Master's Thesis of Computer Science

How Differently Do People Hate? Understanding The Linguistic Difference Of Regional English Hate Speech

Group :
Web Information Systems

TU Delft

Master's Thesis of Computer Science

How Differently Do People Hate? Understanding The Linguistic Difference Of Regional English Hate Speech

by

Author: **Baitian Zhang**
Thesis Advisor: **Prof. Avishek Anand**
Daily Supervisor: **Prof. Jie Yang**
Daily Co-Supervisor: **Dr. Sarah Carter**

TU Delft

Contents

1

Acknowledgements

Hereby I shall express my gratefulness to those who have supported my research and my life in the Master's study at Technische Universiteit Delft. I am highly thankful for the knowledge and financial and emotional support provided to me in the past two years. I could not have finished my thesis without this help.

I would like to thank Professor Jie Yang for being the supervisor of my thesis. Professor Yang instructed me on how the thesis shall be designed and offered valuable suggestions for each step of the research.

I would also like to thank Dr. Sarah Carter for being my daily supervisor. Dr. Carter communicated with me weekly, together with Professor Yang, to discuss the research progress and provide suggestions for the research design, which kept me on the right track of research.

I am grateful to Professor Stefan Buijsman that he inspired me in the selection of thesis topics. I could only have picked this suitable thesis topic with his help.

I would like to express my gratitude to Professor Avishek Anand, and Professor Burcu Kulahcioglu Özkan for attending my thesis defense.

I thank all of my family, including my father, mother, brothers, sisters, grandfathers, grandmothers, uncles, and aunts for their warm care and support. They are my strongest backbone and my eternal supporters.

I thank all of my friends. We share laughs and knowledge, and we help each other. Because of your friendship, my life has become bright and joyful.

I thank all the professors and instructors who have taught me in my Master's study. I appreciate your teaching and guidance.

I thank all TU Delft staff for working hard to keep the university running orderly.

2

Introduction

This research, which aims at understanding the linguistic characteristics of regional English hate speech and identifying online hate speech, provides insight into the fundamental components of the English language and looks into the linguistic difference in regional English hate speech from lexical, semantic, and syntactic perspectives. It finds marked disparities in vocabularies, pronouns, part-of-speech tags, lexical relations, phrasal usage, and syntax transitional probability. It also utilizes these linguistic features for training machine learning classifiers (Random Forest Classifier, Neuro Symbolic Classifier with Long-Short-Term Memory, and Decision Forest Classifier) for automated hate speech detection, and their performances are compared. Finally, it examines the efficiency of the aggregation of linguistic features and machine learning models and reveals the difficulty in selecting useful linguistic features for the model training. Nevertheless, it provides practical suggestions for refining research design and future work.

2.1. Motivation

We are living in an era where the use of the English language is no longer restricted to one nation. English is either the major language or one of the official languages of many nations and cultures globally, such as Australia, India, and Singapore [13, 26, 37]. The regional differences among these English variants have been intensively studied, from aspects of history and social function to their vocabularies and phonologies. **Nelson et al.** [35] introduce the varieties of English, e.g. Australian and New Zealand English sharing similar historical timeline, and similar pronunciations such as raised short front vowels and the development of an onglide for long high vowels, or the use of weird nicknames in Australian English, for example, the use of “Bazza” or “Baz” for “Barry”. At the same time, this characteristic is not found in New Zealand English. In Singaporean English, the use of colloquial Chinese words and the missing noun inflection ending (plural -s and genitive -s) makes it a unique variant of the English language.

In this era, the size of online content has expanded dramatically, and the content uploaded to the Internet each day is more than a man can ever see in life. On Twitter, as early as 2013, the daily new tweets had reached 500 million [4]. On Reddit, an American online content platform, there are around 1.7 billion publicly available comments from year 2007 to 2015 [2]. At the same time, English web content has made up 50% of all content [39], indicating its linguistically dominant role on the Internet.

The downside that comes with the abundance of online data is the prevalence of online hate speech. We use X (formerly Twitter) as an example: The average rate of hate speech on the platform is 0.6% [24], and around 34% of tweets are in English. The average hate speech rate of English tweets is around 0.35% [19], though not the highest compared to other languages (French 0.75%, Arabic 0.51%), it still means that English has the highest number of hate tweets due to its overall amount.

More hateful speech means a higher chance of encounters, and it brings a considerable negative impact on Internet users. **Altman** [9] claims that hate speech degrades the dignity of certain classes of people and damages their belief in being treated fairly. In the book "*Understanding words that wound*", **Delgado** claims that the victims of hate speech will be impacted psychologically and emotionally, facing distress [18]. Besides real-life hate speech that is spoken face to face, hate speech has been creating aggression and hostility on the Internet. One of the recent examples comes from the COVID pandemic, 2019, when the number of hateful tweets mentioning either "China" or "Chinese" spiked, frequently combined with the words "Virus", "F**k", and for most of the time in 2020, the percentage of hate tweets against china was well above average rate [42, p. 6]. Sometimes the hate speech may even incite domestic terrorism acts, as suggested by Piazza [36, p. 19], the frequent hate speech by politicians can raise the rate by as much as 900%.

2.2. Research Question

The English language makes up the majority of online content. It is spoken globally in many nations, and its variants differ in phonology, vocabulary, and grammar, signaling a higher possibility of its hate speech being different. With all these factors combined, we conclude that it is necessary to examine the hate speech of different English-speaking nations. We would like to know their linguistic differences and how we could utilize them for model training.

Thus, we propose the first research question:

- **How does hate speech vary between English-speaking nations**

To answer it, we break it down into three parts:

- **What is the lexical difference in hate speech among English-speaking nations?**
- **What is the semantic difference in hate speech among English-speaking nations?**
- **What is the syntactic difference in hate speech among English-speaking nations?**

The word "lexical" means anything related to words and vocabulary [5]. Therefore, to talk about "lexical" is to talk about words, or a more professional term, "lexicon" [32], meaning the vocabulary. The lexicon plays an important role in computational linguistics, as it helps understand the relation between words and their usage rules. It helps natural language processing applications generate sentences that meet with the grammatical rule and meaning [38, p. 272-277].

The word "semantic" can refer to many things, but in this research, we focus on "meaning". Studying the semantic difference will then be the study of meanings.

The term syntax is about how the linguistic elements are combined to form constituents [7]. In other words, syntax is concerned with the form of sentences, without taking into account the effects of uttering sentences in a context [15, p. 97]. To study the syntactic difference is to study how the sentence is constructed differently.

The **English-speaking nations** refer to these six nations. English is either their official language or one of their major/official languages.

- Australia
- New Zealand
- Singapore
- United Kingdom
- South Africa
- India

Apart from the linguistic study, we are also aware of the urgency of building up an efficient automated hate speech detection system due to the enormous size of online data. We then propose the second research question:

- **Can the linguistic finding help build a better hate speech detection system?**

An automated hate speech detection system exempts people from filtering comments by reading each piece of text. It improves efficiency, reduces people's exposure to harmful text, and maintains mental health.

2.3. Limitation Of Our Research

This research incorporates linguistic knowledge into a hate speech study, presenting rather social-science-oriented results. Thus, it pays less attention to natural language processing; A deeper investigation of the linguistic features of text such as the inflection of verbs or how different types of clauses serve as the Noun Phrase or Prepositional Phrase in the hate speech (**Burton-Roberts and N.**[15, p. 200]) is not done due to the limit of the dataset and the scale of research; In addition, the results could over-generalize the regional difference without analyzing the context of the data; The models trained with these data could be less effective in changing scenarios, such as when classifying comments from different online forums, or the comments are from a different time compared to the texts used for the model training.

3

Literature Review

3.1. Cross-Cultural Studies

In recent years, the difference between hate speech among different cultures has been studied by some researchers. They denote the reality that a language is not always used only by people of a certain culture or a certain nation but by people of many cultures and many nations.

The difference in hate speech annotation consensus of English-speaking nations has been examined by **Lee et al.** [27]. They measure the consensus rate between people from different English-speaking nations (Australia, Singapore, South Africa, the United Kingdom, and the United States) when they annotate the text as hate/non-hate. The result shows a a significant disparate consensus rate of 56.2% among all countries, in hate speech annotation. The research improves the classifier by adding the country label to the training data, achieving up to 8.2% performance increment. However, this research does not look into the linguistics of hate speech, giving no insight into its lexical or syntactic components.

The urgency to expand the non-English cross-cultural dataset is stated by **Arango Monnar et al.** [11]. They collect tweets from Chilean Spanish to enrich the Spanish hate dataset, adding cultural diversity. Monnar adapted a cross-lingual setup and found that the existing pre-trained models performed better on the Spanish dataset than on the Chilean Spanish dataset, indicating a lack of representation of the Spanish language spoken outside Spain. This proves that the mono-cultural data cannot represent a multi-cultural language well enough.

Mubarak et al. [33] take Arabic dialects into account, mentioning the difficulty of classifying Arabic offensive speech by spellings. The research proposes a new method to collect offensive tweets regardless of their topics by using emojis to identify emotions. This research looks into the grammatical structure of hate speech, and compares the performances of multiple machine learning classifiers and language models on the new Arabic dataset with emojis, showing that the fine-tuned Arabic transformers have the best overall performances by at most 9.02% higher than multilingual XLM-RoBERTa. The result indicates the necessity of expanding the corpus with regionally varied data.

Frenda et al. [23] take different English-speaking regions into account. They collect data that reflects various demographic perspectives (Australian, British, Indian, Irish, American) through crowd-sourcing work, and demonstrate the value of perspective-aware models in irony detection tasks. This research confirms the necessity to expand the dataset with

language varieties.

3.2. Automated Hate Speech Detection With Artificial Intelligence

Artificial intelligence is a regular guest in hate speech studies. The reality of hate speech using more than certain hate words (e.g. no explicit hate words in "*When will these dark people leave our country?*"), makes AI an ideal tool for classification.

Traditional machine learning classifiers are commonly used in hate speech detection. **Frenda et al.** [22] and **Magu et al.** [28] both use Support Vector Machines, for which the first research includes n-gram, hand-crafted lexicons (keywords for different linguistic features like vulgarity, hashtag, abbreviations) and shows promising results of identifying misogynous and sexist speech which outperforms the basic model, signifying the importance of training classifiers with various features, and the second research successfully identifies the pattern of coded words in a racist and nonracist context, and manages to train the classifier that achieves 0.795 precision rate and 0.794 recall rate. It provides insight into uncovering the context under subtle and evasive language.

Besides traditional machine learning approaches, researchers use language models such as BERT and RoBERTa. Some researchers do a comparative analysis of performances of traditional classifiers and the cutting-edge language models, like **Chiril et al.** [17], where Long Short Term Memory, Convolutional Neural Network, ELMo, and BERT are spontaneously evaluated with multiple types of hate speech (Sexism, Misogyny, Xenophobia, Racism), with the BERT model getting the highest scores when trained on any dataset. This reveals the exceeding power of the language model, compared to traditional approaches.

Interestingly, the syntactic feature can also be integrated with artificial intelligence, as suggested by **Mastromattei et al.** [30], where the syntax tree of hate speech is used as a feature for the neural network training called "KERM-HATE", and the model gets the highest F1 score and average accuracy against all other models including BERT, RoBERTa, and XL-Net. This research proves the value of traditional linguistic knowledge in natural language processing.

Emotional analysis can be another useful feature to train hate speech classification models. **Martins et al.** [29] bind emotional analysis to NLP, adding emotional dimension as a feature to the model training, achieving double precision rate from original 41% to 80.64%, highlighting its potential in hate speech classification task.

Álvarez-Carmona et al.[44] examine hate speech in Mexican Spanish, and use the text and images for author profiling and aggressiveness detection. The paper builds new corpora considering tweets from Mexican Twitter users. Yet the best results of author profiling were achieved using a text-based instead of a multi-modal method. For aggressiveness detection, the best performance is also achieved by a simpler method instead of complicated algorithms, implying that the high complexity of detection models is not always equal to high performance.

The robustness of models under different scenarios is a popular metric when measuring model performance. **Florio et al.** [21] test the temporal robustness of a BERT model for the Italian language, finding that as time goes on, the model trained with old data performs eventually worse on the new data, with at most 0.1 drop in the F1 score, highlighting the difficulty of maintaining classification ability when the data is heavily influenced by events.

This result implies the necessity to collect data from a wide period.

Nascimento et al. [34] examine the difficulty in automated hate speech detection task. They summarize various detection methodologies like SVM, Logistic Regression, and Neural Networks, and highlight the main challenges including the subtle forms of hate that impact the performance of detection algorithms.

3.3. Linguistics

In this section, the focus shifts to the exploration of linguistics. Since our research questions are about lexical and syntactic differences, we must understand relevant linguistic concepts to answer them.

3.3.1. Lexical Analysis

3.3.1.1. Pronoun

The pronoun is a significant part of hate speech. It is a reference to people and entities. To express hate, one must point out the target, whether it is "you", "us", "them" or "them". An analysis of pronouns will help us understand the reference pattern (e.g. There is more plural personal pronoun *We, They, Them, Our* in Nation A's hate speech than Nation B's, suggesting more group references). There are multiple categories of pronouns [1, p. 1]:

- Personal Pronoun: (e.g. "*I, you, he, she*") In hate speech, the personal pronoun is the reference to people (e.g. "*They are s**t; I hate them*").
- Demonstrative (e.g. "*This, These, Those*"). The demonstrative pronouns refer to certain people/entities. (e.g. *This man is full dumb; Those people shall go to h*ll*)
- Relative (e.g. "*Who, Which*"). Relative pronoun indicates the relation to the people/entities. (e.g. *Only those who are ret*rds will like this music.*)
- Indefinite (e.g. "*all, every, some*"). The indefinite pronouns refer not to certain people/entities but to general people/entities. (e.g. "*Every **** is dirty and mean"*)
- indefinite relative (whoever/whomever/whatever). Similar to indefinite pronouns, it refers to people/entities that have certain characteristics or perform certain actions (e.g. "*Whoever likes them is stupid; Whatever is dark-skinned is disgusting."*).
- Reflexive (e.g. *Myself, Yourself, Themselves*). It is used when the subject is the object, usually for emphasis. (e.g. *They shall K*ll themselves.*)
- interrogative (e.g. *Which, What*). It indicates questions. (e.g. *What kind of st**id people are they?*)
- Reciprocal (e.g. *"each other, one another"*). It describes a mutual reference. (e.g. *Those two can talk to each other and d*e.*)

3.3.1.2. Part Of Speech

"Analysing a language grammatically involves breaking it down to a variety of elements and structures: phonemes, morphemes, and words, and within the words, syntactic categories of various sorts. Among these categories are the parts of speech tags (also known as lexical

or grammatical categories)" [38, p. 288]. In simpler words, the part of speech is about what type a word is. To explain it, we use the sentence "**I like eating apples**", in which "I" is a pronoun (PRP: Personal pronoun), referring to the subject, a person or thing, "like" is a verb (Verb, non-3rd person singular present), showing the action of the main subject, "eating" is the present particle of "eat"(Verb, gerund or present participle), serving as a gerund to function as a noun, and "apples" is a plural noun (NNS: Noun, plural), referring to the object. There are other grammatical categories such as [38, p. 309]:

- Adjective
- Adverb
- Preposition
- Auxiliary
- Determiner
- Complementiser

These categories are called "**Part-Of-Speech tags**", or "**POS tags**". They help us understand the composition of sentences from the perspective of meaning, and word preferences (e.g. Nation A uses question markers *who, what, which, why, how* more than Nation B). A more detailed list of part of speech tags is presented below [14], including most of the common word types:

