assigned to the content producers based solely on their visual appearance. Besides, we also cannot make assumptions about the geographical location of the individuals.

[3] https://github.com/Guilherme26/dissertation/blob/main/data/youtube_data_description.csv

[4] https://support.google.com/youtube/thread/70343381

In conclusion, such efforts allowed us to build and make available a dataset comprising subtitles from 637 YouTube videos, with 45 from Black Women, 106 from Black Men, 171 from White Women, and 315 from White Men. Out of those videos, we total of 30 distinct Black Women creators, 57 Black Men, 84 White Women, and 177 White Men.

3.3 CORAAL

The Corpus of Regional African American Language [35] is a long-term corpus developed and maintained by the University of Oregon with the support of the National Science Foundation. The dataset is comprised of more than 150 socio-linguistic interviews with African-American English speakers born between 1891 and 2005. The dataset contains the orthographic transcriptions of interviews, together with the person's age, gender, and city they live in. Thus, each interview from the corpus encompasses many subjects from a given city/community.

Differently from the YouTube data, the transcriptions available here represent the sentence in its entirety, accounting for complete punctuation, line-level notes, and even non-linguistic sounds. Beyond that, the data also tracks the interviewer's voice in the dialog. The interviews allow the speakers to talk freely about different topics, an interesting feature that emulates the diversity of daily interactions and mood variations.

In its last version, the dataset aggregates five major sub-corpora from different locations of the United States of America, namely, Atlanta (2017), Washington (1968, and 2016), Lower East Side (2009), Princeville (2004), Rochester (2016), and Valdosta (2017). The number of speakers within each one of the mentioned areas is, respectively, 13 (5 women / 8 men), 116 (54 women / 62 men considering both releases), 10 (5 men / 5 women), 16 (9 women / 7 men), 15 (9 women / 6 men), and 12 (6 women / 6 men), respectively.

3.4 Buckeye

The Buckeye [51] corpus is an effort started in 1999 and supported by the National Institute on Deafness and Other Communication Disorders, as well as the Office of Research at Ohio State University. The initial goal was to gather approximately $300,000$

words of speech conversation from central Ohio speakers, keeping track of time and phonetic information. To reach that objective, researchers selected a group of 40 middle-class Caucasian speakers, stratified by sex and age (under thirty and over forty).

Similar to the YouTube dataset, Buckeye sentences are not punctuated. However, instead of automatically generated captions, this corpus employed transcribers who were explicitly instructed not to use punctuation of any kind within the utterances and also, not to try to correct possible speech “errors”.

Different from CORAAL, Buckeye does not present heterogeneity with regard to age and gender. Nevertheless, given that the dataset focuses on Caucasian speakers from a small geographical region, it enabled us to contrast results with CORAAL. Thus we capture AAE (CORAAL) and Non-AAE (Buckeye) interviews.

Chapter 4

Data Preprocessing

To present an analysis with fewer biases, our goal is on assessing models with as little influence of external variables as possible. To do so, we clean up our datasets in order to have a similar setup. That is, before feeding sentences to the models, we clean up and process the dataset to get punctuated sentences in a standardized fashion.

We also employ different tools such as Linguistic Inquiry and Word Count (LIWC) [48], Part-of-Speech (POS) taggers, and AAE dictionary as tools to interpret our findings. Details on these steps are now presented.

4.1 Data Segmentation and Cleaning

As exposed in Chapter 3, with the exception of CORAAL, the datasets we explore are either free-form texts from Twitter or were not transcribed following correct orthographic rules. This is particularly the case for YouTube and Buckeye, where sentences were not segmented according to their inherent meaning, but to silent intervals (not necessarily long ones) after a continuous pronunciation of words. This comes from the nature of the dataset (e.g., YouTube data contains closed captions). Given that such segmentation is not necessarily aligned with the correct punctuation, the meaning of the resulting sentences may be misrepresented. To reduce the impact of incorrect segmentation in our analysis, we employed a machine learning-based segmentation to YouTube and Buckeye datasets. Twitter is an exception given that tweets are self-contained messages. Finally, since CORAAL was correctly segmented by humans, in an attempt to standardize the semantic segmentation with the contrasting dataset (i.e. Buckeye), we completely ignored the original segmentation and also employed the same machine learning segmentation to CORAAL.

The segmentation task was performed with NVIDIA's Punctuation and Capitalization model made available at NeMo Toolkit [37]. This solution is essentially a composition of token-level classifiers on top of a pre-trained language model, by default a Hugging-

Face[1] Transformer. Below we can see the same text snippet in its original form, and then segmented with the aforementioned tool.

Original Snippet

Um oh i was very naive on that i mean i mean you see it on the news and you but the actual relationships all uh but again do you need to know that unless it pertains to you and then even if it does pertain to you right right i think uh i didn't know the difference so it didn't affect me at all it's just they're just another person um i think more uh as i'm being exposed to patients with the different lifestyles and knowing how to educate their caregivers and what they need um i felt somewhat uh ill equipped the first few times but um knowing what the expectations are especially like with spinal cord injury patients and when you go into sexuality and teaching them after injury i mean as a nurse you need to be able to educate them i felt pretty inadequate but the uh that was even through nursing school i mean those weren't things that were discussed [...]

Segmented Snippet

Um, Oh, I was very naive on that. I mean, I mean you see it on the news, and you, but the actual relationships. All? uh, but again, do you need to know that unless it pertains to you, and then, even if it does pertain to you right right, I think. Uh, I didn't know the difference, so it didn't affect me at all. It's just they're just another person. Um, I think more, uh, as I'm being exposed to patients with the different lifestyles and knowing how to educate their caregivers and what they need, Um, I felt somewhat uh, ill equipped the first few times, but um, knowing what the expectations are especially like with spinal cord injury patients, and when you go into sexuality and teaching them after injury, I mean as a nurse, you need to be able to educate them. I felt pretty inadequate. But the uh, that was even through nursing school. I mean, those weren't things that were discussed. [...]

4.2 Swear Word Identification

In our results, we found that the presence of swear words is a major factor that drives models toward scores of either more toxicity or fewer sentiments. For example, the use of the term "*f*cking*" in the sentences can drastically change the toxicity scores

[1] https://huggingface.co/

even when not applied negatively, e.g., *I f*cking love you*, and *I love you* when assessed by Perspective API results in, respectively, 0.4712 and 0.0255 scores. Given that swear words do not necessarily indicate toxic behavior, we decided to analyze sentences both with and without swear words. This choice was made to make sure that our results did not stem directly from the presence of such words, but from other factors we explored. The words were taken from the No Swearing project[2], a cooperative effort to help programmers remove unwanted language from their applications. At the time of writing, there were 363 curse words listed by the project.

4.3 Linguistic Features

The most relevant aspect of our analysis derives directly from linguistic features drawn from the available sentences. There are different ways to analyze text, however, this research focuses only on word classes, or Part-of-Speech (POS), (e.g., Verb, Noun, Adjective, etc.) and language dimensions that represent psychological aspects of communication (e.g., Anger, Hate, Happiness, etc.).

The word class analysis takes into consideration the function of each token in the sentence in which it is applied. The word *smile* can be considered a verb, however, it can also be considered an adjective when used in certain scenarios as in " *The smiling baby is really cute*". This information can help us to understand the composition of the sentence in terms of its word classes. To classify the tokens according to their POS categories, we employ a black-box model [3].

The language dimensions, on the other hand, help to understand the intended meaning behind the uttered sentence. In this case, a single token can be assigned to many suitable categories. For example, the word *cried* is a 10-categories term (i.e., Affect, Positive Tone, Emotion, Negative Emotion, Sad Emotion, Verbs, Past Focus, Communication, Linguistic, and Cognition). Hence, a single sentence can be seen as a counting vector where each dimension is the number of occurrences of the respective language dimension within the given utterance. For this study, we employ Linguistic Inquiry and Word Count (LIWC) software [48] in its 2015 release. The Linguistic Inquiry and Word Count tool is a research effort aimed at mapping psychological features (i.e., language dimensions) of speech that uses lists of seed words and their synonyms to identify the language dimensions aggregated to each term. More on the whole building procedure applied to LIWC's project, as well as the complete list of LIWC language dimensions can

[2] https://www.noswearing.com/
[3] spacy.io

be found in [12].

To complement the linguistic analysis mentioned above, we develop a counting of AAE terms with the aid of Black Talk Dictionary described in Section 4.4. Finally, it is worth noticing that LIWC has a language dimension used to categorize swearing. However, since LIWC allows a single term to be assigned to different language dimensions, we believe that many tokens could be misrepresented as swearing (even though not majorly used to such purposes), causing the swear word filtering based on LIWC to be more aggressive. For example, the word *bloody* can be assigned to *body* and *health* categories, however, could also fit into *informal* and *swear*. Due to such behavior, we decided to apply a set of tokens majorly recognized as swearing as referenced in Section 4.2.

4.4 African-American English Terms

To identify and count African-American English terms occurrences we employ the Black Talk dictionary[56]. This dictionary is an active effort to compile written language surveys along with terms mostly used in spoken format among African-American English speakers. Since African-American English first emerged as an oral language, the main intent of this dictionary was not to define the etymological history of terms, instead, it concentrates on the words/sentences meaning and significance for the speakers' community.

There are a few other alternatives to Smitherman's dictionary [17, 8, 65]. The choice of using [56] instead of the aforementioned references is due to its more general nature (it encompasses more than just slang), and a more recent revision if compared to the remaining works.

This dictionary comprises more than 1800 entries. Since some entries are sentences instead of single terms, they may be applicable to different pronouns. In such cases, the possible subject use cases are listed. For example, "BREAK HIM/HER/THEM OFF SOMETHING". Our transcription of the entries considers every possible combination presented. The entire list of terms and expressions can be found in Appendix A.