Tag	Description
CC	Coordinating conjunction (e.g., "and", "but")
CD	Cardinal number (e.g., "one", "two", "3")
DT	Determiner (e.g., "the", "a", "an")
EX	Existential there (e.g., "there is")
FW	Foreign word (non-English) (e.g., "Déjà vu")
IN	Preposition/Subordinating conjunction (e.g., "in", "on", "because")
JJ	Adjective (e.g., "big", "quick", "blue")
JJR	Adjective, comparative (e.g., "bigger", "quicker", "bluer")
JJS	Adjective, superlative (e.g., "biggest", "quickest", "bluest")
LS	List item marker (e.g., "1.", "2.", "A.", "B.")
MD	Modal (e.g., "can", "will", "could", "would")
NN	Noun, singular or mass (e.g., "dog", "car", "music")
NNS	Noun, plural (e.g., "dogs", "cars", "bottles")
NNP	Proper noun, singular (e.g., "John", "London")
NNPS	Proper noun, plural (e.g., "Americans", "Indians")
PDT	Predeterminer (e.g., "all the", "both the")
POS	Possessive ending (e.g., "'s", "'")
PRP	Personal pronoun (e.g., "I", "you", "he", "she", "it")
PRP$	Possessive pronoun (e.g., "my", "your", "his", "her", "its")
RB	Adverb (e.g., "quickly", "softly", "well")
RBR	Adverb, comparative (e.g., "higher", "better")
RBS	Adverb, superlative (e.g., "highest", "best")
RP	Particle (e.g., "up", "off", "over")
SYM	Symbol (e.g., "$", "%", "&")
TO	to (e.g., "to go", "to buy")
UH	Interjection (e.g., "uh", "wow", "ouch")
VB	Verb, base form (e.g., "take", "run")
VBD	Verb, past tense (e.g., "took", "ran")
VBG	Verb, gerund or present participle (e.g., "taking", "running")
VBN	Verb, past participle (e.g., "taken", "run")
VBP	Verb, non-3rd person singular present (e.g., "take", "run")
VBZ	Verb, 3rd person singular present (e.g., "takes", "runs")
WDT	Wh-determiner (e.g., "which", "that")
WP	Wh-pronoun (e.g., "who", "what")
WP$	Possessive wh-pronoun (e.g., "whose")
WRB	Wh-adverb (e.g., "where", "when")

Table 3.1: List of Part-Of-Speech Tags

3.3.2. Semantic Analysis

3.3.2.1. Lexical Relations

Lexical relation is about how a word relates to other words in languages [38, p. 136]. When constructing sentences, speakers must choose among various lexical items, for example choosing *"kitchen"* over *"Restaurant"* for describing a food place [38, p. 136], or to choose

"animals" over *"creatures"* when describing a zoo. In the past, such relations were only available in dictionaries, but since 1985, a project named WordNet, an online lexical database has been built for representing and organizing lexical semantic information in a psychologically realistic form [38, p. 272]. In WordNet, for example, one can retrieve synonyms of an English word, e.g. searching *beat* returns the following synsets: {beat, flatten}, {beat, throb, pulse}, {beat, flog, punish}, {beat, shape, do metalwork}, {beat, baffle}, {beat, stir, whisk} [38, p. 290]. WordNet includes not only synonyms (words with similar meanings) but other meaning relations like "meronym" (a part of a bigger whole, e.g. "wheel" is part of "car"), "hyponym" (the word that is included in the meaning of another word, e.g. "rabbit" is the hyponym of "animal", "holonym" (the word included in a larger part, e.g. the holonym of "car" is "traffic", as the car is included in the traffic) [38, p. 136-150, 445]

3.3.3. Syntactic Analysis

Several fundamental concepts are crucial to the syntactic analysis, including constituents & phrases, conjunctions, and clauses. They help with understanding the structural characteristics of languages (e.g. Nation A tends to use multiple clauses in hate speech while Nation B uses less). We will further convert them to features for model training.

3.3.3.1. Constituents & Phrases

The constituents are the parts a sentence can be divided into [15, p. 6]. In a sentence like "The teacher read a book in the library", each word is not equally related to the words adjacent to it [41, p. 4]. The word "**read**" has no direct relationship with "**a**". Instead "**a**" is directly related to "**book**" as an article, and "**the**" is directly related to "**library**" as an article as well. These words are organized into units called "**constituents**"(e.g. *"a book, the teacher"*), which are then organized into larger units called "**constituent structure**" [41, p. 5].

The phrase is a **sequence of words** that can function as constituents [15, p. 15]. A constituent composed of a noun and an article is a **Noun Phrase** [NP] [41, p. 5]; A preposition following an **noun phrase** is a **Prepositional Phrase** [PP] [41, p. 5]; A verb following an **noun phrase** is **Verb Phrase**[VP][41, p. 5]; A group of words putting together as an adjective, like "extremely subtle", or "too modest" is **Adjective Phrase** [ADJP] [15, p. 55-56]; A group of adverbs putting together to function as one adverb is **Adverb Phrase** [15, p. 55-56].

3.3.3.2. Conjunction

Conjunctions are the linking words like *and, or, but, then, because* [16]. Phrases like *as well as, as long as* can be seen as multi-word conjunctions [16]. A **Conjunction Phrase** is a **Coordinating Phrase** followed by any other phrase, such as [CC]+[NP]/[VP]/[ADVP].

3.3.3.3. Clause

A clause is a sentence that contains sentences as constituents. An easier way to describe it is"**with sentential recursions**" [15, p. 171]. For example: "**Georgette said she burned the fritters**", in this sentence, the constituent structure is in the following graph, where there is another sentence **"She burned the fritters"** following **"that"** [15, p. 171].

3.4. Research Gaps Of Existing Studies

Existing studies on hate speech/multi-cultural hate speech often overlook the integration of linguistic characteristics and computer science approaches. They either focus primarily on the categorization of hate speech[22], using these categories as features for model training and asserting that this approach enhances model performance. Some concentrate on comparative analysis of model effectiveness[17, 33]. Therefore, our research attempts to amend this gap by introducing various linguistic features into the computer-science-based hate speech study.

4

Methodology

4.1. Hypothesis Formulation

Considering the variations of linguistic features in different English-speaking nations, the complexity of the grammatical structure, and the richness of lexicons, the following hypotheses are proposed.

4.1.0.1. Lexical Differences

- Despite the regional difference, there is a significant overlap in the vocabulary used in the hate speech of all nations (either by meanings or by spelling, e.g. F**ked, F**king can be seen as the same word by meaning). This comes from the observation of online hate speech, that certain words are frequently used in offensive, abusive, and hate languages, like "Sh*t", and "A**h*le". We assumed that even if split by nations, such a pattern would persist. Examining this hypothesis provides an overview of different nations' hate speech, and proves whether it is possible to identify the nationality of the author of hate speech with a purely vocabulary approach.

4.1.0.2. Syntactic Differences

For the syntactic differences of hate speech, we had the following hypotheses.

- The sentence lengths of hate speech in Singapore and South Africa tend to be shorter than those of the other nations. The hypothesis is built upon the fact that their English is more mixed with local languages [35], and it gives an impression that there are more regional words used, which exemplifies the need for complicated sentence structures.

- Despite regional dialectal differences, the use of phrases (e.g., noun phrases, verb phrases, prepositional phrases) in hate speech is similar across the six nations. In detail, this suggests a similar pattern in the phrase transition, with each nation having a similar transitional probability from one phrase to another (e.g. $prob\{[NP] \Rightarrow [VP]\} = 0.3$). The hypothesis originates from the fact that English is the primary language in all of these nations, being used in their educational systems, suggesting that the grammar is potentially similar.

- Australia and New Zealand have the closest pattern of part-of-speech. United Kingdom's pattern is less similar but still close to them. India, South Africa, and Singapore are more distant and far away from them, with India being the closest to the three nations. This hypothesis originates from the fact that Australia and New Zealand share a long piece of history back in the colonization period, and the short geographical distance between the two nations suggests more contact, making the language similar[35].

4.1.0.3. Automated Detection System Training

We expected the classifiers to be slightly more accurate in detecting hate speech with the extra knowledge introduced in the first part of the research. Besides, previous studies have proven it effective to use extra features in the training [30][29].

Examining these hypotheses helps with enhancing the comprehension of linguistic diversity, from both vocabulary (lexical difference) and structural (syntactic difference) perspectives. It also helps with the the development of more sophisticated hate speech detection systems, that these diversities if proven, can be utilized as features to train culture-sensitive detection systems.

4.2. Methodology

We divided the experiment into five parts: data collection, data filtering, understanding lexical differences, understanding syntactic differences, and improving the hate speech detection system.

4.2.1. Data Collection

The data referred to the hateful text written by the English speakers of the six English-speaking nations.

There were many available data sources. However, we must confirm that the data collected was indeed from the six English-speaking nations:

- YouTube. It is not possible to know the nationality of the comment author. However, it is possible to tell under which channel the comment is posted. We assumed that a comment posted under a video on the channel "7NEWS Australia" has a high probability of being written by an Australian English speaker. We also assumed the same for other nations' comments. Lee et al. [27] utilized the same approach, constructing their dataset by copying comments from the channels that are linked to those nations. However, the biggest disadvantage of this approach is obvious: it is inefficient, as one must check each comment to decide whether it is hate speech or not. What is more, not every comment posted is hate speech. One has to browse many videos' comment sections to collect enough hateful comments, as Lee et al. had done, a dataset with only 1580 comments, and 980 of it are from the SBIC dataset which is another hate comment dataset, meaning that they only collected 600 hateful comments from YouTube. We can use a YouTube comment dataset as well, as presented in the YouTube Dataset of 245 million comments [8]. Each entry in this dataset shows the channel ID, which can be traced to a channel. If we make a list of channel IDs which we intend to retrieve comments from, we will be able to get a considerable size of text.

- Twitter. Like YouTube, it is not possible to know the nationality of users, unless the users are willing to show it publicly. Tweets can contain a hashtag such as "#unitedkingdom", "#Singapore", which is not an indicator of nationality, but a keyword written so that the tweet can be found when searching the corresponding hashtags. To retrieve a lot of tweets, TwitterAPI shall be used[3].

- Reddit. It is an online content forum consisting of smaller units named **subreddit**. Examples are **car**, **bike**, **computer**. These subreddits include content relevant to the subreddit names, and users may join the discussion if they are interested in the topics. Fortunately, there is no need to manually copy any comment, since we already have a huge dataset [2] of billions of comments, in which the source subreddit is displayed for every comment. Therefore it is possible to retrieve comments from specific subreddits. It is therefore possible to retrieve comments from **australia**, **new zealand**, **singapore**, **southafrica**, **unitedkingdom**, **india**.

- News Websites. The comment sections of news websites like MSN News, and Yahoo News are good sources of text. Yet there is no available API to retrieve them, nor is there a way to know the nationalities of the comments. We have to manually search for news that is happening in these six nations, assuming that people leaving comments under the articles are mostly English speakers of the corresponding nations.

By comparing each platform's advantages and disadvantages, the Reddit dataset was selected, because it generated the most number of comments, and the fact that it displays which subreddit the comment is in makes the text-to-nationality mapping easier. However, this relies on the assumption that people leaving comments in subreddits of these six English-speaking nations are native speakers from these six nations (Most users in the India subreddit are Indians, most users in the New Zealand subreddit are New Zealanders et cetera). This assumption decides how good the accuracy of future analyses would be, as comments made by people who do not originate in these nations may distort the linguistic characteristics.

4.2.2. Data Filtering

The Reddit dataset [2] contains every publicly available comment from October 2007 to May 2015, segmented by months. Due to the dataset being extremely large (149 Gigabytes) and the computational source limit, we only picked data segments of six months: June 2012, December 2013, January 2014, August 2014, January 2015, and May 2015, among which there are in total 1519110 comments from the six English-speaking nations. We selected these six segments because we wanted to gather the text from the time that is closer to the present. Also, earlier segments (2007-2011) contain significantly fewer comments, making them less suitable for gathering sufficient data.

Each entry in the dataset has the following columns:

- **subreddit**: sections for the discussion of different topics: car, basketball, football, Australia.

- **body**: The comment, a piece of text.

- **controversiality**: 0 or 1, binary attribute, indicating whether the comment is controversial.

Therefore, it was easy to pick out comments from specific nations by checking the subreddit. If it is one of the six nations we selected, we copy them into the new CSV file for future use. The text in "body" remains as it is.

We used the HateXPlain BERT [20, 31] model to classify each comment as **hate speech/ offensive/normal**. The model classified 6502 comments as hate speech, with 3069 from **"australia"**, 975 from **"india"**, 769 from **"newzealand"**, 25 from **"southafrica"**, 1563 from **"unitedkingdom"**, 97 from **"singapore"**. The initial results showed a huge imbalance of data, with South Africa and Singapore having significantly fewer comments than other nations. To reduce bias, we added more sample comments from the CREHate dataset[27], where 150 social media comments from Australia, the United Kingdom, South Africa, and Singapore were manually collected which are deemed "hate speech" by the authors of the paper. By doing so, the size expanded to 7102, with four nations gaining 150 more comments each. The reason we added more entries to Australia and the United Kingdom instead of South Africa and Singapore only was that we believed the manually collected comments could diversify the source of data, being supplemental to the text judged by the HateXPlain BERT model as "hate speech", offering a more solid comprehension and universal pattern for the further analysis.

4.2.3. Understanding Lexical Differences

Firstly, we conducted a vocabulary analysis. It is the most basic analysis of text, or in simpler terms, word frequency analysis. We inspected the most frequent words in each nation's hate speech, to get a grasp of what the major topics were, as the top words could refer to topics such as ethnicity or race (e.g. *Pakis, Black, Asian*);

Secondly, we conducted lexical relation analysis, where we analyzed and distinguished five semantic relations: synonym, meronymy, hyponym, hypernym, and holonym. For every English word, it is possible to find words in each of these five categories. For example, for the word car, we can identify its synonym (vehicle), meronym (engine), hyponym (sedan), hypernym (transportation tool), and holonym (traffic). Thus, for each sentence, we calculated the sum of words of these five semantic categories by adding up the total amount of all lexical relations we retrieved from each word (e.g., "I like apples"), where we calculated the sum of words in five semantic relations, adding them up. English stopwords were removed, leaving out nouns, adjectives, adverbs, and the rest. Finally, we calculated the average count of each of the five semantic relations in the hate speech of each nation. This helped us understand the generality and specificity of the language.

Thirdly, we conducted pronoun analysis. The pronoun plays a major role in the object targeting of hate speech. It is the indicator of the receiver of hate. We compared the average count of all types of pronouns in each sentence separately in hate speech and average speech, and the percentage change in the use of each pronoun from average speech to hate speech. By undertaking these analyses, we identified the pattern of pronoun usage and how the pronoun was deployed to target different groups of targets, on a more abstract perspective, which is unrelated to either "meaning" or "topic".

Fourthly, we conducted the part-of-speech analysis. It turns any English sentence, including hate speech, into a list of part-of-speech tags (e.g. [PRP]-[VBP]-[VBG]-[NNS] for *I*

like eating apples). We calculated the average count of each POS tag in each sentence, and the percentage change in their usages from average speech to hate speech, of each nation. This analysis is crucial, for it shows how regional hate speech tends to be different in word selection and morphology (e.g. Nation A tends to use more past tense verbs, while Nation B tends to use more gerund verbs (-ing)) [38, p. 292-304].

4.2.4. Understand Syntactic Differences

We conducted syntactic analyses to examine the grammatical structure of language, in order to get insight into each nation's sentence structure, which is more fundamental than the topic and meaning.

The analyses were split into two parts: Sentence Length Analysis and Phrasal Analysis.

The sentence length analysis investigated the superficial complexity of sentences. We would like to see whether there is a notable length difference between nations (e.g. Nation A tends to use longer sentences in hate speech than B) and whether there is a huge difference between average and hate speech (e.g. Nation A's hate speech is on average much shorter than its average speech, while B is longer).

The phrasal analysis was designed to uncover regional differences in phrase usage. We determined the proportion of each phrase type with a nation's speech and compared it with others (e.g. In Nation A's speech, 20% of the phrases are Noun Phrases [NP], while only 15% of the phrases are Noun Phrases for Nation B). Additionally, we analyzed the syntax transitional probability, of the likelihood that a phrase will transit to another (e.g. There is a 25% chance that the Verb Phrase[VP] in Nation A's hate speech transits to an Adverb Phrase[ADVP]). By comparing the probabilities of hate speech and normal speech, we discerned the syntactic pattern of regional hate speech.

4.2.5. Improving The Hate Speech Detection System

We combined the previous experimental results and the comprehension of linguistics characteristics with the development of artificial intelligence models for hate speech detection. This helped us build syntactically and lexically aware classification models. For the model training, we used hateXplain [31] to gather non-hate speech and create a training dataset with 36077 comments from the six English-speaking nations, 7110 being hate speech with label 1 that were gathered in the data collection stage, and 28967 being non-hate speech with label 0 which were newly gathered.

Due to limited computational resources, we built only traditional machine learning classifiers, including a random forest classifier, decision forest classifier, and Neuro Symbolic (Long-Short-Term-Memory) classifier. These models are some of the most commonly used classifiers in natural language processing tasks. They are not resource-intensive, which makes them user-friendly. Among the three, we will introduce the LSTM classifier because it uses our custom structure:

- Input Layer, accepting the text input.
- Embedding Layer, transforming the input sequence to a 128-dimension vector.
- LSTM Layer, creating a 128-dimension vector from the input of the previous embedding layer.
- Dense Layer 1, with 128 neurons and ReLU activation.

- Dense Layer 2, with 64 neurons and ReLU activation.
- Output Layer, with 1 neuron and sigmoid activation.

We divided the model training into three parts. Firstly, we trained the text-only classifiers as the baseline, in which the comments and the hate speech labels were used. We call it a "text-only" model.

Secondly, we trained classifiers with the following six components, including categorical, lexical, and syntactic components that were analyzed in the first part of the research. We call it a "combined" model.

- The comment. This is the basic feature of model training.
- The part of speech tags. We converted the comments into a string of part of speech tags (e.g. I hate you ⇒[PRS][NNS][PRS]).
- The phrases. Similar to the part of speech tags, each comment was converted into a string of phrases (e.g. [NP, VP, ADVP, CONJP]), and then used for the model training.
- The sentence length.
- The nationality. There are nations' names in the **subreddit** column of the Reddit dataset.
- The pronoun count. We counted the total number of pronouns in each comment and used it as one of the features for model training.

Thirdly, we trained the model by utilizing the results from previous linguistic analyses, including the percentage change of the part-of-speech tags, the percentage change of the pronouns, and the syntactic transitional probability. We call it a "weighted" model. The way they were coded to the features is simple: for each comment, we calculated the occurrence of each pronoun and part of speech tag. We computed weighted occurrences based on the previous findings. For each occurrence, we computed two numbers, one named "pronoun_hate" "POS_hate", and another named "pronoun_normal" "POS_normal", where "pronoun" and "POS" were replaced by the actual words themselves. For example, in the sentence "I like you." from the "australia" subreddit, "I" appears once, and the results show that there is a 5% decrease in the frequency of "I" in "australia" hate speech compared to its average speech. We then set "I_hate" as $1 * (1 - 0.05) = 0.95$, and the "I_normal" remains 1. The same computation is performed for all part-of-speech tags and pronouns. For the syntax transitional probability, we first convert the comment into a syntax parse tree: [NP-VP-NP]. Since for each nation, the transitional probabilities from one syntax to another in both hate speech and average speech are given, we could then calculate the probability of the transition being hate and average speech (e.g. In "australia"'s hate speech, the transitional probability from NP to VP is 9%, and from VP to NP is 6%, while in its average speech, they are 8.5% and 8%. Therefore the overall transitional probabilities are $0.09 * 0.06 = 0.0054$ & $0.085 * 0.08 = 0.0068$)

We adopted this method for various reasons. Even though machine learning models can learn without the weighted occurrence, we were still interested in whether the custom weights can amplify the signal by adding features based on the previous analysis. We sought to discover the effectiveness of the integration of linguistic knowledge, and if it would help enhance the sensitivity to specific lexical and syntactic elements.

4.3. Ethical Concerns

There are two main ethical concerns: the leakage of privacy and the exposure to verbal attacks.

4.3.0.1. Leakage Of Privacy

Privacy is an important consideration in the research, especially in the processing of data. There are two possible leakages: the senders' usernames and geolocations, and the private information in the comment. The first leakage refers to knowing the personal information of the comment sender, and the second leakage refers to any kind of private information mentioned in the comment.

By the nature of the dataset, the leakage of usernames is highly unlikely, as there is no username entry in the dataset. The geolocations are nearly impossible to leak either, for there is no direct connection between the subreddit's name and the real geolocation. However, the second type of leakage is possible, though unlikely, as anyone can share any type of information online, of either their own or someone else's (e.g. the passport number, the bank account number and balance, the username, geolocations), and there is no way to prevent this from happening. Thus, corresponding measures will be taken: when there is the need to showcase the data, we will only use the comments that contain no personal information; we will not share or keep comments if they include personal information or if they include enough details to identify the people.

4.3.0.2. Exposure to Verbal Attack

There is no volunteer or participant needed for the research. However, some hate speech was shown for demonstration purposes to the readers, which is seen as offensive or distressing by some people. The asterisk mark * was used to replace part of the speech to prevent direct exposure and such text was shown as few times as possible to minimize the risk. (e.g. F**K, C*NT, D**KHEAD)

For the researcher of this thesis, exposure to reading and analyzing hate speech can impact mental health. Previous research shows that some social media content moderators were diagnosed with PTSD (Post Traumatic Stress Disorder)[40, p. 3]. They report impaired psychological wellness due to having to watch unnerving videos about murders, beheadings, and similar violent content [40, p. 3]. To moderate this negative influence, the hateful text will only be read when sample texts are needed. The automated processes will handle most text processing without human intervention. Also, when there is a need to read many hate texts, periodic rests will be given to the researcher for relief.

4.3.0.3. Imbalanced Representation of Minority Groups

This research evaluates the results on only a national level and may overgeneralize the hate speech differences. The artificial intelligence system trained with these data can be less effective in handling minority groups' speech due to the nature of the dataset (only 'subreddit' for identification), causing bias when deployed for use. The best measure for countering such bias is to only apply the newly trained models to other Reddit comments of the six English-speaking nations, maximizing the accuracy.

5

Experimental Results

5.1. Results Of Linguistic Analyses

5.1.1. Results Of Lexical Analysis

5.1.1.1. Results Of Vocabulary Analysis

We use word clouds for the analysis, which show the 50 most frequent words in each nation's hate speech.

We see a large overlapping in the most frequent vocabularies, and prevalent usage of coarse and bawdy language in all nations' speech, such as "**c**t**", "**f**k**", "**sh*t**". In the top ten most frequent words of each nation's hate speech, five words are the same; The word **stupid** appears in four nations, **ass** appears in three nations. The word frequency indicates a superficial similarity. Overall, the words related to violence, nations, regional slang, and race are prevalent, with each nation leaning towards certain topics.

For South Africa, the top words are "**boer**", "**terrorist**", "**kill**", "**c**t**", "**f**k**", "**people**", "**white**", "**black**", "**monkey**", suggesting its hate speech having a strong connection with racial topics, as seen in comments like "*Nothing better than when Americans...Release the black youth from your prisons...shut the fuck up and just keep making movies.*", "*When Penny Sparrow wrote about blacks as monkeys she was fined R150000.*", and the words "**kill**", "**terrorist**" are indicating aggression and violence. This is seen in sentences like "*Kill yourself race traitor scum.*", "*No really, kill yourself you f** ken fa*get pigf**ker.*", either asking people to kill themselves, or claiming that some people acted killing, and in comments like "*ANC doesn't care about our safety and security everyone just flock to SA even terrorists are here.*", "*Nobody...that the ANC are nothing more than a bunch of terrorists and have been terrorizing the people of this nation for decades.*", convey the idea of some organization/groups of people being terrorists.

For the United Kingdom, the top words are: "**c**t**", "**people**", "**f**k**", "**want**", "**make**", "**right**", "**time**", "**say**", "**wa**", suggesting a similar preference in using vulgar words.

For Singapore, the top words are "**china**", "**chinese**", "**gay**", "**slur**", "**people**", "**malay**", "**singapore**", "**user**", Comments like "*I giggle when I hear China struggles*", "*not USA , killing people for nothing, quenching for blood, China will takeover Usa soon.*", "*When she pulls the I have Malay and Indian friends card you know what's going down*" demonstrate a clear focus on ethnically specific topics.

For New Zealand, the top words are "**c**t**", "**people**", "**f**king**", "**s**t**", "**kiwi**", "**day**",

Figure 5.1: Word Clouds Of **Australia** (Top Left), **United Kingdom** (Top Right), **New Zealand** (Middle Left), **Singapore** (Middle Right), **South Africa** (Bottom Left), **India** (Bottom Right)

"**time**", "**want**", "**good**", suggest a similar pattern in using vulgar words, and the regional terms like "*kiwi*" in hate speech.

For India, the top words are "**india**", "**c**t**", "**ha**", "**people**", "**fuck**", "**muslim**", "**hindu**", "**paki**", "**law**", "**make**", indicating its preference in religion and ethnicity. Examples are "*The Pakistani people are idiots.*", "*pakis can piss off*", and "Sir what are your views on Hindus who blame Indian Muslims for problems created by Pakistanis?"

For Australia, the top words are "**c**t**", "**people**", "**f**king**", "**s**t**", "**bogan**", "**australia**", "**wa**", "**time**", "**want**", "**right**", showing similarity to New Zealand's pattern.

These results prove our hypothesis of the lexical differences, that there is a significant overlapping in the vocabulary used in the hate speech of all nations.

5.1.1.2. Results Of Part-Of-Speech Tag Analysis

We used NLTK, "Natural Language Toolkit"[6] for processing the text. This Python library tokenizes a piece of text and turns it into a list of part-of-speech tags. We calculated the amount of each tag in one nation's speech, divided it by the total number of comments of the nation, and got the average count of each type of tag in each comment of the nation.

We set n as the number of comments from a certain nation c of a specific POS tag t. $C_{t,k}$ is the count of tag t in comment k, the average $\text{Avg}_{t,c}$ of tag t for nation c is calculated with the formula:

$$\text{Avg}_{t,c} = \frac{1}{n} \sum_{k=1}^{n} C_{t,k}$$

This calculation is performed for each POS tag and each nation, and it results in a mapping of each nation to the average occurrence of each POS tag in its hate comments.

The table [5.1] presents the average frequency of each type of part of speech tag.

POS Tag	Australia	UK	India	New Zealand	Singapore	South Africa
CC (e.g., "and", "but", "or")	**1.18**	1.16	1.09	1.06	0.67	0.46
CD (e.g., "one", "two", "3")	**0.14**	0.12	0.10	0.12	0.07	0.02
DT (e.g., "the", "a", "an")	**3.84**	3.64	3.15	3.22	1.99	1.98
EX (e.g., "there is")	**0.08**	0.07	0.07	0.05	0.05	0.02
FW (e.g., "Déjà vu")	0.01	0.01	**0.02**	0.02	0.01	0.02
IN (e.g., "in", "on", "because")	**4.11**	3.96	3.54	3.62	2.27	1.93
JJ (e.g., "big", "quick", "blue")	**2.84**	2.61	2.59	2.49	1.87	1.33
JJR (e.g., "biggest", "quickest", "bluest")	0.13	**0.14**	0.10	0.13	0.09	0.04
JJS (e.g., "1.", "2.", "A.", "B.")	**0.10**	0.07	0.07	0.08	0.07	0.05
MD (e.g., "can", "will", "could", "would")	0.49	**0.51**	**0.51**	0.46	0.30	0.28
NN (e.g., "dog", "car", "music")	**5.88**	5.36	5.61	5.29	4.22	2.93
NNP (e.g., "dogs", "cars", "bottles")	1.97	1.78	**2.26**	1.80	1.85	1.85
NNPS (e.g., "John", "London")	**0.05**	0.04	0.03	0.02	0.03	**0.05**
NNS (e.g., "Americans", "Indians")	**2.31**	2.19	2.01	2.01	1.31	1.16
PDT (e.g., "all the", "both the")	**0.05**	0.04	0.04	**0.05**	0.01	0.03
POS (e.g., "'s", "'")	0.0003	0.0000	**0.0010**	0.0000	0.0000	0.0000
PRP (e.g., "I", "you", "he", "she", "it")	2.27	**2.26**	1.92	2.06	1.40	1.33
PRP$ (e.g., "my", "your", "his", "her", "its")	**0.69**	0.65	0.68	0.59	0.42	0.29
RB (e.g., "quickly", "softly", "well")	**2.07**	2.06	1.68	1.79	1.53	1.08
RBR (e.g., "better", "faster")	**0.07**	0.06	0.05	0.05	0.02	0.02
RBS (e.g., "best", "fastest")	**0.02**	0.02	0.01	0.01	0.02	0.01
RP (e.g., "up", "off", "over")	**0.27**	0.25	0.19	0.24	0.12	0.07
SYM (e.g., "$", "%", "&")	0.0	0.0	0.0	0.0	0.0	0.0
TO (e.g., "to go", "to buy")	**0.98**	0.95	0.90	0.86	0.68	0.40
UH (e.g., "uh", "wow", "ouch")	0.02	0.02	**0.03**	0.02	0.01	**0.03**
VB (e.g., "take", "run")	1.77	**1.77**	1.66	1.64	1.28	1.01
VBD (e.g., "took", "ran")	**0.81**	0.77	0.66	0.66	0.43	0.48
VBG (e.g., "taking", "running")	0.96	**0.98**	0.84	0.92	0.51	0.54
VBN (e.g., "taken", "run")	0.63	**0.69**	**0.69**	0.54	0.34	0.37
VBP (e.g., "take", "run")	1.69	**1.71**	1.51	1.59	1.17	0.95
VBZ (e.g., "takes", "runs")	1.06	1.00	**1.10**	0.86	0.73	0.64
WDT (e.g., "which", "that")	0.18	0.17	0.12	**0.15**	0.11	0.05
WP (e.g., "who", "what")	0.27	**0.30**	0.30	0.24	0.14	0.19
WP$ (e.g., "whose")	**0.002**	0.001	0.001	0.001	0.000	0.000
WRB (e.g., "where", "when")	**0.26**	0.24	**0.26**	0.20	0.22	0.21

Table 5.1: Average POS tag frequencies in hate speech across six nations

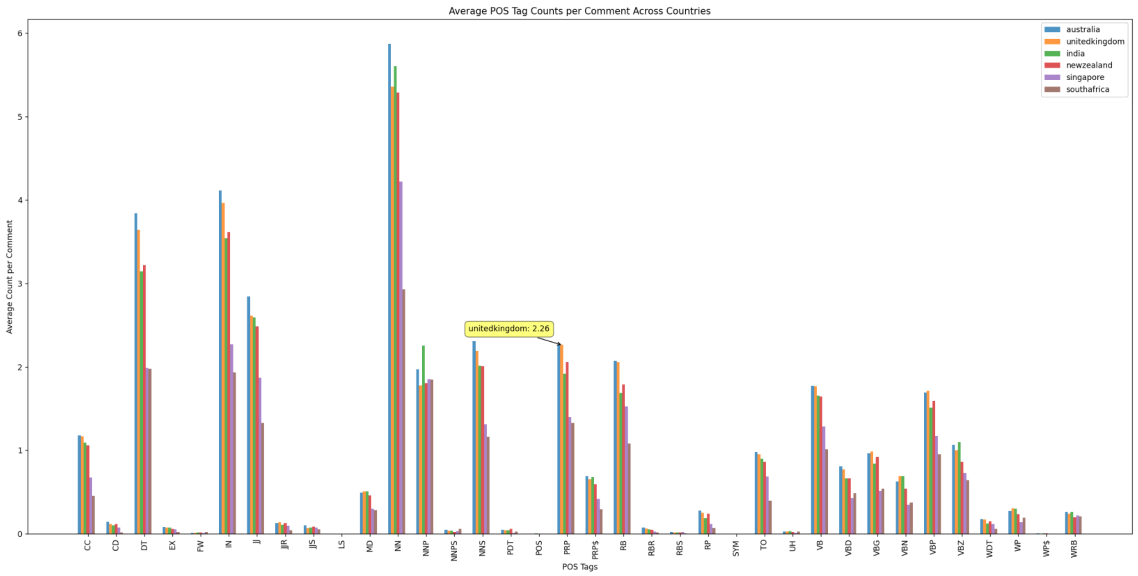


Figure 5.2: Bar Chart of Average Frequency of POS tags In Hate Speech

Overall, all six nations show high frequencies in the usage of most common POS tags including Nouns (NNP, NNPS, NN), Verbs (VB, VBD, VBG, VBN, VBZ), Pronouns (PRP, PRP$), indicating that a strong focus on entities and actions. For noun usage, the average frequency of singular nouns (NN) is the highest for all nations, with Australia having 5.88 and South Africa having 2.93. The proper nouns (NNP, "London", "Jack") are most frequent for India (2.26), indicating more references to certain people or entities in Indian hate speech. Base form verbs (VB) are most frequent in the UK and Australia (1.77) and least frequent in South Africa (1.01). Present tense verbs (VBP, VBZ) are more common in Australia and India compared to other countries.