Chapter 5

Methodology

This research follows a structured methodology that helps to visualize possible biases against some demographic groups. Considering we want to analyze and understand how language models treat disparately different parts of society, the datasets must comprise the demographic categories of the utterances. Here we seek to develop complementary visualizations taking into consideration the possible confounding variables comprised in any observational study.

Our methodology aims to isolate the impact of the usage of curse words. The same approach is taken to analyze the sheer impact of using African-American English terms. Posteriorly, we analyze how does linguistic features, and also AAE terms, impact the overall toxicity of sentences in each demography. Finally, we employ the Word Movers Distance in the process of finding sentences with similar meanings in order to compare their toxicity/sentiment scores. Below we provide a more detailed description of the experiments and their shortcomings.

Score Distribution In this experiment, we intend to have a general understanding of the toxicity and sentiment scores' distribution of both African American English (AAE), and Standard English (SE) sentences. A major concern here was to consider the distribution of scores in their raw form (the entire dataset), and after removing curse words. This visualization helps to comprehend the disparate treatment between the groups and also points to whether or not the differences are motivated by a more intense use of curse words.

Impacts of African-American English Terms Besides the influence of curse words, we also want to understand the impact of having African-American English terms within the sentences. As referenced in Section 2, historically, some of these terms are strictly associated with negative meaning when used by people outside the community historically marked with such discrimination. These experiments show the biases against the uses of African-American English terms independently of the demographic group.

Logistic Regression Logistic Regressions are statistical methods capable of drawing complex functions to describe a data distribution in terms of their variables. Usually, this framework enables the calculation of numeric coefficients that express the importance of variables within the process, as well as their aggregated p-value.

To execute this experiment we initially construct semantic and POS tagging features from the sentences, along with information about race and the number of African-American English terms. After feature engineering, we fit one Logistic Regression for each toxicity/sentiment analysis model having their scores as the response variable. With this experiment, we intend to statistically understand the relevance of each feature to the process of predicting their scores and also get the statistical relevance of the outputted coefficients.

Word Movers Distance Word Movers' Distance (WMD) is a technique used to measure the distance between sentences in a latent space generated by methods such as [44, 49, 50]. The WMD technique calculates the similarity of two sentences based on their location in a latent space relative to the semantic meaning of the words comprised in the training corpus. Since words with related meanings tend to fall within the same vicinity in the resulting embedding space, sentences do not need to necessarily have words in common to present some similarity in meaning and a valid distance score.

As we know, we have contrasting groups (African-American English and Standard English) of sentences. However, the direct comparison of two sentences (one from each group) may not be so fair since they potentially differ in meaning. To control this effect, and make sure sentences of different groups have similar semantic values, we use WMD.

In the experiment, we, for each sentence in the African-American English group, select the sentence with the closest meaning in the Standard English group. After defining the pair of sentences, we calculate the difference in their scores and visualize the distribution of differences per WMD score. The visualizations allow us to assert whether African-American English sentences are treated differently than SE sentences when representing similar content.

Our results are presented in four major steps. First, we analyze the score distribution of each model with and without sentences containing curse words. Second, we contrast the score distributions of sentences with and without African-American English terms. Third, through the usage of Logistic Regression models, we analyze the relationships of LIWC, POS Tagging, African-American English terms, and race attributes with the scores from each model. Lastly, we assess the difference in toxicity/sentiment scores between the sentences taking into consideration, for each sentence of a demographic group, the semantically closest utterance of the comparison group. We decided to present the results below in a way that the first three methods regard sentiment analysis models, whereas the last three, are toxicity models.

5.1 Limitations

Our methodology is developed upon observational data. The data from observational studies, by nature, is not generated following a well-defined structure. To make such datasets useful for certain purposes, we need to investigate the underlying behavior in the data while controlling for the possible confounding variables. Consequently, to make the datasets as comparable as possible and posteriorly draw valid assumptions, we need to clearly define the contexts in which data was created and become aware of environmental features. However, most public datasets available today do not comprise enough information to control completely the confounding variables, posing an inherent limitation to every analysis performed on such data.

One important aspect to be aware of when analyzing datasets is eventual sampling biases. To soften the impacts of sampling biases we have chosen to assess machine learning methods in some datasets built upon different social media and populations. The analysis depends on contrasting different demographic groups, however, not every dataset is capable of differentiating between some demographic features. Most of the time, the absence of such information can be beneficial if we take into consideration the potentially harmful downstream applications that could be built, for example, a credit loan machine learning model that receives race as one of the features and regards such variable in the decision process. Nonetheless, it also handicaps well-intended researchers and companies. We overcome these limitations by synthesizing information with the available data. Consequently, the final results are likely to differ slightly if compared to the analysis performed with the real demographic labels.

Another important limitation in our analysis regards non-exhaustive lists used to identify curse words and African-American English terms. Besides not encompassing every possible occurrence available, the terms' written form can have typos or even be written slightly differently causing the procedure to not be able to count every occurrence. On the other hand, we also have to be aware of counting terms that are not good representatives of the group being measured. For example, "gay", "hell", and "nigga" are not necessarily curse words in the same way "african american", "bear", and "peace' are not exclusively African-American English terms.

Another important aspect to take into consideration is the inherent nature of Logistic Regression models used in the experimentations. Despite being a useful technique to understand the influence of the features in the prediction of the response variable, we need to be aware of its inability to unveil more complex relations between the variables. In other words, we need to know that the nature of the model is restricted to the domain of linear interactions between the features.

Finally, we have to be aware that Word Mover's Distance is a method built on top of embedding models and, as described in Section 2.1, they are also subject to some of the biases described throughout this research effort. For example, some professions may be more frequently associated with men than women and, consequently, this behavior pattern may cause misleading comparison scores.

Chapter 6

Results

The experimental setup aims to gather complementary results in a way to evidence the inherent biases in the written form of African-American English (AAE), while also reinforcing that such biases are not necessarily bound to the use of curse words. However, they may be tightly linked to the use of African-American English terms. It is worth noticing that every Complementary Cumulative Distribution Function (CCDF) presented below reaches statistical significance in the Kolmogorov-Smirnov's Test with p-value < 0.05, except when the plot is explicitly marked with a red symbol. In other words, the samples are likely to have come from different distributions.

6.1 CORAAL vs. Buckeye

In Figure 6.1 we can see the comparison between African-American English and Standard English sentences through the Complementary Cumulative Distribution Functions (CCDF) of scores from each demographic group. The x-axis of the figure shows the score, whereas the y-axis captures the fraction of sentences for that demographic group with scores greater than the value on the x-axis, i.e. the empirical probability.

These visualizations help to outline the scores' behavior from the demographic groups. To analyze the overall behavior of the scores **we have to consider the value interval from each model**. Whenever the scores range from [-1, 1], we have to pay attention to the curve that stays predominantly under the other one through the [-1,0] interval. This means that the group at hand is seen more negatively than the other one. The same behavior when found in the [0,1] interval means that the group is seen less positively. Nonetheless, for any model with scores ranging in [0,1], the curve of a group that appears consistently above is seen more negatively.

For this dataset, we can see a slight tendency of toxicity models to consider African-American English sentences more toxic than Standard English sentences. However, the same tendency is not seen in sentiment analysis models. Instead, they tend to present

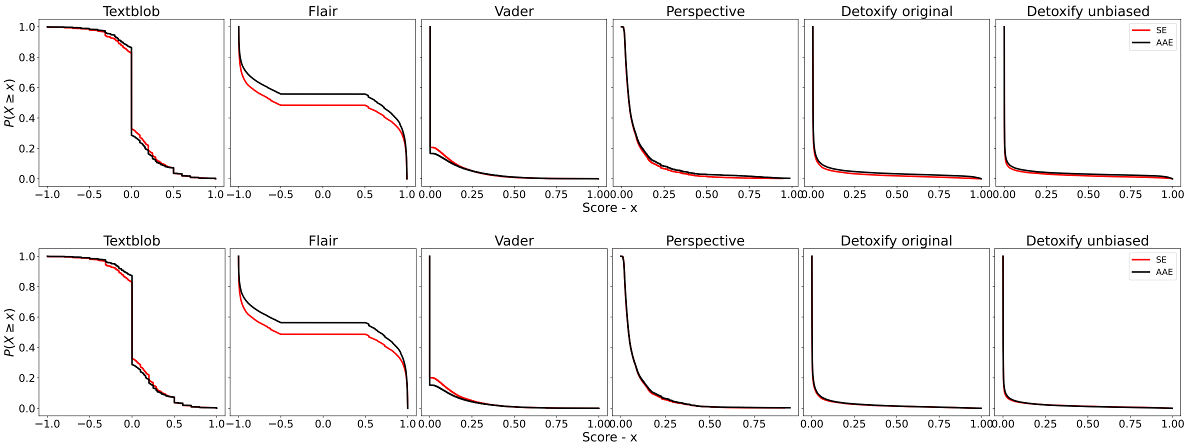


Figure 6.1: Scores distributions from CORAAL/Buckeye dataset with (above) and without (below) sentences containing curse words.

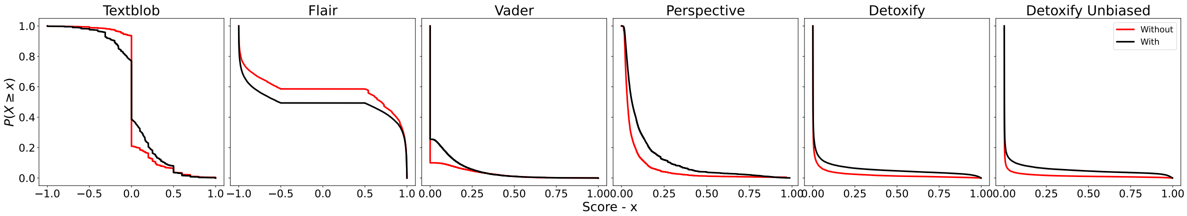


Figure 6.2: Score distributions for sentences with and without African American Terms in the CORAAL/Buckeye dataset.

biases against Standard English speakers, here represented by the sentences from Buckeye. As we can see, curse word usage seems to have little to no effect on the scores distributions.