It is noticed that the average frequency of each type of POS tag is generally consistent. South Africa and Singapore consistently stand out with the lowest numbers in almost all categories except the Plural Proper Noun (NNPS, 0.05, South Africa). This phenomenon suggests a significant difference in the average length of hate speech, as shorter sentence length means fewer words are used, which will be verified later in the syntactic analyses part. It is also obvious that Australia and the United Kingdom are the most frequent in almost all POS tags, which means these two nations use more varied language in their hate speech.

Given these results, we then performed an analysis of the percentage difference in POS tag usage from average speech to hate speech. We defined $\text{count}_{t,c,k}$ as the count of POS tag t in comment k from nation c in a dataset. First, we computed the normalized count for each tag within each nation's dataset by the total count of each tag across all comments and then divided it by the total number of POS tags found in those comments:

$$\text{Normalized}_{t,c} = \frac{\sum_{k=1}^{n} \text{count}_{t,c,k}}{\sum_{k=1}^{n} \text{Total}_{c,k}}$$

$\text{Total}_{c,k}$ is the total count of all recognized POS tags in comment k of the nation c, and n is the total number of comments in the nation c.

Next, we calculated the percentage difference in the normalized counts between the

normal and hate speech datasets for each tag and each nation:

$$\text{Percentage Difference}_{t,c} = \left(\frac{\text{Normalized}_{t,c}^{\text{hate}} - \text{Normalized}_{t,c}^{\text{normal}}}{\text{Normalized}_{t,c}^{\text{normal}}} \right) \times 100\%$$

The analysis of the percentage difference in POS tag usage from average speech to hate speech, in figure [5.2] below, offers deeper insights.

POS Tag	Australia	UK	India	New Zealand	Singapore	South Africa
CC (e.g., "and", "but", "or")	-1.64	-10.51	**11.57**	0.38	-40.50	-69.39
CD (e.g., "one", "two", "3")	-0.44	-21.04	**-9.28**	-9.89	-38.05	-89.72
DT (e.g., "the", "a", "an")	4.72	-8.50	**15.94**	1.10	-30.73	-51.36
EX (e.g., "there is")	**-17.95**	-32.12	-12.28	-39.22	-41.87	-83.01
FW (e.g., "Déjà vu")	-1.69	-39.34	-19.56	**60.26**	-32.13	9.58
IN (e.g., "in", "on", "because")	-4.13	-16.05	**4.19**	-6.05	-41.36	-61.41
JJ (e.g., "big", "quick", "blue")	3.39	-9.85	**17.51**	3.28	-26.18	-58.49
JJR (e.g., "bigger", "quicker", "bluer")	**-22.74**	-24.01	-5.02	-15.01	-38.58	-78.30
JJS (e.g., "biggest", "quickest", "bluest")	**9.32**	-33.17	1.53	1.11	-23.13	-60.14
LS (e.g., "1.", "2.", "A.", "B.")	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
MD (e.g., "can", "will", "could", "would")	-10.65	-16.50	**13.28**	-4.74	-39.57	-55.34
NN (e.g., "dog", "car", "music")	2.58	-8.73	**14.29**	3.13	-17.96	-53.75
NNP (e.g., "John", "London")	**-4.80**	-14.92	-7.40	-11.72	-19.24	-38.49
NNPS (e.g., "Americans", "Indians")	**22.37**	-25.89	-8.19	-15.56	-22.85	-26.98
NNS (e.g., "dogs", "cars", "bottles")	11.93	0.77	**22.32**	14.74	-26.08	-48.40
PDT (e.g., "all the", "both the")	16.73	-3.58	20.40	**53.08**	-77.95	-41.57
POS (e.g., "'s", "'")	120.47	-100.00	**486.46**	-100.00	-100.00	-100.00
PRP (e.g., "I", "you", "he", "she", "it")	-1.75	-10.21	**1.47**	-5.38	-38.37	-54.29
PRP$ (e.g., "my", "your", "his", "her", "its")	-0.53	-10.65	**23.84**	-6.42	-43.81	-65.41
RB (e.g., "quickly", "softly", "well")	**-10.30**	-21.12	-3.48	-15.44	-34.05	-61.96
RBR (e.g., "better", "faster")	**-11.66**	-37.43	-5.70	-38.67	-68.58	-82.70
RBS (e.g., "best", "fastest")	10.98	-27.68	-1.21	-24.19	**12.78**	-78.40
RP (e.g., "up", "off", "over")	27.25	13.10	**37.91**	15.46	-43.19	-70.76
SYM (e.g., "$", "%", "&")	-100.00	-100.00	-100.00	-100.00	-100.00	**0.00**
TO (e.g., "to go", "to buy")	-9.54	-19.36	**14.83**	-9.41	-31.37	-66.46
UH (e.g., "uh", "wow", "ouch")	-5.52	-6.71	**64.06**	-21.26	-68.79	7.73
VB (e.g., "take", "run")	-3.38	-11.57	**14.63**	-0.09	-27.96	-51.23
VBD (e.g., "took", "ran")	**-8.74**	-21.97	-14.81	-21.64	-46.89	-53.49
VBG (e.g., "taking", "running")	-1.12	-5.78	**19.59**	4.70	-36.26	-47.48
VBN (e.g., "taken", "run")	-19.33	-21.35	**-2.25**	-19.81	-40.85	-53.97
VBP (e.g., "take", "run")	7.30	-0.16	**16.51**	10.97	-23.12	-50.72
VBZ (e.g., "takes", "runs")	**-9.34**	-17.66	5.72	-15.16	-23.15	-52.33
WDT (e.g., "which", "that")	-12.48	-22.81	-12.88	-13.44	-20.99	**-73.82**
WP (e.g., "who", "what")	13.43	14.33	**42.01**	15.50	-30.76	-27.19
WP$ (e.g., "whose")	**19.26**	-74.91	-63.15	-10.21	-100.00	-100.00
WRB (e.g., "where", "when")	-5.18	-19.37	**14.02**	-17.70	-14.05	-29.30

Table 5.2: Percentage difference in POS tag frequency from average speech to hate speech across six nations (in percentage)

Most nations show a drastic decrease in various categories, with Singapore and South Africa showing on average the most decrease among the six nations. Australia, the United Kingdom, and New Zealand show a decrease in most categories while India leans toward a significant increase in many categories. South Africa shows the highest decreases across most POS tags. (**Results on.1 page**)

We analyzed each nation for a better understanding of the differences:

- **Australia**: Possessive endings (POS) raised by 120.47%, indicating more reference to abstract and physical possession of people and entities (e.g. "*John's, worker's*"). List

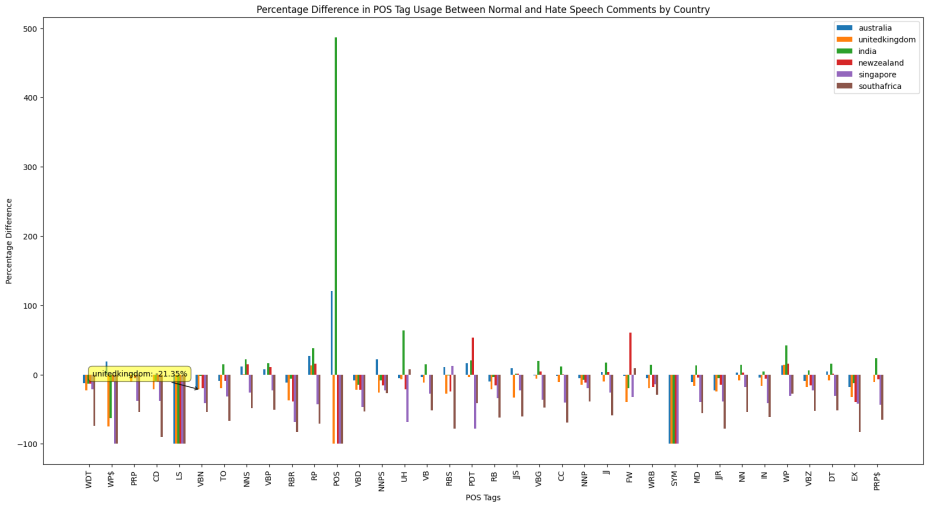


Figure 5.3: Percentage Difference of POS Tag Categories from average speech to hate speech

items (LS) and symbols (SYM) are completely missing, indicating that there is no listing of items in the hate speech (e.g. "*You have three choices: A. Buy the phone B. Buy the game console, C. Buy the concert ticket.*")

- **United Kingdom**: Possessive Wh-pronouns (WP$, e.g. whose), superlative pronouns (e.g."*biggest*"), comparative adjectives (JJR)(e.g. "*bigger*") are sharply decreased, indicating fewer comparisons in its hate speech. Possessive endings (POS), list items (LS), and symbols (SYM) are missing, suggesting a missing item listing and reference to the possession of people or entities.

- **India**: Possessive pronouns (PRP$)("'s"), and basic adjectives (JJ) are sharply increased, indicating more reference to the physical or abstract possession of people or entities(e.g. *"f**k of co**ie swine, keep eating ob*ma's s**t"*, *"I disagree with Bhusan's stand, however, freedom of speech, b**ch!"*), and more descriptions of the characteristics. However, the increase in interjection may indicate various changes dependent on the type of interjection [10]: The primary interjection is standalone words or sounds (e.g. *"Ouch"*, *"Wow"*), the secondary interjection is words with meaning and can express the mental attitude of state(e.g. *"Help!"*), the expressive interjection are vocal gestures (e.g. "Eww!"). It is not clear which type of these interjections are most prevalent.

- **New Zealand**: Predeterminer (PDT, e.g."all the", "both the") and foreign words (FW) sharply increased, indicating more references towards whole groups of people or entities, and a preference to adopt non-English words in hate speech.

- **Singapore**: Possessive Wh-pronouns (WP$), present participle verbs (VBG), comparative adverb (RBR), preposition (Preposition or subordinating conjunction), interjection (UH, e.g. "wow"), and possessive pronouns (PRP$) are sharply decreased.

- **South Africa**: For almost all POS categories, there is a sharp decrease, especially for possessive Wh-pronouns (WP$), possessive endings (POS), and foreign words (FW).

We utilized the previous results to calculate the Euclidean Distance to better measure their similarities. Euclidean distance helps measure the differences between samples in high dimensional space and better define the difference than the absolute difference[25]. We converted them into a matrix where each row is a nation and each column is a type of POS tag. If the data for any tag is missing, we set that number as 0 as a filler number.

$$\textbf{POS Tag Average Frequency Matrix (PTAFM)} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix}$$

f_{ij} represents the average frequency of the j-th POS tag in the i-th nation, m is the number of countries, and n is the number of POS tags.

We used this formula to compute the Euclidean distance between any two countries i and k , with their corresponding POS tag frequency vectors from **PTAFM**. The distance f_{ik} is calculated through the formula:

$$f_{ik} = \sqrt{\sum_{j=1}^{n} (f_{ij} - f_{kj})^2}$$

This calculation is performed for every pair of nations, ending up in an Euclidean Distance Matrix **EDM**.

$$\textbf{Euclidean Distance Matrix(EDM)} = \begin{bmatrix} 0 & f_{12} & \cdots & f_{1m} \\ f_{21} & 0 & \cdots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \cdots & 0 \end{bmatrix}$$

In which each (f_{ik} represents the Euclidean distance between the POS tag frequency vectors of nation_i and nation_k. The heat map representation is shown as follows:

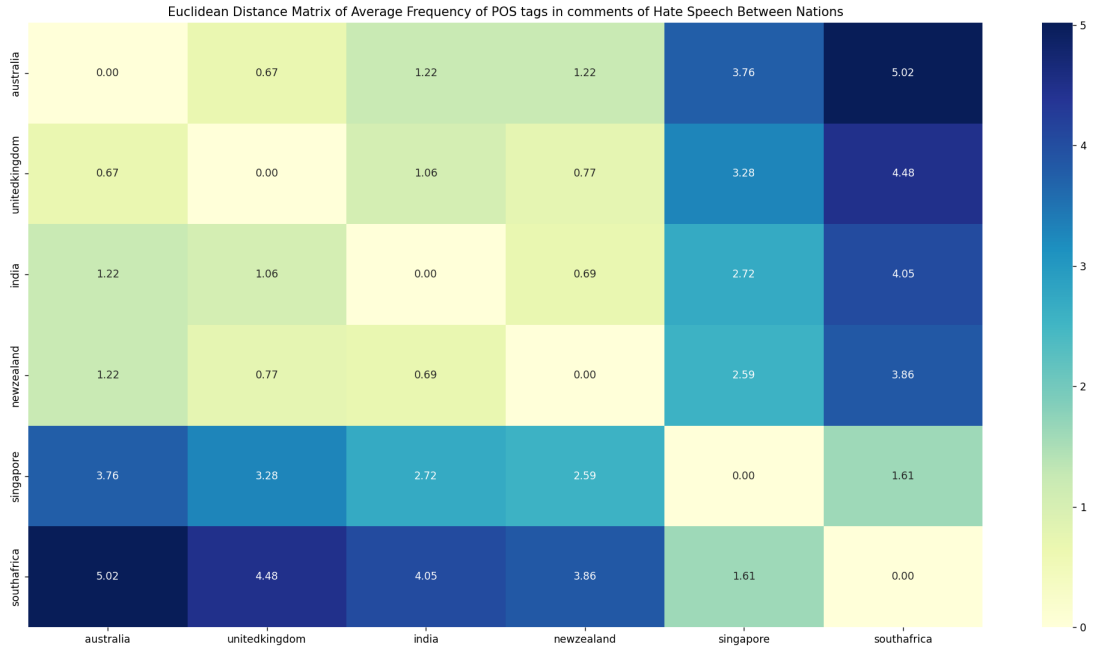


Figure 5.4: Euclidean Distance Matrix of Average Frequency of POS tags in comments of Hate Speech Between Nations

Figure [5.4] shows that judging by the Euclidean distances of average frequency of part-of-speech tags, **Australia**, **United Kingdom**, **India**, **New Zealand** are similar in the POS tag usage, suggesting similar language style in hate speech, while Singapore, South Africa are distant from the majority, and the distance between them is even further than between any of the first four nations;

Australia being close to the United Kingdom, Australia, and New Zealand is not surprising, considering their shared history and cultural background[35]. Surprisingly, **India** is closer to **New Zealand** than to any other nation, and vice versa. This reveals an unrealized similarity in the discourse style of hate speech between India and New Zealand.

Using the same method, we calculated the Euclidean distances of each nation's percentage difference in POS tags from average speech to hate speech, to see how similarly they change when writing hate speech (e.g. One nation changes more drastically than the others). The results are displayed in a heatmap: (**Figure Below**)

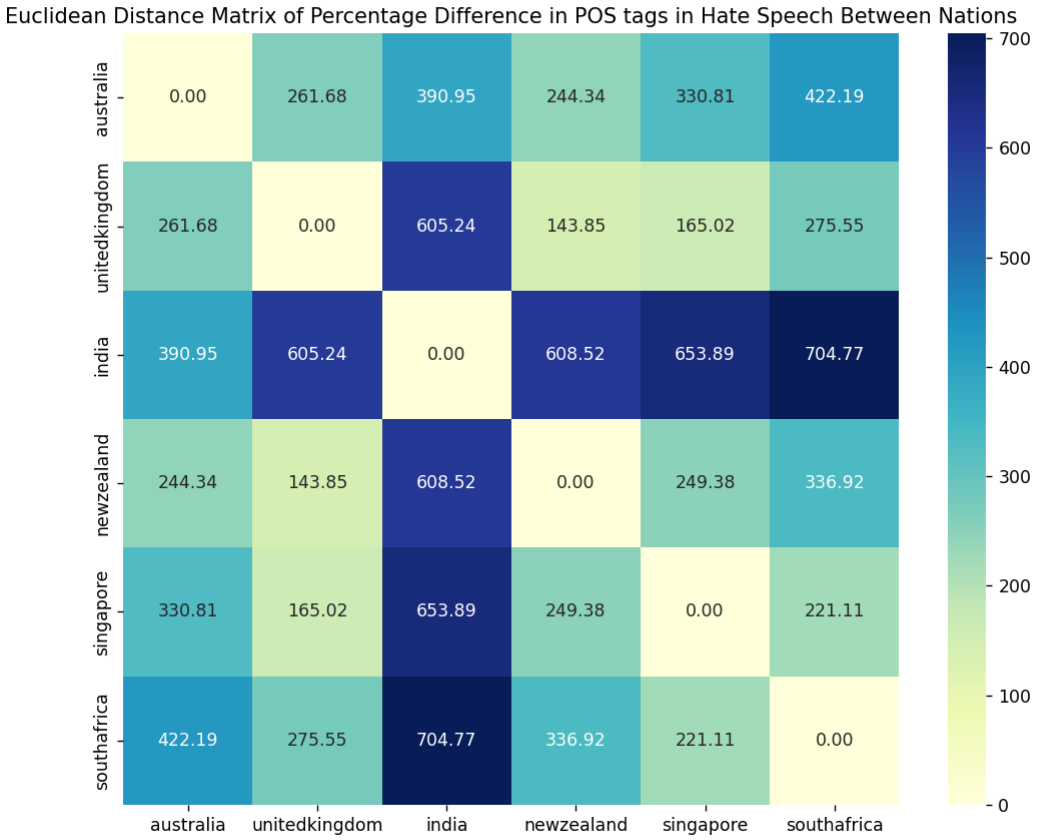


Figure 5.5: Euclidean Distance Matrix of Percentage Difference in POS tags in Hate Speech Between Nations

Figure [5.5] indicates that **India** has the most drastic change in the POS tag usage, with at most 704.77 distance and at least 390.95 distance, suggesting a huge difference in the language style between its average speech and hate speech. This also indicates a fact, that Indian hate speech tends to use similar language to the predominantly English-speaking nations (Australia, New Zealand, United Kingdom), heavily modifying the style of their average speech. We see that the shortest distance is **143.85** between **New Zealand** and **United Kingdom**, and the average Euclidean distance to other nation is 374.2835, meaning each nation is changing in a considerably different way when they shift from the average speech to hate speech.

The results above prove our hypothesis about the proximity of Australia and New Zealand's Part-Of-Speech Tag usage pattern.

5.1.1.3. Pronoun Analysis

We calculated the percentage of pronouns in the hate speech for each nation c and pronoun p, with the following formula:

$$\text{percentage}(n,p) = \frac{\text{pronoun\_count}(n,p)}{\text{total\_pronouns}(n)} \tag{5.1}$$

in which

- pronoun_count(n, p) is the count of pronoun p in nation n
- total_pronouns(n) is the total count of all pronouns in nation n

If any pronoun is missing, that is pronoun_count(n, p) = 0, we set:

$$\text{percentage}(n, p) = 0 \tag{5.2}$$

To avoid division by zero.
The following table shows the percentage in decimal numbers of each pronoun:

Pronoun	Australia	UK	India	New Zealand	Singapore	South Africa
i	0.0945	0.0898	0.0739	0.1172	0.1232	0.0823
you	0.0936	0.0987	0.1060	0.1134	0.1258	0.1265
he	0.0187	0.0217	0.0246	0.0109	0.0174	0.0221
she	0.0034	0.0035	0.0064	0.0023	0.0027	0.0020
it	0.0739	0.0755	0.0592	0.0739	0.0602	0.0803
we	0.0352	0.0309	0.0334	0.0338	0.0321	0.0482
they	0.0590	0.0586	0.0497	0.0450	0.0402	0.0703
me	0.0149	0.0152	0.0136	0.0178	0.0295	0.0181
him	0.0078	0.0058	0.0081	0.0057	0.0054	0.0080
her	0.0039	0.0034	0.0057	0.0046	0.0067	0.0020
us	0.0095	0.0096	0.0132	0.0103	0.0067	0.0100
them	0.0250	0.0317	0.0264	0.0278	0.0308	0.0120
my	0.0187	0.0148	0.0172	0.0238	0.0241	0.0080
your	0.0263	0.0244	0.0328	0.0304	0.0281	0.0361
his	0.0114	0.0114	0.0152	0.0077	0.0094	0.0080
its	0.0343	0.0361	0.0238	0.0307	0.0402	0.0181
our	0.0135	0.0095	0.0189	0.0158	0.0107	0.0181
their	0.0275	0.0303	0.0354	0.0203	0.0228	0.0181
mine	0.0005	0.0005	0.0002	0.0009	0.0013	0.0000
yours	0.0003	0.0001	0.0007	0.0000	0.0000	0.0040
hers	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ours	0.0002	0.0001	0.0002	0.0000	0.0000	0.0000
theirs	0.0001	0.0005	0.0000	0.0003	0.0013	0.0000
myself	0.0009	0.0009	0.0007	0.0017	0.0000	0.0020
yourself	0.0021	0.0015	0.0009	0.0020	0.0027	0.0080
himself	0.0010	0.0005	0.0011	0.0009	0.0000	0.0000
herself	0.0001	0.0001	0.0002	0.0000	0.0000	0.0020
itself	0.0007	0.0006	0.0024	0.0003	0.0013	0.0020
ourselves	0.0004	0.0003	0.0004	0.0003	0.0000	0.0000
yourselves	0.0002	0.0003	0.0002	0.0000	0.0013	0.0000
themselves	0.0034	0.0024	0.0035	0.0009	0.0040	0.0040
this	0.0479	0.0418	0.0565	0.0390	0.0549	0.0743
that	0.0869	0.0938	0.0765	0.0899	0.0803	0.0823
these	0.0181	0.0167	0.0297	0.0115	0.0174	0.0161
those	0.0129	0.0126	0.0112	0.0178	0.0094	0.0060

Table 5.3: Average Pronoun Usage In Hate Speech Comments By Nations

Pronoun	Australia	UK	India	New Zealand	Singapore	South Africa
who	0.0275	0.0306	0.0312	0.0229	0.0187	0.0281
whom	0.0002	0.0002	0.0013	0.0003	0.0000	0.0000
whose	0.0005	0.0001	0.0002	0.0003	0.0000	0.0000
which	0.0098	0.0096	0.0103	0.0100	0.0107	0.0060
what	0.0251	0.0290	0.0306	0.0281	0.0268	0.0422
anyone	0.0036	0.0050	0.0040	0.0014	0.0000	0.0060
anybody	0.0003	0.0003	0.0007	0.0011	0.0000	0.0000
anything	0.0034	0.0042	0.0048	0.0037	0.0027	0.0040
everyone	0.0039	0.0050	0.0026	0.0034	0.0027	0.0080
everybody	0.0004	0.0003	0.0013	0.0006	0.0013	0.0000
everything	0.0022	0.0015	0.0015	0.0006	0.0013	0.0040
someone	0.0072	0.0084	0.0033	0.0074	0.0000	0.0020
somebody	0.0003	0.0014	0.0011	0.0009	0.0000	0.0000
something	0.0069	0.0066	0.0062	0.0066	0.0040	0.0080
no one	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
nobody	0.0009	0.0009	0.0009	0.0006	0.0013	0.0020
nothing	0.0057	0.0050	0.0046	0.0043	0.0107	0.0120
each	0.0025	0.0018	0.0022	0.0009	0.0013	0.0000
either	0.0025	0.0031	0.0022	0.0023	0.0067	0.0020
neither	0.0008	0.0005	0.0002	0.0006	0.0000	0.0000
one	0.0186	0.0183	0.0198	0.0178	0.0147	0.0040
another	0.0050	0.0040	0.0048	0.0032	0.0040	0.0020
other	0.0138	0.0116	0.0103	0.0129	0.0107	0.0040
others	0.0013	0.0027	0.0029	0.0014	0.0013	0.0000
such	0.0032	0.0061	0.0075	0.0043	0.0094	0.0020
all	0.0314	0.0324	0.0323	0.0352	0.0214	0.0382
any	0.0117	0.0115	0.0112	0.0137	0.0094	0.0000
more	0.0191	0.0179	0.0161	0.0180	0.0094	0.0100
most	0.0100	0.0076	0.0062	0.0072	0.0134	0.0040
some	0.0242	0.0186	0.0196	0.0221	0.0134	0.0120
few	0.0054	0.0051	0.0035	0.0049	0.0040	0.0080
many	0.0053	0.0058	0.0079	0.0069	0.0107	0.0020
several	0.0004	0.0009	0.0002	0.0000	0.0000	0.0000
each other	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
one another	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5.4: Average Pronoun Usage In Hate Speech Comments By Nations (Part 2)

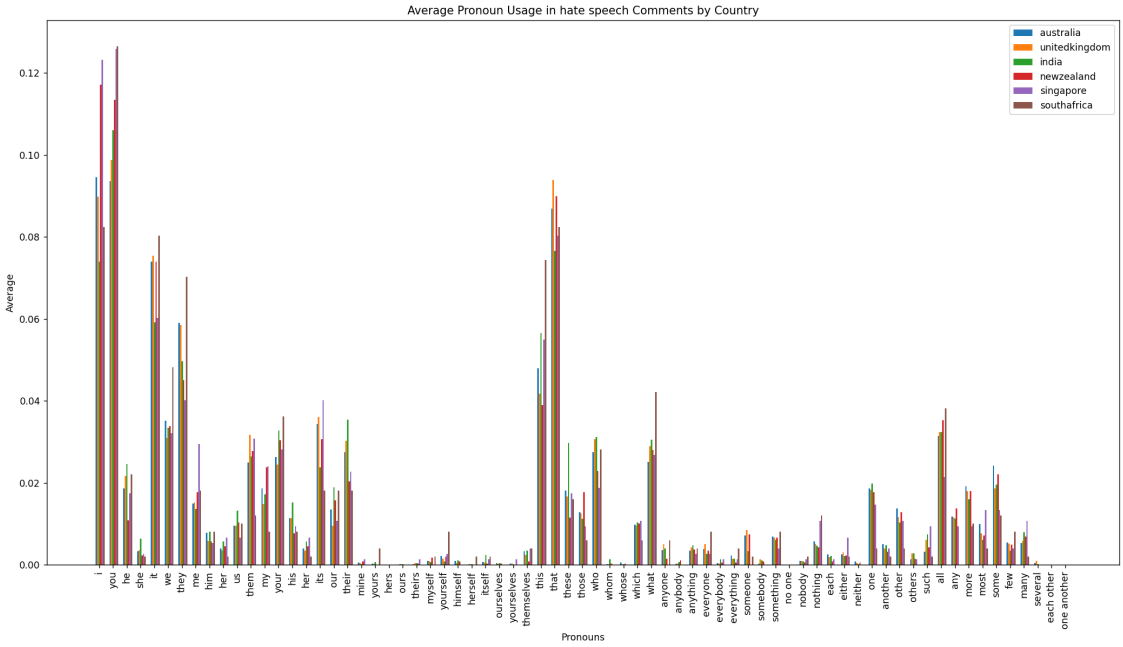


Figure 5.6: Average Pronoun Usage In Hate Speech Comments By Nations (In Percentage By Decimal Number)

We see that personal pronouns like *i, you, he, it, they* are frequently used; while the usage of feminine pronouns like *she, her, herself* is much less, which may suggest less direct targeting towards women, while it may also suggest that words like "women""lady" are replacing the feminine pronouns; the usage of possessive pronouns are notably high, like *your, his, its, our, their, my*, indicating critiques of certain people/entities like "*We dislike our government favoring those who take illegal means*", or comparison like "*It's our fault not theirs*". However, the reflexive pronouns, especially *myself, ourselves, yourselves, itself* are much less being used, while *themselves, yourself* are comparatively higher (0.0034, 0.0021), suggesting fewer self-references; The higher usages of interrogative (*who, which, what*) and demonstrative (*this, that*) pronouns lead to two possible explanations: 1, There are more clauses used in the hate speech, where the pronouns are used for mentioning new entities or people; 2, There are more questionings in the hate speech.

Besides these similarities among all nations, there are regional variations that make their hate speeches unique:

- South Africa and India show imbalanced usages of **i** and **you**, with the percentage of "i" being significantly lower than "you". The other nations use them nearly equivalently. This suggests less self-reference and a stronger blaming or targeting of others.
- The usage of "*this, that, these, those*" indicates a preference for targeting a singular person/entity, with a higher preference for "that" than "this". However this does not always suggest further targeting in the hate speech, but sometimes the prevalence of that-clauses (e.g. *This is the car that was sold.*).

To have a clearer interpretation of the data, we once again calculated the Euclidean distances of the pronoun usage pattern of each nation, resulting in the following heat map:

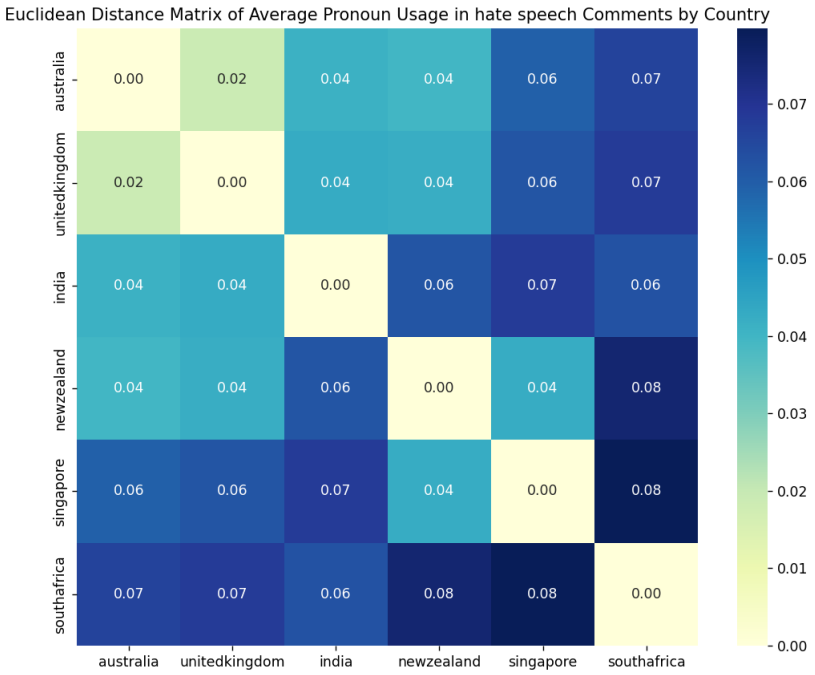


Figure 5.7: Euclidean Distance Matrix Of Average Pronoun Usage In Hate Speech Comments By Nations

From the heat map, it is easy to tell that Australia and the United Kingdom are once again closer to each other than to any other nation. India and New Zealand are the second closest to Australia and the United Kingdom. At last, every nation has the longest Euclidean distance with either Singapore or South Africa, suggesting their unique patterns of pronoun usage in hate speech.