To better understand such an effect, we developed a deeper investigation and verified that, instead of what we previously thought, Buckeye does not appear to be a great representative of Standard English due to its intense use of African-American English tokens. Table 3.1 we see the Buckeye dataset presenting a higher AAE Terms Ratio than CORAAL's. This has the potential to narrow the differences in the distributions of the scores by aggregating the biases of the minority groups into the Standard English group. Indeed, this behavior seems to be more evident with Flair scores.

When considering solely the impact of African-American English terms, we can see a generalized trend against African-American English speakers. Due to a higher zeroed-score rate for sentences without AAE terms, Textblob displays an odd behavior in this data. As we can see in Figure 6.2, there is a visual tendency in the model to output more negative scores, but also to output more positive scores to sentences with African-American English terms than to those without them.

Like the CCDFs presented previously, the coefficient analysis must be divided into models with scores ranging within the [-1,1] interval, and the ones in [0,1]. In the tables that follow, the first category of models will be visually marked in red, whereas the second marked in blue. For the models marked with the red color, a negative coefficient means a direct relation to negative sentiments, whereas for blue cells, a positive coefficient represents a direct relation to toxicity/negative sentiments. To improve our visualization, we only highlight the five most important coefficients along with the African-American English terms count. The coefficients are only presented when they reach a minimum p-value threshold.

In Table 6.1, we can see the predominance of LIWC features, with greater prominence of Swear, Sexual, Death, Body, and Anger. Each one of the features manifests a direct relation with higher toxicity and negative sentiment scores. Here, the African-American English terms count feature influences positively the negativity/toxicity in every model, with Flair assuming the higher coefficient. Everything is in line with the information presented in Figure 6.2.

Features	**Textblob**	**Flair**	**Vader**	**Perspective**	**Detoxify**	**Detoxify Unbiased**
LIWC_SWEAR	-	-	**0.8555**	**3.0254**	**4.0274**	**4.0881**
LIWC_SEXUAL	-1.0563	**-3.0443**	-	**0.9234**	**1.174**	**0.9254**
LIWC_DEATH	-	-1.7256	**1.1814**	**0.7102**	**0.7121**	**0.6416**
LIWC_BODY	-1.0645	**-3.3617**	0.1652	**0.6734**	**0.8216**	**0.4898**
LIWC_ANGER	-	-1.0377	**0.2716**	**0.6184**	**0.7269**	**0.8434**
LIWC_NEGEMO	-1.0261	-1.8000	**0.7942**	0.4245	0.2816	0.3007
AAE_TERMS_COUNT	-	-0.9432	-	0.3291	0.1754	0.2126
LIWC_SAD	0.5181	-	0.1442	-0.2251	-0.1433	-0.1722
LIWC_SEE	-0.2984	-	-	0.1787	0.1971	0.2505
LIWC_REWARD	**1.5792**	-	-	0.1583	-	0.0669
LIWC_WORK	-	-	-	-0.1219	-0.0867	-0.0685
LIWC_INGEST	**-1.1640**	**-2.7596**	-	0.1127	0.1753	-
LIWC_POWER	1.0633	-	0.0651	0.1124	-	-
LIWC_FOCUSPAST	-	-1.2460	-	-0.0993	-0.1475	-0.1197
LIWC_TIME	-	-	-	-0.0986	-0.0701	-0.0849
LIWC_DRIVES	-1.0721	-	-	-0.0927	-	-0.0827
LIWC_DISCREP	-	-0.7990	0.0624	0.0817	-	0.069
LIWC_ANX	0.5183	-	0.2125	-0.0783	-0.1628	-0.1625
LIWC_HEAR	-0.1722	-0.7780	-	-0.0770	-	-
LIWC_HEALTH	-1.0454	**-2.5510**	0.1684	-	0.1515	-0.0911
LIWC_BIO	**1.1858**	**2.3735**	-0.0716	-	-0.1228	-
LIWC_TENTAT	-	-1.6149	0.0768	-	0.0756	0.0642
LIWC_POSEMO	**1.7323**	1.3032	**-0.4082**	-	-	-
LIWC_AFFILIATION	**1.0946**	1.0698	-	-	-0.1133	-
LIWC_FOCUSPRESENT	-	0.8681	-0.0553	-	-	-0.0633

Table 6.1: Logistic Regression coefficients for the CORAAL/Buckeye dataset with p-value < 0.01. The five most relevant coefficients of each model are presented in bold, whereas not statistically significant coefficients were omitted. Note: due to score ranges, the negative coefficients in the red cells present a positive influence in toxicity/negative sentiment.

As described in 5, the Word Movers Distance (WMD) is a technique based on the comparison between word embeddings that helps to identify sentences with similar meaning, even in scenarios where they do not have a single word in common. In this experiment, for each sentence in the African-American English corpus, we find the sentence with the closest meaning in the contrasting group and then subtract their scores. Since they both have similar meanings, they both should receive similar scores. Since

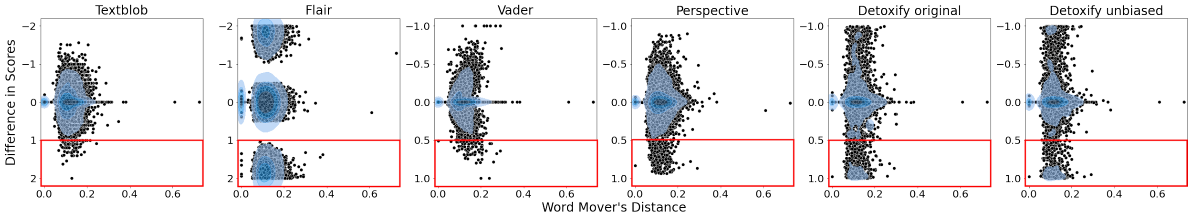


Figure 6.3: Distribution of the differences in the scores for the sentences with the closest meaning according to Word Mover's Distance in the CORAAL/Buckeye dataset.

the contrasting sentences have similar meanings, the subtraction of their scores should result in zero. However, in practice, it can result in three scenarios, that is, bias against Standard English speakers (a negative number), bias against African-American English speakers (a positive number), and no bias (zeroed result). Hence, in these visualizations, the closer a point is to zero on the x-axis, and the more positive it is on the y-axis, the bigger the bias against African-American English speakers.

In the previous subsections, we have already seen that African-American English terms cause sentences to reach higher negative scores. When analyzing sentences of similar meaning in Figure 6.3, this behavior appears to emerge more drastically within toxicity models (higher points density with the red square) which may point to a more stable judgment by sentiment analysis models over sentences spoken in conversations taken in person. As we can see, there are many cases where every model presents disparate treatment against both groups of speakers. Furthermore, we have also verified that lexical-based models appear to have a more egalitarian behavior, i.e., the differences in scores are less acute.

6.2 YouTube

In this dataset, we see a stronger tendency in the models, if compared to the previous datasets, towards seeing African-American English sentences as more negative/toxic than those from Standard English. In Figure 6.4 we verify that every model, with the exception of Vader, presents at least a slight deviation in behavior against the African-American English. Once again, lexical-based models (Texblob and Vader) appear to have a more balanced behavior, with Vader even presenting a small bias against Standard English speakers. As in the previous dataset, here we also see that the models are not so sensible to the use of curse words.

As expected, the distributions of the scores are negatively impacted by the use

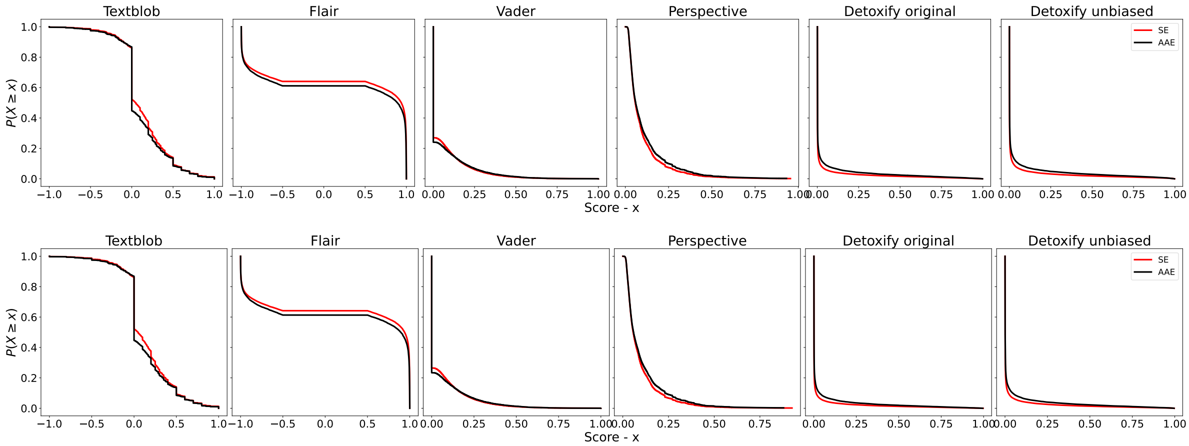


Figure 6.4: Scores distributions from YouTube dataset with (above) and without (below) sentences containing curse words.

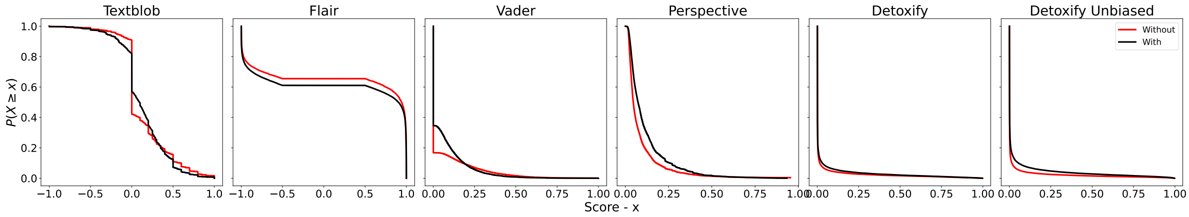


Figure 6.5: Score distributions for sentences with and without African American Terms in the YouTube dataset.

of African-American English terms. Even though this dataset appears to display less biased distributions, in Figure 6.5 we can visualize a clear tendency to deem sentences with African-American English terms more toxic than those without them. As seen in the previous dataset, here we also see a generalized behavior toward considering sentences with African-American English terms more negative/toxic than those from Standard English.

In Table 6.2, we again see a major agreement when assessing coefficients, for both toxicity and sentiment analysis models. Here we also see the exclusive prominence of LIWC features Swear, Death, Sexual, Body, and Negative Emotions. In the vast majority of cases, we see the features mentioned above presenting a direct relation with negativity/toxicity. Furthermore, we also see the African-American English terms positively influencing negativity/toxicity scores on Vader, Perspective, and Detoxify Unbiased.

When dealing with sentences of similar meaning, every model displays many scenarios of disparate treatment against both groups (African-American English and Standard English). However, in Figure 6.6 we see an emerging pattern of discrimination over African-American English sentences by Detoxify (both models) and Flair. As in the previous dataset, Vader appears to have a more balanced judgment than the remaining models, i.e., the sentiment scores' differences tend to be higher when the sentences are not so similar. TextBlob and Perspective also appear to behave similarly to Vader, however,

Feature	Textblob	Flair	Vader	Perspective	Detoxify	Detoxify Unbiased
LIWC_SWEAR	-	-	0.2137	**1.4529**	**2.102**	**2.1819**
LIWC_DEATH	0.3438	**-3.1559**	**1.8787**	**1.0922**	**1.5901**	**1.3711**
LIWC_SEXUAL	-0.4023	**-2.1333**	0.0712	**0.7733**	**0.7051**	**0.8716**
LIWC_BODY	**-1.6790**	**-2.4150**	0.0782	**0.6038**	**0.5397**	**0.4376**
LIWC_NEGEMO	**-1.2651**	**-3.3423**	**0.9376**	**0.5267**	0.2902	0.3222
LIWC_ANGER	0.2931	-	**0.3694**	0.3485	0.2662	**0.4039**
LIWC_FILLER	-	-1.4227	-	-0.2525	**-0.3458**	-0.2979
AAE_TERMS_COUNT	-	-	0.0617	0.2488	-	0.1334
LIWC_ASSENT	0.2790	-0.7194	-0.0938	-0.2394	-0.3079	-0.2824
LIWC_NONFLU	-	-	-0.0686	-0.2303	-0.3301	-0.2987
LIWC_INFORMAL	-	-	0.0775	0.2236	0.3146	0.2887
LIWC_SAD	0.4311	-	0.2289	-0.1992	-0.1175	-0.1392
LIWC_INGEST	**-1.1896**	-1.5451	-	0.1893	-	-
LIWC_RELIG	-	-0.9880	-	0.1764	0.2261	-
LIWC_NETSPEAK	-	-	-	-0.1568	-0.2241	-0.2578
LIWC_SOCIAL	-	-	-	0.1381	0.1342	0.0701
LIWC_WORK	-	-	-	-0.1141	-0.0878	-0.0494
LIWC_FEEL	-0.3244	-1.1777	0.0791	0.1124	-	-
LIWC_HEALTH	**-1.8130**	-2.0063	0.1785	0.1099	0.1295	-
LIWC_FOCUSPAST	-	-1.5012	-	-0.0815	-0.115	-0.0778
LIWC_DISCREP	0.2604	-	-	0.0798	-	0.071
LIWC_AFFILIATION	0.7533	1.2576	-	-0.0753	-0.097	-
LIWC_NEGATE	-0.2661	**-3.7329**	**0.4872**	-	-	-
LIWC_RISK	0.4176	-1.2727	**0.7053**	-	-	-
LIWC_ACHIEV	**1.3981**	0.9477	-	-	-0.0507	-

Table 6.2: Logistic Regression coefficients for the YouTube dataset with p-value < 0.01. The five most relevant coefficients of each model are presented in bold, whereas not statistically significant coefficients were omitted. Note: due to score ranges, the negative coefficients in the red cells present a positive influence in toxicity/negative sentiment.

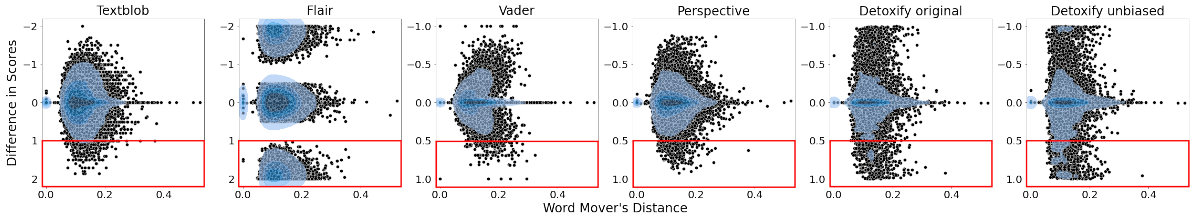


Figure 6.6: Distribution of the differences in the scores for the sentences with the closest meaning according to Word Mover's Distance in the YouTube dataset.

with more acute differences in scores.

Note that even though some models appear to be more well-behaved than others, the scenarios in which the disparate judgments arise can be harmless to a group while being extremely nocuous to the other group due to historical discrimination events.

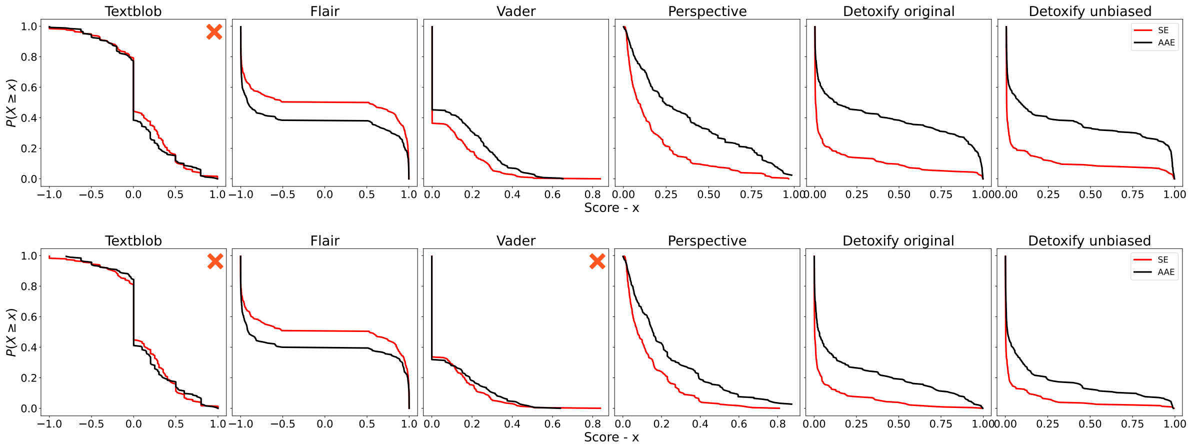


Figure 6.7: Scores distributions from Twitter dataset with (above) and without (below) sentences containing curse words.

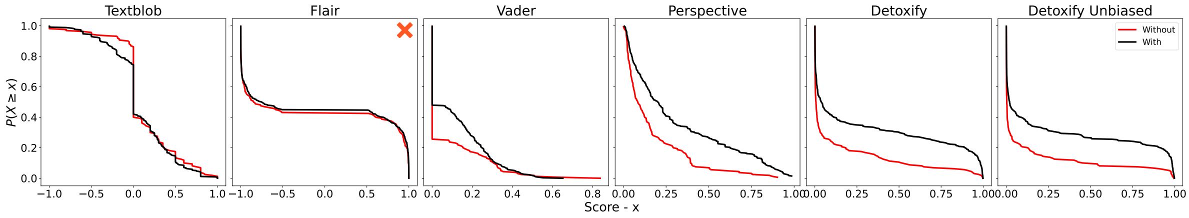


Figure 6.8: Score distributions for sentences with and without African American Terms in the Twitter dataset.

6.3 Twitter

Consistently to the previous datasets, the models in Figure 6.7 manifest biases against the African-American English sentences. In this case, we are able to see a more clear tendency in every model, with the exception of Textblob. However, as explicitly marked in the plots, the Textblob distributions do not reach statistical significance in the Kolmogorov-Smirnov test, suggesting us to reject the Null Hypothesis in favor of the Alternative Hypothesis, i.e., both samples come from the same distribution. Differently from the experiments we have seen before, the curse word usage appears to amplify the disparity between the two linguistic groups.

In Figure 6.8 we show scores for sentences with and without African-American English terms regardless of the linguistic group from which the sentence was drawn. The sentiment analysis models appear to have a less biased output if compared to toxicity models when assessed on sentences containing African-American English terms. However, every model still presents a consistent bias against the use of African-American English terms, with the exception of Flair.