Another analysis is performed on the difference in pronoun percentage between average speech and hate speech. As we already know the percentages of each pronoun in the hate speech, and by computing the percentages of each pronoun in the average speech, we are able to know the percentage difference between the two types of speeches. The total count of pronouns across all comments is calculated with the following formula:

$$\text{total}_{\text{nation}} = \sum_{\text{pronoun}} \text{count}_{\text{nation, pronoun}}$$

The average number of each pronoun appears per comment is computed with:

$$\text{average}_{\text{nation, pronoun}} = \frac{\text{count}_{\text{nation, pronoun}}}{\text{total}_{\text{nation}}}$$

The percentage difference between normal and hate speech is calculated with:

$$\text{percentage difference}_{\text{nation, pronoun}} = \left(\frac{\text{average}_{\text{hate, pronoun}} - \text{average}_{\text{normal, pronoun}}}{\text{average}_{\text{normal, pronoun}}} \right) \times 100\%$$

The result is shown in the following table (**Next Page**):

Pronoun	Australia	UK	India	New Zealand	Singapore	South Africa
I	-8.30	-15.91	-32.25	-5.89	-17.53	-37.56
You	5.92	20.50	8.41	20.82	9.75	29.27
He	-7.45	15.29	-10.32	-39.06	30.16	24.30
She	-38.37	-39.76	-30.02	-62.57	-52.10	-59.86
It	-19.66	-20.30	-28.37	-21.78	-25.22	-5.62
We	24.89	14.06	36.93	36.79	79.89	88.11
They	6.97	4.76	12.58	-4.35	22.77	57.09
Me	2.90	3.14	-22.63	-1.56	25.22	-2.05
Him	28.16	-2.45	-6.02	-5.74	0.55	26.86
Her	-20.06	-30.91	-34.71	-17.19	15.78	-58.47
Us	11.36	27.43	37.82	46.72	-5.15	29.12
Them	23.35	44.81	51.55	46.93	97.64	-30.90
My	-14.00	-33.90	-28.19	-15.97	-43.42	-73.28
Your	10.21	17.87	16.98	18.85	-27.57	30.67
His	3.87	18.51	-2.70	-19.27	15.61	-24.98
Its	-9.80	-12.65	-9.16	-19.22	12.10	-47.51
Our	26.10	15.48	107.53	66.60	48.81	92.08
Their	8.51	24.22	65.82	-2.70	31.65	-8.73
Mine	-43.87	-29.90	-70.26	-16.18	23.43	-100.00
Yours	-12.85	-65.77	53.43	-100.00	-100.00	909.85
Hers	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
Ours	-17.89	-49.78	7.80	-100.00	-100.00	-100.00
Theirs	-29.01	156.49	271.18	101.39	1059.61	-100.00
Myself	-17.71	-35.78	-42.55	9.75	-100.00	19.92
Yourself	25.50	-6.74	-49.67	12.38	-5.44	248.86
Himself	30.22	-37.55	0.88	27.39	-100.00	-100.00
Herself	-12.30	-35.57	-8.37	-100.00	-100.00	1270.51
Itself	-38.13	-52.91	45.29	-68.32	27.73	108.56
Ourselves	49.59	4.41	53.26	24.03	-100.00	-100.00
Yourselves	365.03	471.64	322.93	-100.00	1998.35	-100.00
Themselves	53.17	-3.82	96.80	-48.63	204.60	108.56
This	2.11	-4.20	-0.05	-12.22	41.12	57.31
That	-18.20	-10.91	-12.78	-9.63	-5.14	-18.47
These	107.65	110.31	167.98	67.27	186.00	119.91
Those	44.42	46.89	66.97	115.83	38.51	-16.82
Who	44.83	47.64	53.65	31.11	24.88	85.13
Whom	-29.60	-25.16	152.30	52.94	-100.00	-100.00
Whose	29.45	-73.02	-67.54	-8.79	-100.00	-100.00
Which	-24.83	-39.75	-33.96	-25.69	-4.95	-43.35
What	-6.45	12.31	7.99	9.90	5.98	58.63
Anyone	-10.69	15.63	32.28	-61.15	-100.00	74.43
Anybody	-22.35	-15.07	14.15	224.15	-100.00	-100.00
Anything	-35.13	-21.61	4.34	-23.11	-40.29	-12.79
Everyone	-4.18	12.55	-19.06	-13.78	-26.06	99.87
Everybody	-11.30	-32.27	106.02	43.75	188.01	-100.00
Everything	-17.94	-38.84	-40.86	-76.18	-50.04	47.03

Table 5.5: Percentage Difference In Pronoun Usage Between Normal And Hate Speech Comments By Nations, Part 1

Pronoun	Australia	UK	India	New Zealand	Singapore	South Africa
Someone	16.44	16.89	-32.61	12.47	-100.00	-61.55
Somebody	-33.93	135.68	91.74	91.29	-100.00	-100.00
Something	-8.92	-16.97	-8.08	-18.17	-48.60	-0.46
Nobody	-9.61	-31.33	-24.97	-37.12	51.95	225.21
Nothing	23.67	15.78	8.40	6.38	238.96	243.65
Each	19.17	-17.82	1.96	-59.16	-44.01	-100.00
Either	-28.24	-20.08	-15.53	-32.79	133.64	-33.15
Neither	33.61	-31.15	-70.81	36.97	-100.00	-100.00
One	-1.53	-2.30	-4.47	-10.67	-27.97	-79.04
Another	23.20	8.79	28.37	-15.95	4.50	-46.40
Other	13.17	-4.10	-22.41	7.56	-15.22	-68.70
Others	-38.07	28.04	-4.40	-24.83	-45.12	-100.00
Such	-31.11	22.57	9.40	11.03	76.97	-53.43
All	19.40	24.76	26.17	43.91	-0.81	62.39
Any	-8.42	-13.82	-24.89	17.26	-24.97	-100.00
More	-13.75	-25.06	-8.92	-16.06	-56.87	-50.75
Most	10.68	-20.31	-29.03	-17.50	35.38	-59.35
Some	54.23	18.88	18.65	27.68	-22.56	-35.03
Few	10.15	1.31	-26.57	-12.39	-25.94	38.04
Many	-23.89	-19.67	2.80	10.20	43.65	-74.59
Several	-46.69	3.29	-73.20	-100.00	-100.00	-100.00

Table 5.6: Percentage Difference In Pronoun Usage Between Normal And Hate Speech Comments By Nations, Part 2

By analyzing the data, we see that the use of plural and collective pronouns is significantly increased. The use of the second personal pronoun *You, Your* are increased in most nations and *They* increased in all nations except New Zealand. The personal pronoun *I* however, is decreased across all nations, with India (-32.25%) and South Africa (-37.56%) with the biggest changes, suggesting a depersonalization process in the hate speech. Meanwhile, the feminine pronoun *She* is drastically decreased, notably in New Zealand (-62.57%) and South Africa (-59.86%). What is more, demonstrative pronouns *These, Those* show great increases across all nations. The reflexive pronouns like *yourself, himself* increase while *myself* decrease, indicating fewer self-references and more target-references.

We analyzed the percentage change for each nation individually for a better understanding of the pattern:

- **Australia**: A noticeable increase in collective pronouns (*we, our, ourselves, us, them, themselves, yourselves*), and a decrease in individual pronouns (*i, she, her, mine, hers, myself*), suggesting stronger group targeting and less individual targeting of certain types.
- **United Kingdom**: High increase in *you, your, them, their, theirs, those*, specifically 44.81% of increase of "them", 156.49% of "theirs", suggesting stronger group targeting; noticeable decrease in individual pronoun (*i, she, her, mine, hers, myself*) just like Australia, suggesting less individual targeting.
- **India**: High Increase in *we, us, our, ourselves, them, their, themselves, those*, suggesting a trend to mention and make boundaries between different groups of people (specifically our group V.S. their groups, e.g. *this cu*t will also pay for their upbringing?*).
- **New Zealand, Singapore, South Africa**: High increases in *we, us, our, and them*, like India.

Overall, the patterns imply a shift from individual to collective entities in the context of hate speech, meaning a generalization or stereotype against certain groups of people/entities, indicating a stronger "Us vs. Them" narrative.

5.1.2. Results Of Semantic Analysis

5.1.2.1. Lexical Relation Analysis

The average number of each lexical category for a given nation is calculated with the following formula:

$$\text{Average}_{\text{category}} = \frac{\sum_{i=1}^{n} \text{Count}_{\text{category},i,w}}{n}$$

in which

- $\text{Count}_{\text{category},i,w}$ is the count of a particular lexical category (e.g., synonyms, hyponyms) of each word added up for the i-th comment.
- n is the total number of comments of the nation.

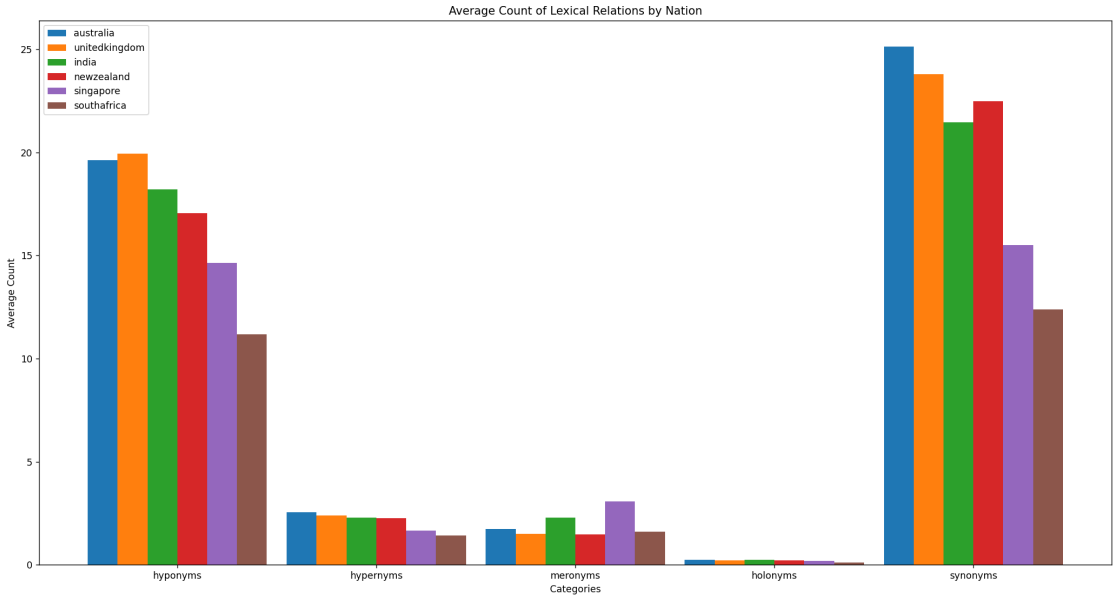


Figure 5.8: Average Count Of Lexical Categories By Nations

An example is: *I like apples*, the synonyms of "*like*" can be "*love, admire, appreciate*", and the synonyms of "*apples*" are *olive*, making the total count of synonyms 4 (I is a stop word removed in the preprocessing of the text). Then we perform the same calculation for each lexical relation, each comment, and each nation. With all of these calculations, we get the result with the following graph and chart:

Category	Australia	UK	India	New Zealand	Singapore	South Africa
Hyponyms	19.62	19.94	18.20	17.04	14.63	11.18
Hypernyms	2.54	2.39	2.28	2.27	1.66	1.44
Meronyms	1.74	1.51	2.29	1.46	3.07	1.61
Holonyms	0.25	0.23	0.23	0.21	0.19	0.12
Synonyms	25.14	23.79	21.46	22.47	15.51	12.38
Total	16,095	8,565	4,875	3,845	1,235	910

Table 5.7: Average count of lexical categories by nation

All countries show relatively high counts of synonyms compared to other categories. Australia has the highest average count at about 25.14, followed by the United Kingdom at 23.79 and New Zealand at 21.46, India at 22.47, while Singapore and South Africa show lower numbers, suggesting a lower semantic diversity. For hypernyms and holonyms which are respectively the more specific terms and the terms encompass parts, we see all nations have similarly low numbers, with Singapore and South Africa constantly lower than the others. However, Singapore has a relatively higher count of meronyms (3.07), indicating a higher frequency to mention terms that are parts of a whole or to use part to refer to the whole instance.

5.1.3. Results Of Syntactic Analysis

5.1.3.1. Average Sentence Length

We calculated the average sentence length (word count) of the average speech and hate speech of all nations.

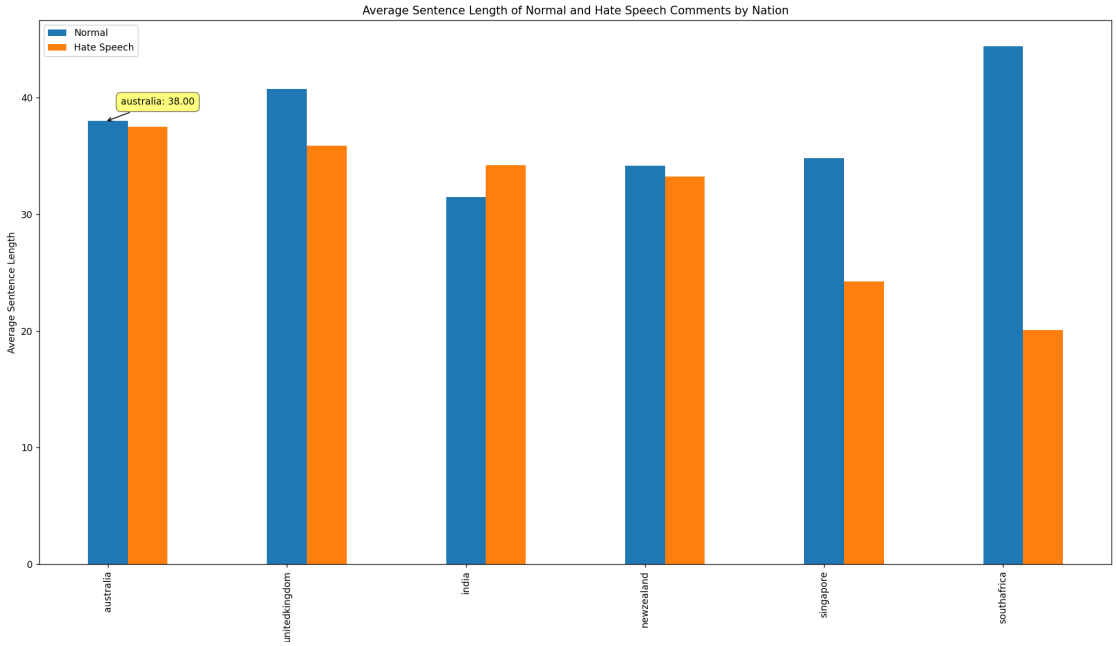


Figure 5.9: Bar Chart of Average Sentence Length of Normal and Hate Speeches

nation	Average Speech		Hate Speech	
	Comment Count	Avg Length	Comment Count	Avg Length
Australia	404,532	38.00	3,219	37.51
United Kingdom	413,886	40.75	1,713	35.88
India	468,872	31.48	975	34.19
New Zealand	149,835	34.15	769	33.22
Singapore	66,801	34.83	247	24.23
South Africa	15,058	44.42	182	20.06

Table 5.8: Comparison of Average Sentence Length by Nation

For all nations except India, the average length of hate speech decreased. Singapore and South Africa are having the largest decrease by −30.43% and −54.84% compared to the average speech, United Kingdom by −11.95%, India by 8.61%, and smaller changes for New Zealand and Australia by −2.72% and −1.29%. These results prove our hypothesis of the syntactic difference, which the sentence lengths of hate speech in Singapore and South Africa tend to be shorter than those of the others.

However, we are not rushing to the conclusion. We computed the variance of the sentence length, and found more interesting results:

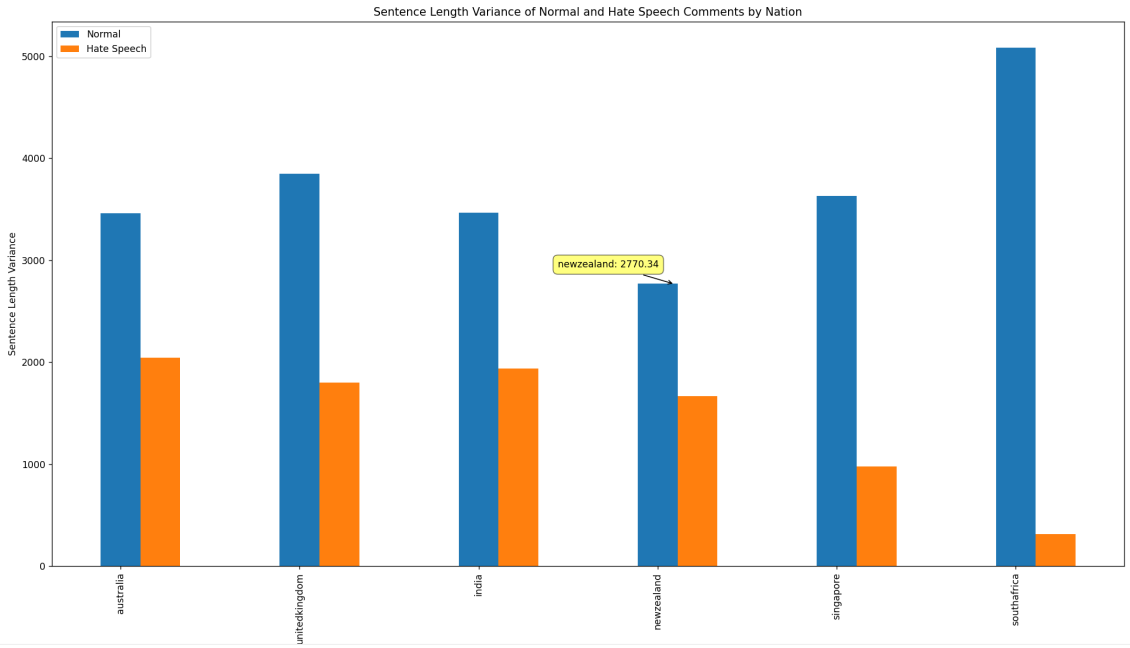


Figure 5.10: Sentence Length Variance of Normal and Hate Speech Comments by Nation

Nation	Variance Of Average Speech Length	Variance Of Hate Speech Length
Australia	3460.53	2043.51
United Kingdom	3848.40	1797.11
India	3466.93	1936.79
New Zealand	2770.34	1668.75
Singapore	3631.53	976.85
South Africa	5084.93	315.90

Table 5.9: Variance Of Sentence Lengths Of Average And Hate Speech By Nations

For all nations, the variance of hate speech sentence length is significantly lower than the variance of average speech sentence, which suggests a more focused and repetitive language used in the hate speech of all nations (e.g. similar expression results in similar sentence length). Also, the languages used by Singapore and South Africa are much longer in the average speech than in their hate speech, meaning a more focused and direct use of the language when making hate speech than the other four nations.