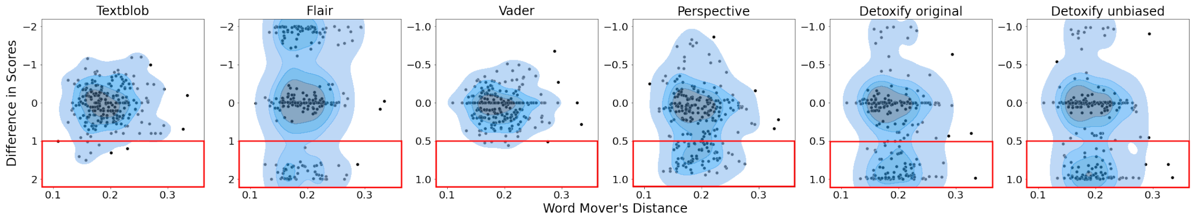


Figure 6.9: Distribution of the differences in the scores for the sentences with the closest meaning according to Word Mover's Distance in the Twitter dataset.

This experiment evidences the prominence of LIWC features such as Swear, Sexual, Netspeak, and Informal. In Table 6.3 we notice an agreement among the estimated importance given by toxicity models. An interesting thing to notice in such experiment, is the importance of the Informal category. At least for one sentiment analysis model, we see the mentioned feature assuming a positive coefficient, i.e., helps sentences to be considered more positive, while for toxicity models, this feature plays a contrary role. Once more, we see the African-American English terms assuming a direct relation with negativity/toxicity scores, something in line with the previous experiments in every dataset.

Features	**Textblob**	**Flair**	**Vader**	**Perspective**	**Detoxify**	**Detoxify Unbiased**
LIWC_SWEAR	-	-	-	**0.8567**	**0.9492**	**1.2792**
LIWC_SEXUAL	**-0.3354**	**-1.3448**	-	**0.4657**	**0.5942**	**0.5609**
LIWC_NETSPEAK	-	**-2.6379**	-	**-0.4496**	**-0.8121**	**-0.9239**
LIWC_INFORMAL	-	**2.5228**	-	**0.4386**	**0.7988**	**0.946**
POS_X	**0.5655**	-	**-0.1636**	**-0.3857**	-0.4255	-0.3066
AAE_TERMS_COUNT	-	-	**0.0934**	0.2238	0.1779	-
LIWC_NEGATE	-	**-0.8331**	-	0.2075	-	0.1855
POS_DET	-	-	-	0.1623	0.3369	0.3268
LIWC_ASSENT	-	-	-	-0.1599	-0.2614	-0.1985
LIWC_MALE	**-0.3237**	-	-	-0.1505	-0.1968	-0.191
RACE	-	-	-	0.0508	0.1262	0.0762
LIWC_FILLER	-	**-2.1683**	-	-	**-0.6053**	**-0.6377**

Table 6.3: Logistic Regression coefficients for the Twitter dataset with p-value < 0.05. The five most relevant coefficients of each model are presented in bold, whereas not statistically significant coefficients were omitted. Note: due to score ranges, the negative coefficients in the red cells present a positive influence in toxicity/negative sentiment.

In Figure 6.9, we can see a tendency of points accumulating at the bottom of the plots, with more intensity with Textblob, and the toxicity models. As we saw in the previous datasets, the models appear to have a deviant behavior for both linguistic groups, however, more accurately for the African-American English group. Even though the Standard English group also gets impacted by the deviant behavior, such disparities might not be seen as a problematic issue once the linguistic variation is not historically associated with a discriminated demographic group.

Chapter 7

Conclusions

In this dissertation, we proposed to investigate the capability of sentiment analysis/toxicity models to skillfully disambiguate harmful situations and normal events regarding African-American English speakers. We analyzed the performance of six different off-the-shelf models in the light of three different datasets. Our datasets encompassed online texts from Twitter, single-speaker closed captions from YouTube, and spoken English interviews that depicting different daily life situations. Overall, our analysis was performed in an attempt to isolate confounding variables from our main focus, the use of African-American English.

7.1 Discussions and Future Work

Considering the late, or nonexistent, introduction of African American English (AAE) in machine learning datasets, there likely exists a non-neglectable under-representation rate in the data used to create these models. We argue that the biggest problems derive directly from the absence of context in the sentences, since in many cases a highly toxic utterance for a specific group may not be seen as toxic at all to others. Our results evidence a more prominent persistent bias within the more powerful approaches, such as the recent transformer-based solutions, when compared to the more rigid solutions such as lexical-based models. Something in line with previous research [52].

As is the case with most observational studies, our results are likely to be impacted by the data sampling strategy. In order to mitigate this fact, overall, we attempted to present a broad-scale analysis focusing on datasets of different natures. Moreover, one of our datasets, YouTube, was gathered with the focus of controlling for confounding variables (i.e., content and dialogue). Finally, we present our analysis on several different models and find similar findings in all of them, though with different effect sizes.

In our results, we see more significant biases in the Twitter dataset, followed by the Youtube data, and finally by the CORAAL and Buckeye. We believe these differences derive primarily from the target audience and the type of communication intended by the author. For example, when interacting via Twitter, a user is usually speaking to a closer audience with a possible real-life acquaintance with the followers. This type of interaction may lead to a substantially informal communication which tends to manifest more dialectal traces of the languages. Similarly, YouTubers may develop their contents based on planned scripts and with a specific audience in mind, not to mention that the platform is often an income source and their content usually is constrained by the community baselines. Lastly, we have the CORAAL/Buckeye dataset that tracks the communication between an interviewer and a speaker.

When dealing Twitter and YouTube datasets, it is worth noticing that we are analyzing only posts and captions in compliance with their respective platforms' community guidelines. Hence, the bias emerging from our analysis is at its best already softened by previous filtering performed by the community and extreme content moderation mechanisms.

With regard to the Logistic Regressions, the African-American English terms count appeared as an important aspect, with a positive impact, on the toxicity/negative sentiment score generated by the models. This result was complemented by the comparison between sentences with and without AAE terms. Such results suggest that this vernacular is stained with biases and the absence of context can be prejudicial to the labeling processing, causing usual and harmless sentences to be deemed toxic/negative.

We have also seen that biases do not necessarily emerge as a sole consequence of curse word usage. Results point to a direct relation between curse words and higher toxicity/negative sentiment scores. However, there seems to be a persistent bias even though the sentences do not comprise such words. It is worth noticing that some terms, despite being a curse word, are not used with a toxic or negative intent (e.g., "this is so f*cking good."). However, this single term bias is capable of playing an important role in the whole sentence score. In other words, there are biases in the sense that not every occurrence of swearing is an authentic toxic/negative utterance.s

Finally, the results present a more restrained/cautious behavior in the Lexical-based models. This pattern is probably a consequence of the construction of the respective lexicons given this endeavor is a conscious and grounded effort usually involving researchers and linguists. Hence, such solutions appear to be less sensitive to the usage of AAE terms.

Appendix A

Black Talk Terms and Expressions

- ace
- ace boon coon
- ace kool
- afraamerican
- afriamerican
- african
- african american
- african centered
- african holocaust
- african people's time
- afrikan
- afro
- afroamerican
- after hour joint
- ain a thang
- ain studyin
- airish
- ak
- all in
- all in the kool aid and don't even know the flava
- all is well
- all that
- all that and then some
- all the way live
- all the way through
- alley ball
- altar call
- amen corner
- amp
- an you know that
- angel dust
- ankh
- ann
- the anounted
- applause
- apple
- a rab
- are you right
- around the way
- as god is my secret judge
- ashy
- ass
- ass from a hole in the ground
- ass on one's shoulder
- ass on his shoulder
- ass on her shoulder
- ass out
- atl
- attitude
- audi 5000
- aunt hagar's chillun
- aunt jane
- aunt thomasine
- aw ight
- b ball

- b boy
- b more
- baby
- baby dady
- baby factory
- baby girl
- baby momma
- baby sis
- back
- back slidin
- backing the numbers
- bad
- bad hair
- bad mouth
- bad nigga
- bag
- bail
- bailin
- ball out
- balla
- ballin
- ballistics
- ballroom
- bamma
- banana
- banger
- bangin
- banjy boy
- bank
- bankroll
- barefoot as a river duck
- bars
- base
- basehead
- be about
- be bout
- be somebody
- beam on
- beam up
- beamer
- bear
- bear witness
- beast
- beastly
- beat down
- beaucoup
- beautician
- beauty shop
- be bop
- bee yotch
- beef
- bees
- befoe god get the news
- behind
- benjamins
- bent
- benz
- benzo
- bet
- betta ask somebody
- betta recognize
- bid
- biddy
- bidness
- big
- the big apple
- big d
- big faces
- big foe
- big four
- big fun
- big lips
- big momma
- big paper
- big time
- big timin it
- big ups

- big willlie
- bill
- big bam thank you mam
- bird
- bitch
- bite
- bittin
- bitty
- bk
- black
- black and tan
- black bottom
- black than
- blacker than thou
- the blacker the berry the sweeter the juice
- blaze
- blazin
- blessed
- blob
- block
- block boy
- blondie
- nlood
- blow
- blow out
- blow the glass
- blow up
- blow up the spot
- blowed
- blue
- blue eyed devil
- blue eyed soul
- blue light special
- blues
- blunt
- bmt
- bmw
- bnic
- bo dick
- bo jack
- boards
- bodacious
- body bag
- body shop
- bogard
- bogue
- da bomb
- bone
- bone out
- boned out
- boo
- boo boo
- boo coo
- boody
- boody call
- boody green
- boody queen
- boodie
- boodie down
- doogie woordie
- booguh bear
- boojee
- book
- boom
- boom box
- boomin
- boones
- boost
- boot
- boot up
- bop
- boppin
- born again
- boost
- boss like hot sauce
- boston