Combining the two analyses, we came up with the following conclusions:

- Australia, United Kingdom, and New Zealand: They both show a reduction in the variance of hate speech sentence length, while still having a similar sentence length compared with average speech, suggesting a tightened use of language without significantly shortening the length.

- India: Increase in the average length of hate speech, decrease in the variance of hate speech length, suggesting a more consistent and slightly more extended dialogue in hate speech.

- Singapore: Drastic decrease in both average sentence length and variance in hate speech, suggesting a contrast: much shorter yet far more consistent.
- South Africa: Similar to Singapore, with even higher contrast.

5.1.3.2. Phrasal Analysis

We analyzed the percentage of each type of phrase in the hate speech.

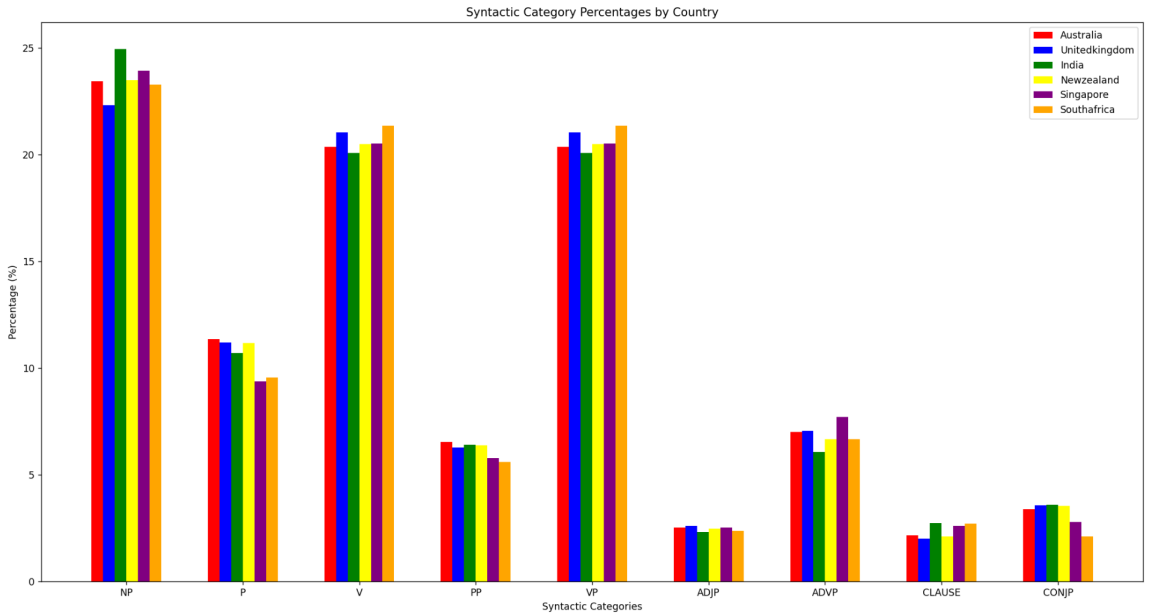


Figure 5.11: Syntactic Category Percentages by Nation

Category	Australia	United Kingdom	India	New Zealand	Singapore	South Africa
NP	23.44	22.31	24.96	23.49	23.93	23.30
P	11.36	11.19	10.71	11.19	9.39	9.56
V	20.38	21.05	20.08	20.50	20.52	21.36
PP	6.53	6.29	6.41	6.40	5.79	5.60
VP	20.38	21.05	20.08	20.50	20.52	21.36
ADJP	2.54	2.61	2.31	2.48	2.53	2.38
ADVP	7.01	7.07	6.06	6.68	7.71	6.66
CLAUSE	2.17	2.01	2.74	2.13	2.61	2.70
CONJP	3.40	3.58	3.61	3.54	2.80	2.10

Table 5.10: Syntactic Category Percentages by Nation

Overall, there is no significant difference in all types of phrases, indicating generally similar phrasal usage. No nation tends to use more of certain phrases than the others do. This proves our hypothesis about the usage of phrases, that the six nations use them similarly despite regional dialect differences.

We then analyzed the transitional probability of phrases in both hate speech and average speech. There are four steps in this process.

- Firstly, for a sentence, tokenization and part-of-speech tagging can be represented as:

$$\text{Tokens}, \text{POS} = \text{Tokenize\_and\_Tag}(s)$$

- The parsing into syntactic categories is then performed on the tagged tokens:

$$\text{Tree} = \text{Parse}(\text{Tokens}, \text{POS})$$

- For each syntactic category C_i in a nation's comments, the transition counts to another category C_j are calculated with the formula:

$$\text{TRANSITION\_COUNTS}[\text{nation}][C_i][C_j] \mathrel{+}= 1$$

  C_j is a child node of C_i in the parse tree.

- The transition probabilities from syntactic category C_i to C_j are calculated by normalization of the transition counts:

$$P(C_j \mid C_i) = \frac{\text{TRANSITION\_COUNTS}[\text{nation}][C_i][C_j]}{\sum_k \text{TRANSITION\_COUNTS}[\text{nation}][C_i][C_k]}$$

  where $\sum_k \text{TRANSITION\_COUNTS}[\text{nation}][C_i][C_k]$ is the sum of all transitions from category C_i.

By the calculation, we got the following transitional probability graph (**Next Page**):

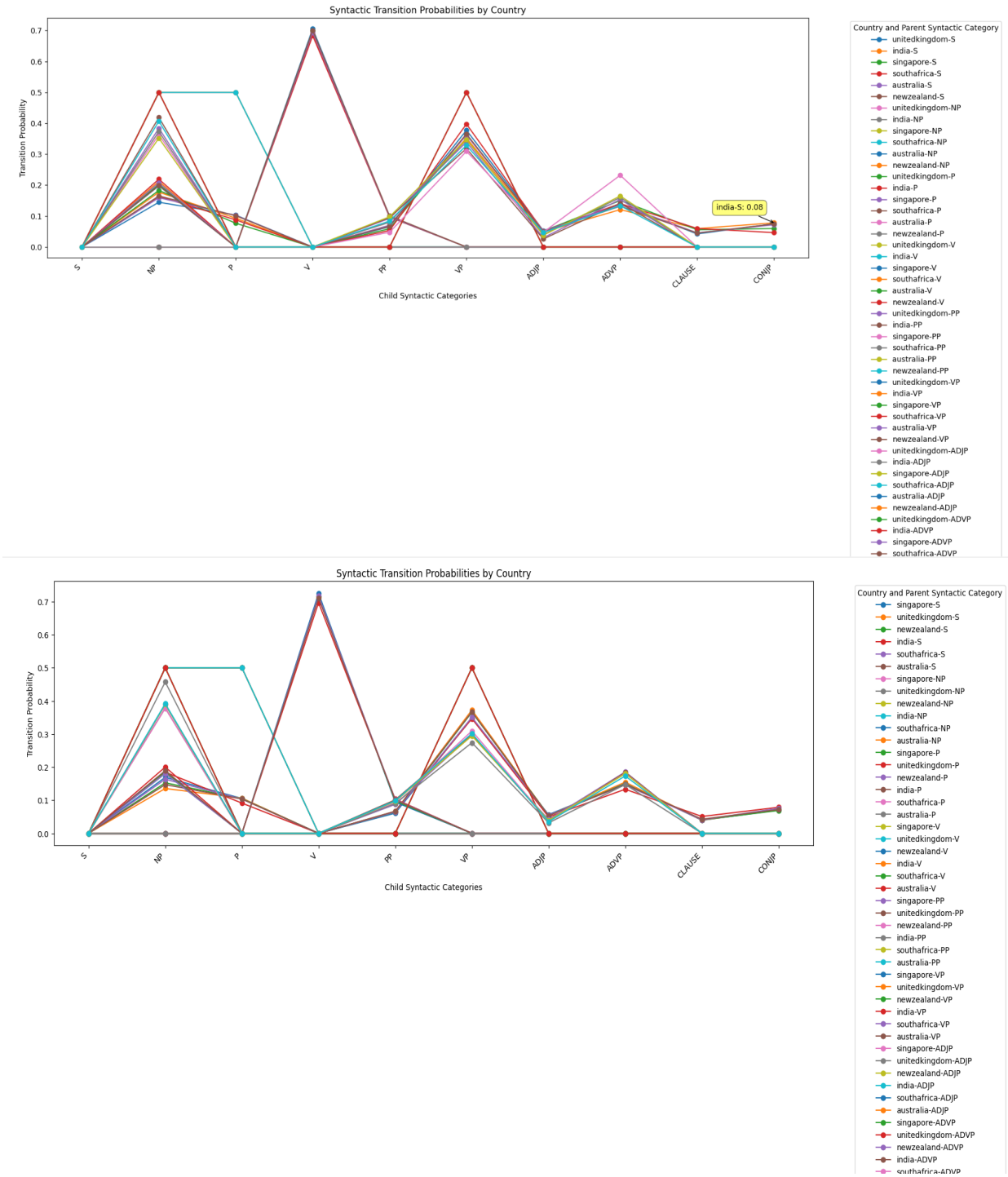


Figure 5.12: Syntactic Transition Probabilities Of Hate Speech By Nations (**Up**), Syntactic Transition Probabilities Of Average Speech By Nations (**Down**)

The most notable change from average speech to hate speech is the increase of transitions to Noun Phrase and Verb Phrase in the hate speech, compared with the average speech. This indicates a stronger emphasis on verbs and targeting entities.

We also observed the following regional differences:

- Disparity in Conjunction Phrase to Noun Phrase transition ($[CONJP] \rightarrow [NP]$):

India:0.4186, New Zealand: 0.4067, United Kingdom: 0.3841, South Africa: 0.3730, Singapore: 0.3630, Australia: 0.3519. There is a 0.0667 gap between the highest and the lowest. We see that India and New Zealand have close numbers and both above 0.40, while the rest are all below. Even though this number is not overwhelmingly large, it suggests a preference to use a compound structure where additional information is included following the conjunctions.

- Huge disparity between Singapore's Conjunction Phrase to Prepositional Phrase ($[CONJP] \rightarrow [PP]$) and others: Singapore: 0.0479, South Africa: 0.0800, New Zealand: 0.0841, United Kingdom: 0.0927, India: 0.0952. This indicates the lesser usage of extensive prepositional phrases.

  However, this is also possibly caused by the insufficient amount of Singaporean comments, which leads to the irregularly small number.

- We observed a very high transitional probability in the United Kingdom's data. There is a **0.5** probability that its Prepositional Phrase[**PP**] will transit to Noun Phrase[**NP**] ($[PP] \rightarrow [NP]$) and its Clause transiting to Verb Phrase[**VP**] ($[CLAUSE] \rightarrow [VP]$). The first result indicates a more complicated sentence structure and more descriptive language because the prepositional phrase is more likely to be followed by a noun, meaning more description for entities; The second result indicates its language being more action-directed, suggesting the clause is often the setup for the action, which are described through verbs later.

5.2. Results Of Machine Learning Model Training

Firstly we examined the results obtained from the random forest classifiers. We used 80% of randomly shuffled data to train and 20% for testing.

Table 5.11: Classification Report Of Random Forest Classifier

Metric	Text-Only		Combined		Weighted	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Precision	0.89	0.95	0.87	0.96	0.81	0.47
Recall	0.99	0.50	1.00	0.37	0.98	0.08
F1-Score	0.94	0.66	0.93	0.54	0.89	0.13

The text-only model shows overall the best performance, with 0.89 precision on non-hate speech and 0.95 on hate speech, yet the recall rate for the hate speech is low, meaning that many hate comments are mistakenly classified as non-hate comments. The combined model showed similar results, where it ended up with a 0.37 recall rate. The weighted model has the worst performance, significantly in detecting hate speech in both precision and recall rates.

In the following table, we have the results from the neuro-symbolic classifier, where Long-Short-Term Memory is used.

Table 5.12: Classification Report Of Neural-Symbolic Classifier

Metric	Text-Only		Combined		Weighted	
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Precision	0.81	0.24	0.84	0.41	0.80	0.21
Recall	0.81	0.23	0.86	0.38	0.65	0.36
F1-Score	0.81	0.24	0.85	0.40	0.72	0.27

The weighted classifier once again underperforms in hate speech classification. However, the text-only classifier performs equally worse. The combined model has the best performance overall while having a low performance in detecting hate speech as well.

Table 5.13: Classification Report Of Decision Forest Classifier

	Text-Only		Combined		Weighted	
Metric	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
Precision	0.81	1.00	0.68	0.56	0.80	0.85
Recall	1.00	0.00	0.94	0.15	1.00	0.01
F1-Score	0.90	0.00	0.79	0.24	0.89	0.02

For the decision forest classifier, one commonality is the extremely low recall rates for all versions of models when classifying hate speech, and its comparatively better performance in classifying nonhate speech.

It is worth noticing that the decision forest classifier does not give a direct prediction of 1s and 0s, instead, it gives a decimal score between 0 and 1 for each entry, and for all versions of the classifiers above, we use 0.5 as the decision boundary, that a score lower than or equal to 0.5 is putting the comment in the non-hate speech category, and a score higher than 0.5 is putting it in the hate speech category.

Reviewing the results above, we observed the following fact: the weighted classifiers that use the linguistic differences are performing worse except in the Decision Forest classifier by a 0.02 difference of F1 score, which makes it only slightly better than the text-only classifier which has a 0.00 F1 score. This is likely caused by the curse of dimensionality, where extra features may not improve the performance of models and even decrease it because of the sparsity of data. There are 239 features in the training dataset of weighted classifiers, and for each comment, most of these features remain 0, as it has only some of the pronouns and part of speech tags. A sparse matrix like this will thus lead to weak statistical inferences of the given data[12]. Also, the hateXplain BERT model we used, since trained on a different dataset, is possibly not accurate when classifying our dataset, causing the issue together with the curse of dimensionality.

What's more, there is a huge performance gap between identifying non-hate speech and hate speech, for all versions of models. The most possible explanation is the imbalanced amount of non-hate speech and hate speech, that the dataset used for the classifier training consists of only 19.7% of comments labeled as hate speech, while the rest are labeled as non-hate speech. In some research, this makes the classifier biased and performs worse when classifying the minority class[43].