- bottom
- bounce
- bout
- bout it
- box
- box on that fox
- boy
- boyfriend
- bozard
- bra strap
- braids
- brang ass to git ass
- brang it on
- bring it on
- bread
- break
- break down
- break him off
- break her off
- break him off something
- break her off something
- break them off something
- break it down
- break it off
- break on somebody
- break out
- break somebody's face
- break wide
- breakdown
- brew
- brick
- brick house
- bright
- bring the noise
- brang the noise
- bro
- broad
- broccoli
- broke
- broke ass j
- broke down
- brotha
- brown skin
- bs
- buck
- buk whylin
- buck wild
- bucket
- bucket of blood
- bud
- bruddha
- bruddha grass
- blussa soldiers
- buffalo stance
- bug
- but out
- bugging
- bulldagger
- bullet
- bum rush
- bump
- bump one's fums
- bump up
- bumper kit
- bumpin
- bun
- buppier
- burn
- burner
- bus
- bus a cap
- bus a rhyme
- bus on somebody
- bus one's nuts
- bus some cards
- bus somebody

- bush
- business
- busta
- busted
- bustin out
- bustin rhymes
- bustin suds
- the butt
- butta
- butta from the duck
- buy a wold ticker
- buzz
- caesar
- cakes
- cali
- call yourself
- call herself
- call somebody out
- call somebody outa they name
- cameo cut
- can I run wit ya
- cabdy cabe
- cane
- can't kill nothin and won't nothin die
- cap'n save a ho
- cap on
- carbon copy
- case
- cat
- cat faces
- cat walk
- catch you later
- cattin
- cave
- ccm
- changes
- charlie
- check
- check a trap
- check it in
- check him out
- check it out
- check her out
- check this out
- check up
- check yosef
- check you
- check you out
- cheddar
- cheese
- chi town
- chicker eater
- chicken head
- chicken shit
- chief
- chill
- chill out
- chill pad
- chilling
- chine white
- chitlin ciruit
- chitlins
- chocolate city
- choke
- choose
- chose
- chronic
- chuck
- chump
- chump change
- church family
- church folk
- claim
- clean
- clip
- clipped

- clock
- clow
- clown
- cluckhead
- co sign
- coal
- coal bloaded
- coal chillin
- coat
- cock
- cock block
- cock diesel
- cock strong
- cock suck
- cocktail
- cold
- cold blooded
- colom
- color scale
- color struck
- colored
- colored people
- colors
- come
- come again
- come backed up
- come correct
- come out
- come out of a bag
- come wit it
- comin up
- commerciel
- community
- company
- coney oney
- conk
- constant
- conversate
- conversation
- cookie
- cookin
- cookin with gas
- coochie
- cool
- cool it
- cool out
- coolin it
- coolness
- cop
- cop a plea
- copasetic
- corn rows
- corny
- cotton
- couldn't hit him in the behind with a red apple
- couldn't hit her in the behind with a red apple
- the count
- counterfeit
- cover
- cp time
- cpt
- crabs
- crack on
- cracked out
- cracker
- crackhead
- cracking but fackin
- crapped out
- crazy
- creep
- crew
- crib
- crimey
- crips
- cronz
- cross out
- cross ovah

- cross the burning sands
- crumb snatchers
- crumbs
- crystal
- cuffed
- curb
- cut
- cut somebody some slack
- cuttin up
- cuz
- d up
- d whupped
- daisy dukes
- damn skippy
- dank
- dap
- dark skin
- a day late and a dollar short
- day one
- dazzey duks
- dbi syndrome
- dead
- dead presidents
- dead rag
- the deal
- dealin
- death eating a soda cracker
- decoy
- deep
- ef
- delive
- den
- deuce
- deuce and a quarter
- deuce five
- deuce deuce
- devil
- diaspora
- dichty
- dick
- dick whipped
- dick whupped
- die
- diesel
- dig
- dig on
- digits
- dime
- dime piece
- dip
- dipped
- dippin
- dippin and dabbin
- dis
- the district
- diva
- dividends
- dj
- dl
- do
- do a bid
- do a big
- do a face
- do a ghost
- do it fluid
- do it to def
- do it tho the max
- he do not play
- she do not play
- do rag
- do yo thang
- do his thang
- do her thang
- dr thomas
- dr watts

- dodgers
- dog
- dog somebody out
- dome
- dome piece
- don't deal in coal
- don't go there
- don't make me none
- doo wah diddy
- doo wop
- doobie
- doodley squat
- doofus
- dope
- dope fiend move
- dot dat eye
- double deuce
- double dutch
- double r
- double ups
- down
- down by law
- down for
- down for mine
- down home
- down low
- down pat
- down south
- down wit
- down wit the nation
- downtown j
- the dozens
- dp
- drag
- drama
- draped
- draws
- dreadklocks
- dreads
- dream book
- drive by
- driving while black
- drop
- drop a dime
- drop a line
- drop a lug
- drop it
- drop science
- drop top
- droppin babies
- duckettes
- dude
- dues
- duke
- dukie braids
- duckie chain
- duks
- dummy up
- dunk
- dust
- dusted
- dutchmaster
- dwb
- dynamite
- eagle flies
- eagle flyin day
- earth
- earthly things
- easy
- ebonics
- edges
- educated fool
- eight ball
- eight rock
- eight sex
- eight track
- e light
- el pee

- elders
- the electric slide
- ends
- the enemy
- esseys
- european american
- european negro
- everythan is everythang
- evil
- evil eye
- extensions
- eye busted
- f in
- fade
- faded
- fag
- fair
- fair skin
- fake out
- fall
- fall out
- fam
- familiar
- fass
- fat
- fat man against the hole in a doughnut
- fat mouth
- fay
- federated
- federellis
- feed somebody the pill
- feel ya
- fell off
- fess
- field nigga
- fiend
- fiendin
- fifty one
- figure
- fine
- finesse
- fire it up
- first mind
- fish
- five
- five and dime
- five hundred
- five o
- five on the black hand side
- five on the sly
- five percent nation
- flaky
- flat top
- flava
- flex
- flip the script
- flossin
- flow
- fly
- foe by
- foe day
- foe one one
- folks
- for days
- for the duration
- fore day
- forty acres and a mule
- forty dog
- forty ounce
- foul
- four by
- foxy
- franklin faces
- freak
- freebase

- freestyle
- fresh
- fried, dyed, and laid to the side
- fro
- frog
- froggy
- from amazing grace to floating opportunity
- from appetite to ashole
- from jumpstreet
- from the git go
- from the jump
- from jump
- from the rip
- front
- front and center
- front on somebody
- front street
- fruit
- fry
- fuck
- fucked up
- fuhgit it
- fuhgit you
- fuhgit that
- fuhgit him
- fuhgit her
- full face
- full of shit
- funds
- funk
- funky
- funky fresh
- g ride
- g thang
- gaffle
- game
- gangbanger
- gangbangin
- gangsta
- gangsta class
- gangsta lean
- gangsta limp
- gangsta roll
- gangsta walk
- gangsta walls
- ganja
- ganja weed
- gank
- ganker
- gap mouth
- gas up
- gat
- gatas
- gauge
- g'd up
- gear
- geared up
- gee mo nitty
- geek
- get a nut get busy
- ghetto
- ghetto bird
- ghetto fabulous
- ghost
- giddyup
- giddayup
- gift
- gig
- gig on
- girl
- girlfriend
- git a nut
- git busy
- git clipped
- git down
- git ghost

- git go
- git good to somebody
- git happy
- git it on
- git it togetha
- git mine
- git yours
- git his
- git hers
- git off my case
- git on the good foot
- git out my face
- git outa here
- git ovah
- git paid
- git real
- git skins
- git some air
- git some boody
- git some leg
- git the ass
- git the spirit
- git up
- git wasted
- git wit
- git yo bes holt
- git my bes holt
- git his bes holt
- give a care
- give it up
- give some head
- give somebody five
- give somebody skin
- give some skin
- give somebody some play
- give somebody some slack
- give somebody some sugar
- give something some play
- give up the ass
- glass dick
- glass house
- glock
- glory
- go back
- go down
- go for
- go for bad
- go for self
- go for yours
- go for his
- go for hers
- go for what you know
- go off
- go out like a sucker
- go ovah
- go to blows
- goal tendin
- god don't like ugly
- goddess
- goin through changes
- gold digger
- gold front
- gone
- gone home
- good hair
- good lookin out
- good to go
- got game
- got his nose
- got her nose
- got it goin on
- got it honest
- got your back
- grandstand
- grapevine

- gray
- grease
- great white hope
- the greatest
- grill
- grip
- grits
- grown
- grown folk
- grown folk bidness
- grub
- gsp
- gumby
- gut bucket
- ha step
- haim
- haincty
- haints
- hair dressed
- half ass
- half step
- half track
- hammer
- hand
- handkerchief head
- handle the ball
- handle your bidness
- handle my bidness
- handle his bidness
- hands down
- hang out
- happy
- hard
- hard leg
- hard rock
- hard headed
- harlem world
- harvest
- hat up
- hata
- hate on somebody
- hate on something
- have church
- hawk
- hawking
- he say he say
- head
- head hunter
- head nigga in charge
- head rag
- head up
- heads
- heart
- heat
- heavy
- heifer
- hella
- hellified
- hello
- hen dog
- herb
- high
- high five
- high roller
- high top fade
- high fella
- hip
- hip hop
- hit
- hit it
- hit me up
- hit on
- hit the number
- hit the lottery
- hit the skins
- hnic
- ho
- ho cake

- how
- hog
- hog maws
- holding down
- hole
- holla
- holler
- holy ghos
- home
- home on high
- home slice
- homefolks
- homegirl
- homeboy
- homegoing
- homes
- homey
- homo
- honey
- honky
- hoo rah
- hoochie
- hood
- hood rat
- hoodoo
- hoodoo man
- hook
- hook something up
- hook up
- hooked
- hoop
- hoopty
- hops
- hot
- hot blooded
- hot comb
- hot curlers
- hot iron
- hot lady
- hot natured
- hot sauce
- hound
- house
- house nigga
- hump
- hump in his back
- humpin
- hung
- hung low
- hunky
- hush yo mouf
- hustle
- hype
- ibwc
- ice
- ice down
- ice people
- if you feel froggy leap
- ig
- ill
- illin
- i’m out
- in da zone
- in effect
- in full effect
- in like flin
- in the day
- in the house
- in the mix
- in the skins
- in the street
- in there
- in yo face
- indo
- ink town
- iron mike
- ish
- issue

- it ain hapnin
- it's on
- jack
- jack d
- jack move
- jack shit
- jack up
- jackleg
- jake
- jam
- jam session
- jammin
- jammy
- jaw jackin
- jaws tight
- jazz
- jb
- jeep music
- jerk somebody around
- jet
- jheri curl
- jigga
- jiggy
- jim
- jimmy
- jim browski
- jim hat
- jimmy hat
- jim jones
- jimmy joint
- jiglin
- jitterbug
- jove
- jock
- jock strap
- jody
- joe chilly
- johnson
- joint
- jones
- jook
- jordans
- jubilee
- juice
- juiced
- jump
- jump bad
- jump salty
- jump sharp
- jumpstreet
- junetennth
- jungle fever
- kep on keepin on
- keepin it real
- kente
- key
- kibbles and bits
- kick
- kick back
- kick butt
- kick down
- kick it
- kick it around
- kick it live
- kick the ballistics
- kick to the curb
- kickin
- kickin it
- kicks
- kid
- killin fields
- kinks
- kinky
- kitchen
- knock
- knock boots
- knocked off
- knot

- know god
- know what uhm sayin
- knuckle up
- kufi
- kwanzaa
- lady
- laid
- lala land
- lame
- lamp
- lampin
- larceny
- large and in charge
- later
- lawd
- lawd have mercy
- lay dead
- lay it down for me
- lay it on me
- lay out
- lay pipe
- lay up
- layin in the cut
- lean
- leave somebody hangin
- led by the head of one's dick
- leg
- legit
- let the door hit you where the good split you
- let's have church
- letter from home
- lifted
- lifts
- light break
- light bread
- light into
- light skin
- light up
- lighten up
- lightweight
- like that
- like to
- like white on rice
- lil bit
- lil man
- lil somethin somethin
- line don't lie
- lip
- liquid juice
- listen up
- live
- liver lips
- livin high off the hog
- livin large
- lizard
- look city
- lock down
- locker number
- locks
- loke
- look for you yesterday here you come today
- loose change
- loot
- a lot of nature
- loud talk
- love
- love bone
- love me some
- low
- low five
- low life
- low low
- low rate
- low ridin
- low sick
- lp

- lug
- lyin
- macaroni
- mack
- mack daddy
- mackin
- mad
- the madison
- main man
- make bank
- make like
- make somebody's love come down
- mamma jamma
- man
- the man
- mandingo
- mannish
- many windows
- marinate
- mark
- mary frances
- mary jane
- maryland farmer
- max
- the max
- may like
- mc
- me and you
- mean
- mecca
- mellow
- member
- mess
- mess around
- mess wit
- mess wit someone's mind
- mf
- mfic
- michael white jackson
- mickey d
- mickey mouse
- mickey mouse is in the house and donald duck don't give a fuck
- mickey t
- midnight hour
- mind's eye
- miss ann
- miss thang
- mission
- mista charlie
- mista franklin
- mista wind
- mo mo
- moanuhs' bench
- mobbin
- mojo
- molded
- momma
- mommy
- mondo
- money
- monsta
- more coal on the fire
- moriney
- mother
- mother hubbard
- mother wit
- motherland
- motherlode
- mother's day
- mother's day pimp
- mothership
- motor
- motor city
- motown
- mouf

- mourners' bench
- mouthpiece
- mug
- muh fuh
- murder mouth
- murderin us
- murphy
- muthafucka
- my bad
- n's
- the n word
- nana
- nap up
- nappy
- naps
- nathan
- the nation
- natural
- natural high
- nature
- nearer my god to thee
- neck
- negro
- neo slavery
- new jack
- new jill
- nice
- nickel
- nickel n dime
- nickel slick
- nigga
- nigga mess
- nigga please
- nigga rich
- nigga toe
- niggamation
- niggas and flies
- nigger
- nigger apple
- nine
- ninety leben
- nip
- nitty gritty
- 'no'
- no count
- no longer than john stayed in the army
- no love
- no nature
- nod
- noi
- noise
- none
- none yuh
- nookie
- nose job
- nose open
- not tryin ta
- number game
- number man
- number one
- number two
- numbers
- nurse
- nut
- nut out
- nut roll
- nut up
- nuts
- oaktown
- od
- oe
- ofay
- 'off'
- off the hook
- off the wall
- og
- oil

- okay
- oke doke
- ol bird
- old head
- old school
- olde english
- ’on’
- on a mission
- on e
- on full
- on it
- on it like a honet
- on override
- on point
- on somebody’s case
- on somebody’s shit list
- on t
- on the block
- on the case
- on the fly
- on the good foot
- on the outs
- on the pipe
- on the rag
- on the strength
- on time
- one eight seven
- one on one
- one mo once
- one time
- opb
- opp
- oprah
- oreo
- out box
- out cold
- out of order
- outa here
- outa sight
- outside kid
- outtie 5000
- ovah
- oerride
- overseer
- overstand
- packer’s club
- packin
- packin chitlins
- pad
- paddy
- paid
- paper
- paper chase
- paper route
- papers
- par tay
- paranoid
- parlay
- partner
- party
- pass
- pay dues
- payback
- pcp
- pe
- peace
- peace out
- peanut butter
- peck
- peckawood
- peel a cap
- peep
- peep things out
- peeps
- pen
- people of color
- perm
- perp

- perpetratin
- perpetrator
- perpin
- phat
- phd
- philly
- philly blunt
- picked yo pocket
- pickin in high cotton
- pick up game
- pick up lady
- pick up man
- picture
- pie
- piece
- pig
- pig latin
- pill
- pimp
- pimp slap
- pimp strut
- pimp walk
- pimped out
- pink toes
- pipe
- pitch a bitch
- play
- play brother
- play sister
- play cousin
- play aunt
- play like
- play out
- play past
- play pussy and git fucked
- play asome bid
- play somebody close
- play somebody for his reaction
- play somebody for her reaction
- play somebody like a piano
- play that
- play the dozens
- play the numbers
- playa
- playa hata
- playa hate
- player
- player hate
- player hater
- playin for blood
- pluck
- plumbing
- po lice
- po po
- point game
- point up
- poison
- poontang
- poot
- poot butt
- pootenanny
- pop
- pop a cap
- pop a car
- pop somebody
- poppy
- posse
- poundin
- pp
- praise him
- praise house
- prayer march
- press
- pressed
- primo

- process
- profile
- program
- promised land
- propers
- props
- psych out
- puffer
- pull a train
- pull shit
- pull someone's coat
- pull someone's hole card
- pump it up
- pumpin
- punany
- punchy
- punk
- punk out
- push come to shove
- push up on
- pussy
- pussy whupped
- put a baby on a man
- put it on him
- put it on her
- put on wax
- put out with somebody
- put shit on somebody
- put somebody in check
- put somebody on front street
- put somebody's bidness in the street
- put the ig on
- put they mouth on you
- puttin on a clinic
- pwt
- quick fast an in a hurry
- quiet as it kept
- quo vadis
- race man
- race woman
- rada
- rag
- ragmuffin
- ragamuffin tip
- raggedy
- rags
- raise
- raise a hymn
- raise cain
- raise sand
- raise up
- rank
- rap
- rap
- rap attack
- raped
- raspberry
- rasta
- rastafaria
- rastafarian
- rat
- rat pack
- raw
- raw dog
- read
- ready
- rebellion
- recognize
- recruiting
- red black and green
- red eye
- red neck
- reefer
- rejoice
- relaxed

- relaxer
- rent a nigga
- rep
- represent
- re up
- revival
- ride
- ride down on
- ride for
- ride on
- ride shotgun
- right hand of fellowship
- right on t
- rightous
- rinky dink
- rip
- rip it
- rip off
- ripped
- rise
- roach
- road dog
- robo cop
- rock
- rock n roll
- rock star
- rock the house
- rokee
- role
- role
- roll em up
- roll up on
- roller
- rollie
- roulie
- rug rats
- run
- run a boston
- run a drag on
- run a train
- run and tell that
- run it down
- run one's mouth
- run out
- run the street
- run wild
- runnin
- runnin off at the mouth
- rush
- sackchaser
- sadiddy
- salty
- sam
- sanctified
- sang the song
- sapphire
- saturday night special
- sausage
- saved
- savin
- say what
- scag
- scandalous
- scank
- scared of you
- scholar
- scholled
- schollgirl
- schoolboy
- science
- scope
- scope something or somebody out
- scoreboard
- scotty
- scratch

- scream on
- second head
- seeds
- sell a wolf ticket
- sell a woof ticket
- sell out
- sell out negro
- selling bellings
- semi black
- send
- senegalese twist
- sent up
- serious
- serious as a heart attack
- serious bidness
- serve
- set
- set book
- set it off
- set it out
- settin hand
- shade tree
- shades
- shake and bake
- sheik
- sherm
- shine
- shit
- the shit
- shit from shinole
- shit hit the fan
- shiz out
- sho you right
- shook
- shoot dice
- shoot some hoop
- shoot the die
- shoot the gift
- shoot the shit
- shootin the rock
- short
- shorty
- shot caller
- shout
- shout out
- show
- show and prove
- show some sign
- showboat
- shuckin and jivin
- shut the noise
- shut up
- sick
- sig
- siggin
- signification
- signify
- signifyin
- silk
- silly
- simp
- single action
- single action lady
- single action man
- sissy
- sista
- sista rea
- skeeze
- skeezer
- skillz
- skin
- skinnin and grinnin
- skins
- skunk
- sky
- slack
- slain in the spirit

- slam
- slam dunk
- slammin
- slammin partner
- slang
- slanguage
- slave
- sleef
- sleep
- slick
- slick shit
- slide
- slippin and slidin
- slob
- slope
- slow jam
- slow your roll
- slow my roll
- slow his roll
- slow her roll
- smack
- smash
- smellin up behind somebody
- smoke
- smoker
- smokin
- smooth
- snake
- snapper
- snappin
- snaps
- snatch
- snort
- snow
- snow bunny
- solid
- some
- sooki sooki
- sooner
- sorry
- sos
- soul
- soul brotha
- soul sista
- soul clap
- soul food
- soul shake
- soul sound
- spade
- spirit
- spleefer
- splib
- sponsor
- spook
- sport
- spot
- springs
- sprung
- square
- square bidness
- squash
- squash
- squash it
- stace adams
- stallion
- stank
- star
- static
- stay in the street
- stay up
- staying on the place
- steady
- steal
- step
- step off
- step show
- step to

- step up
- stepping
- stick
- stick it
- stiff
- stocking cap
- stole yo lunch
- stomp
- stone
- stone to the bone
- stoopid
- storefront church
- story
- straight
- straight up
- straghten
- straighten up and fly right
- strap
- strapped
- strawberry
- stray piece
- street wear
- strides
- stridin
- stroke
- strong
- stronger than red devil eye
- the struggle
- strung out
- strut
- strut yo stuff
- stuff
- stupid
- stylin and profilin
- sucka
- sugar
- the sugar
- sup
- superfly
- supermarket conversation
- sure you're right
- sweat
- sweet
- sweet talk
- sweetie
- swep
- swept
- swoop
- system
- table pimps
- tail on the whale
- take a chill pill
- take a text
- take a care of bidness
- take it to the hole
- take it to the hoop
- take low
- take out
- take shit
- taking no shorts
- talk shit
- talk show shit
- talk that talk
- talkin head
- talkin in tongue
- talkin out the side of your neck
- talkin out the side of your mouth
- talkin smack
- talkin to
- talkin trash
- tall paper
- tap it
- tap that ass

- tarrying service
- taste
- tcb
- tear the roof of the sucka
- telephone number
- tell it
- tell the truth
- tender
- tenderoni
- tent meeting
- terrible
- testify
- tg
- that how you livin
- that you
- that's all she wrote
- that's mighty white of you
- there it is
- thick
- thick lips
- third struggle
- thirty eight
- thought like lit
- threads
- three six nine
- three sixty five
- through
- throw
- throw down
- throw a bricj
- throw bones
- throw the d
- throw the p
- throw the gift
- throw up a brick
- thump
- ti s
- tight
- tight as dick's hatband
- tight as jimmy's hatband
- timbos
- time
- tip
- tired
- tlc
- to put somebody on ice
- tow out the frame
- toe up
- togetha
- token
- tom
- too through
- top of my game
- top of mhis game
- top of her game
- torn up
- totaled
- touch it up
- trey eight
- trick
- trickeration
- trickin
- tricknology
- triflin
- trim
- trip
- trippin
- trip somebody out
- tripple nickels
- truckin
- truth be told
- tryin to make a dolla outa fifteen cent
- tude

- turf
- turkish
- turn a book
- turn somebody out
- turn something out
- twenty cents
- twenty foe seven
- twinkie
- twisted
- two minute brotha
- uaw
- uhm out
- uncle thomas
- uncle tom
- undergound hit
- up on it
- up shit creek
- up south
- ups
- upside yo head
- upside the head
- upside his head
- uptight
- usg
- vamp
- vapors
- vee in
- vega
- verdict
- vibe
- vicious
- vine
- visitation of the spirit
- voodoo
- ws
- wack
- wannabe
- washed in the blood
- waste
- watch da bows
- watch meeting night
- watermelon head
- wave nouveau
- wax
- wax some ass
- we be clubbin
- weak shit
- weak side
- weak face
- wear out one's welcome
- wear you out
- wear them out
- wear it out
- weave
- weed
- weight
- well all right
- wes side
- wham bam thank you mam
- whass crackin
- whass hapnin
- whassup
- whassup with that
- what go round come round
- what it b like
- what it c like
- what set you from
- what time it is
- what up
- what up doe
- what's happening
- what's up
- what you on
- whip
- whis

- whissin
- white on rice
- white white
- whitemail
- whitenization
- whitey
- who yo daddy
- whole lotta yella wasted
- the whole nine
- whore
- whupped
- wifey
- wifin
- wigga
- wigger
- wild
- wilderness
- willie
- windy city
- wit
- wit the program
- with
- with the program
- witness
- wolf
- woman
- womanish
- womanist
- womlish
- womnish
- wood
- woof
- woof ticket
- woofer
- word up
- word is bond
- word to the mother
- you don't hear me though

Bibliography

[1] Abubakar Abid, Maheen Farooqi, and James Zou. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021.

[2] CJ Adams. New york times: Using ai to host better conversations. https://blog.google/technology/ai/new-york-times-using-ai-host-better-conversations/, 2018.

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[4] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759, 2020.

[5] Arnetha F Ball. Cultural preference and the expository writing of african-american adolescents. *Written Communication*, 9(4):501–532, 1992.

[6] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128, 2021.

[7] David Bamman, Chris Dyer, and Noah A Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, 2014.

[8] John Baugh. Runnin'down some lines: The language and culture of black teenagers, 1981.

[9] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

[10] Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. Twitter universal dependency parsing for african-american and mainstream american english. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1415–1425, 2018.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[12] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47, 2022.

[13] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[16] Joey Lee Dillard. Non-standard negro dialects-convergence or divergence?. *The Florida FL Reporter*, 65(2), 1968.

[17] Joey Lee Dillard. *Lexicon of Black English.* ERIC, 1977.

[18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[19] David Garcia, Ingmar Weber, and Venkata Rama Kiran Garimella. Gender asymmetries in reality and fiction: The bechdel test of social media. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[20] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):1–41, 2016.

[21] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[22] A Gomes, D Antonialli, and T Dias-Oliva. Drag queens and artificial intelligence. should computers decide what is toxic on the internet. *Internet Lab blog*, 2019.

[23] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[24] Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.

[25] Lisa J Green. *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[26] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[27] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 789–798, 2022.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[29] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[30] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Unintended machine learning biases as social barriers for persons with disabilitiess. *ACM SIGACCESS Accessibility and Computing*, pages 1–1, 2020.

[31] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[32] Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Measuring gender bias in news images. In *Proceedings of the 24th International Conference on World Wide Web*, pages 893–898, 2015.

[33] Jigsaw. Perspective api. `https://perspectiveapi.com/`. Accessed: 2023-01-30.

[34] Jigsaw. How latin america's second largest social platform moderates more than 150k comments a month. `https://medium.com/jigsaw/how-latin-americas-second-largest-social-platform-moderates-more-than-150k-comm`
2019.

[35] Tyler Kendall and Charlie Farrington. The corpus of regional african american language (version 2021.07). eugene, or: The online resources for african american language project, 2021.

[36] Animesh Koratana and Kevin Hu. Toxic speech detection. *URL: https://web. stanford. edu/class/archive/cs/cs224n/cs224n*, 1194, 2018.

[37] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

[38] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318, 2021.

[39] William Labov et al. A preliminary study of the structure of english used by negro and puerto rican speakers in new york city. 1965.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[41] Steven Loria. textblob documentation. *Release 0.15*, 2, 2018.

[42] Patricia Georgiou Marie Pellat. Perspective launches in spanish with el país. https://medium.com/jigsaw/perspective-launches-in-spanish-with-el-pa%C3%ADs-dc2385d734b2, 2018.

[43] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[45] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21:1081–1088, 2008.

[46] Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. Distinguishing the popularity between topics: a system for up-to-date opinion retrieval and mining in the web. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II 14*, pages 197–209. Springer, 2013.

[47] Daniel Borkan Patricia Georgiou, Marie Pellat. Parlons-en! perspective and tune are now available in french. https://medium.com/jigsaw/perspective-tune-are-now-available-in-french-c4cf1ca198f2, 2019.

[48] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[50] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[51] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005.

[52] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[53] Max Roser, Hannah Ritchie, and Esteban Ortiz-Ospina. Internet. *Our World in Data*, 2015. https://ourworldindata.org/internet.

[54] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[55] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

[56] Geneva Smitherman. *Black talk: Words and phrases from the hood to the amen corner*. Houghton Mifflin Harcourt, 2000.

[57] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. Boosting image sentiment analysis with visual attention. *Neurocomputing*, 312:218–228, 2018.

[58] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[59] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.

[60] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2017. *Cyberemotions: Collective emotions in cyberspace*, 2014.

[61] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[63] Pranav Narayanan Venkit and Shomir Wilson. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*, 2021.

[64] Hao Wang, Doğan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120, 2012.

[65] Maciej Widawski. *African American slang: A linguistic description.* Cambridge University Press, 2015.

[66] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35, 2005.

[67] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[68] Min Yang, Qiang Qu, Xiaojun Chen, Chaoxue Guo, Ying Shen, and Kai Lei. Feature-enhanced attention network for target-dependent sentiment classification. *Neurocomputing*, 307:91–97, 2018.