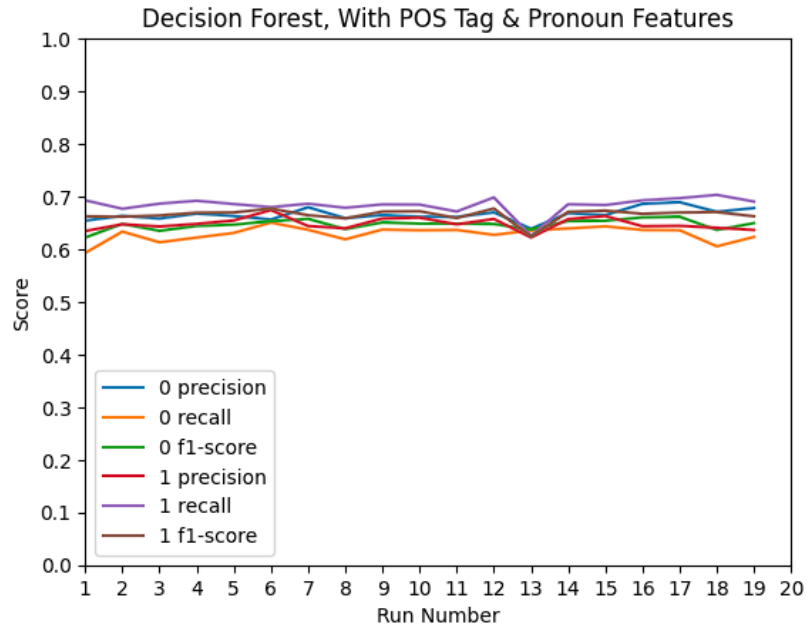
To solve these two problems, we shrank the number of features in the training of weighted classifiers. We did not keep improving the combined model for that it did not utilize the statistics from the first part of the research.

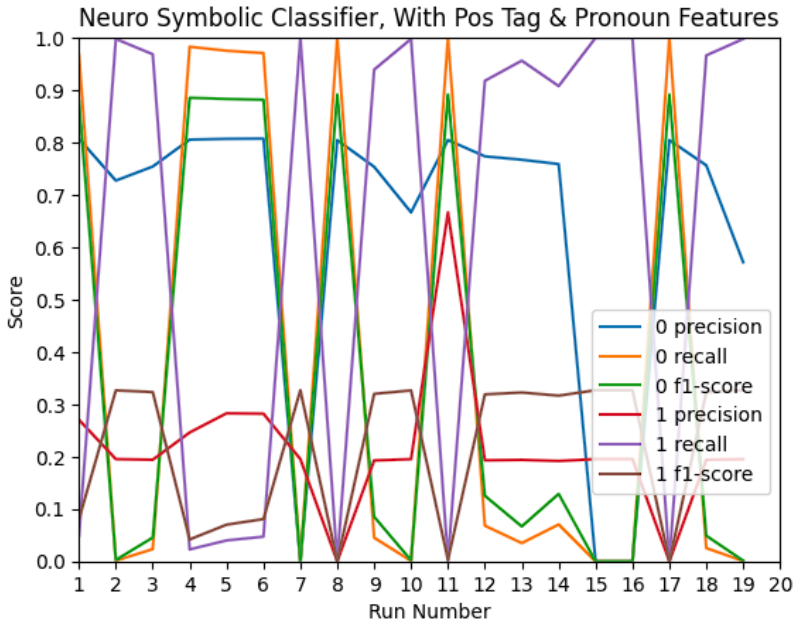
When we shrink the number of features, we must select those that contribute the most to the classification process. In the previous training, all numeric features are used, while in each entry, many features are of value 0. Therefore, we kept those features if the corresponding pronouns or the part of speech tags have enough average occurrences, and the percentage changes from the average speech to normal speech are large enough for the classifier to capture the differences. (e.g. The pronoun "I" consists of around 9-12 % of all pronouns, while from average speech to hate speech, its percentage drops significantly, ranging from -8.30% to -37.56%). Based on this criterion, we picked the following numeric features:

- "NNS_hate","NNS_normal","DT_hate","NN_hate","RB_hate","NN_normal", "RB_normal","IN_hate","DT_normal"
- 'i_hate', 'it_hate', 'these_hate', 'these_normal', 'they_hate', 'they_normal', 'you_hate', 'you_normal', 'them_hate', 'them_normal', 'those_hate', 'those_normal'

These features, along with the nationality of the comment "subreddit" and the comment text "body", are used in the new training session. We intended to confirm the consistency of the performance of these new classifiers. Therefore, instead of charts of values, we will use graphs instead. For each type of classifier, we train 20 times, each time with randomly shuffled data for training and testing. The data, apart from the different numeric features, uses the same nationality and comment text.

We are first presenting the results of the classifiers trained with POS tag and pronoun features.





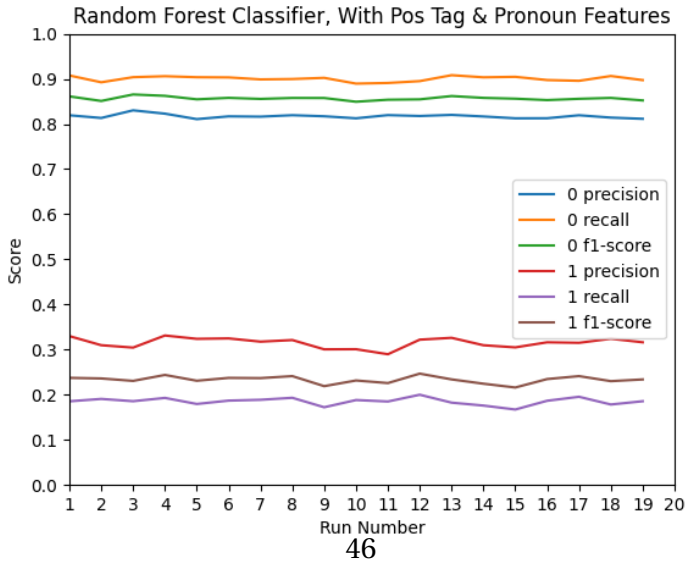


Figure 5.13: The performances of Three Classifiers, Trained With POS Tag Numeric Features & Pronoun Numeric Features

We observed that the decision forest classifier and the random forest classifier delivered comparatively consistent results, where the decision forest classifier carried out resembling results on the classification tasks of both non-hate speech (Class 0) and hate speech (Class 1), while the random forest classifier achieved higher performance on non-hate speech, and much lower on hate speech. Moreover, the neuro-symbolic classifier was unstable, with substantial performance disparity at each round of training. It is worth noticing that the recall rates of positive and negative classes of the neuro-symbolic classifier add up to 1.0. This suggests that the neuro-symbolic classifier (Long-Short-Term Memory in this scenario) is having an error-recall trade-off.

Following these results, we kept deducing the number of features and used the POS tag features exclusively.

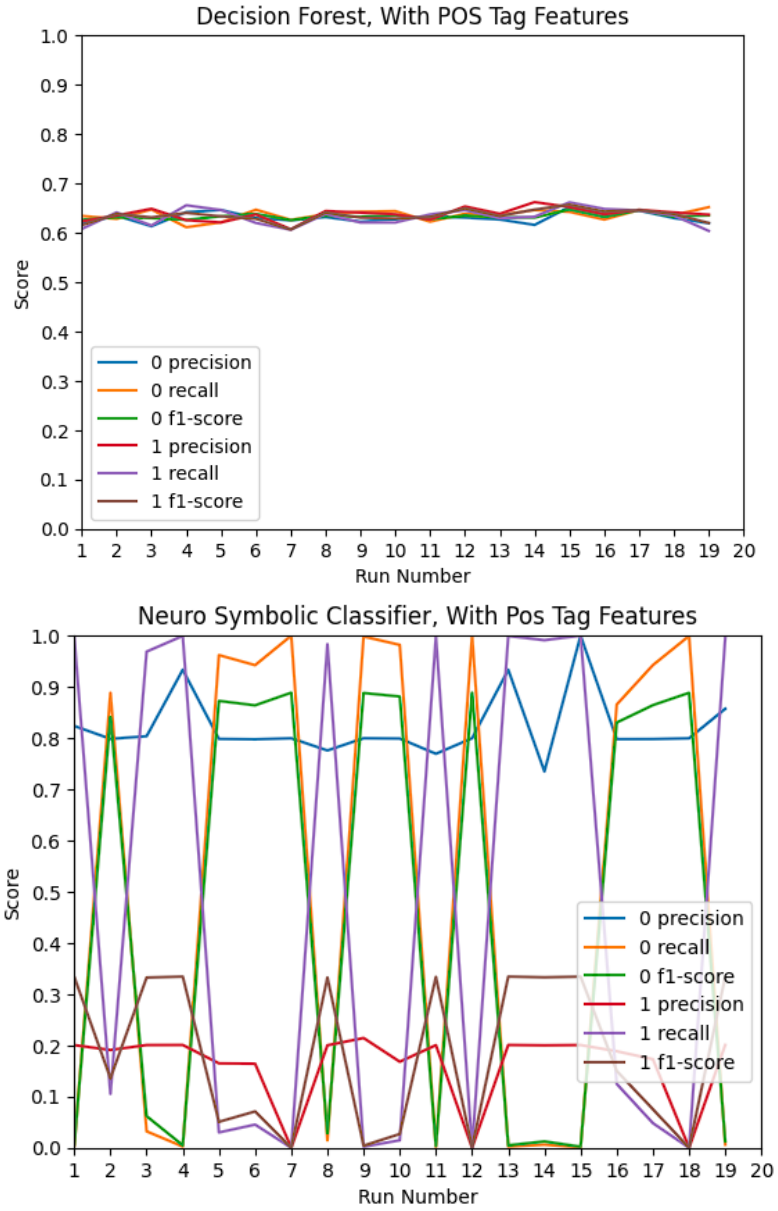


Figure 5.14: Decision Forest Classifier & Neuro Symbolic Classifier, Trained With POS Tag Numeric Features

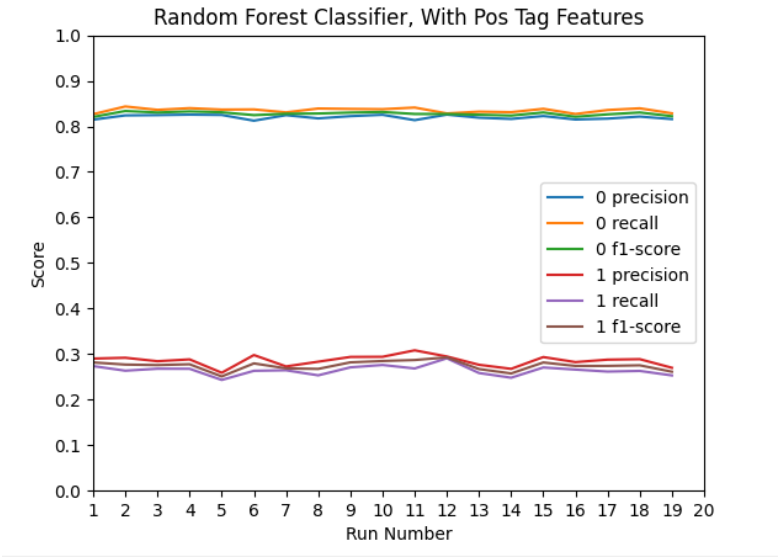


Figure 5.15: Random Forest Classifier, Trained With POS Tag Numeric Features

By reducing the features, the disparity between the precision and recall rates of random forest classifiers and decision tree classifiers is getting smaller. At the same time, the neuro-symbolic classifier is carrying out equally unstable results.

When we only used the pronoun features to train the classifiers, the results were analogous. Although the gap between precision and recall was slightly bigger for the random forest classifier, as can be seen in Fig. 5.16.

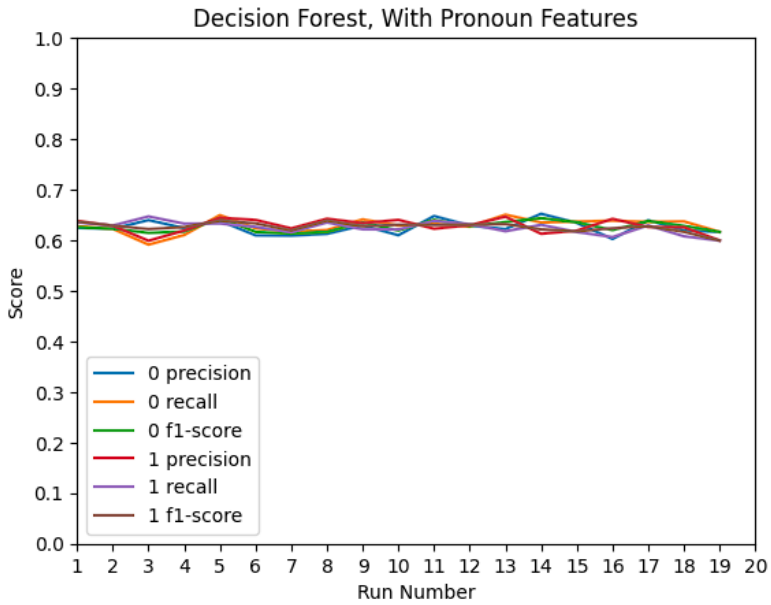


Figure 5.16: Decision Forest Classifier, Trained With Numeric Pronoun Features

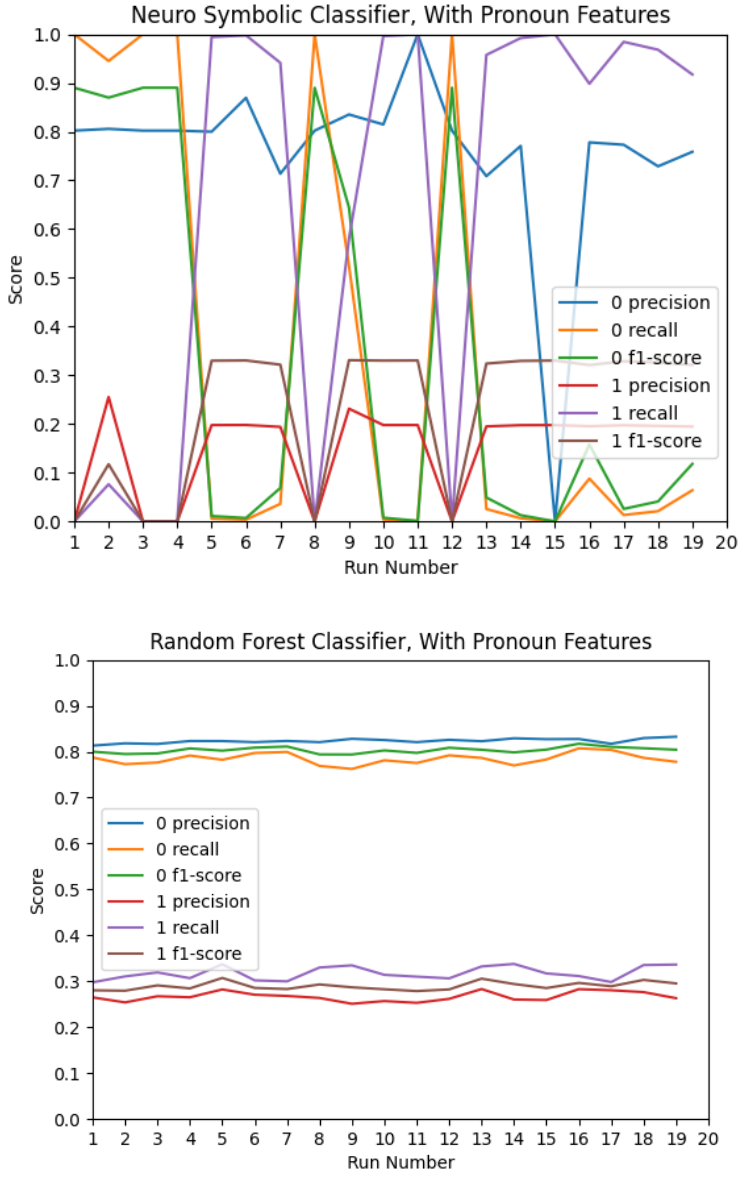


Figure 5.17: Neuro Symbolic Classifier & Random Forest Classifier, Trained With Numeric Pronoun Features

The prevalent instability of the neuro-symbolic classifier is best explained by its sensitivity to class imbalance. However, in each round of classifier training, all of the data was shuffled once, and the ratio between two classes (0 and 1) in the training data was fixed to 23033:5624. Hence the imbalance of classes does not explain the varying results.

To support our argument over the neuro-symbolic classifier, we tested it on its training dataset.

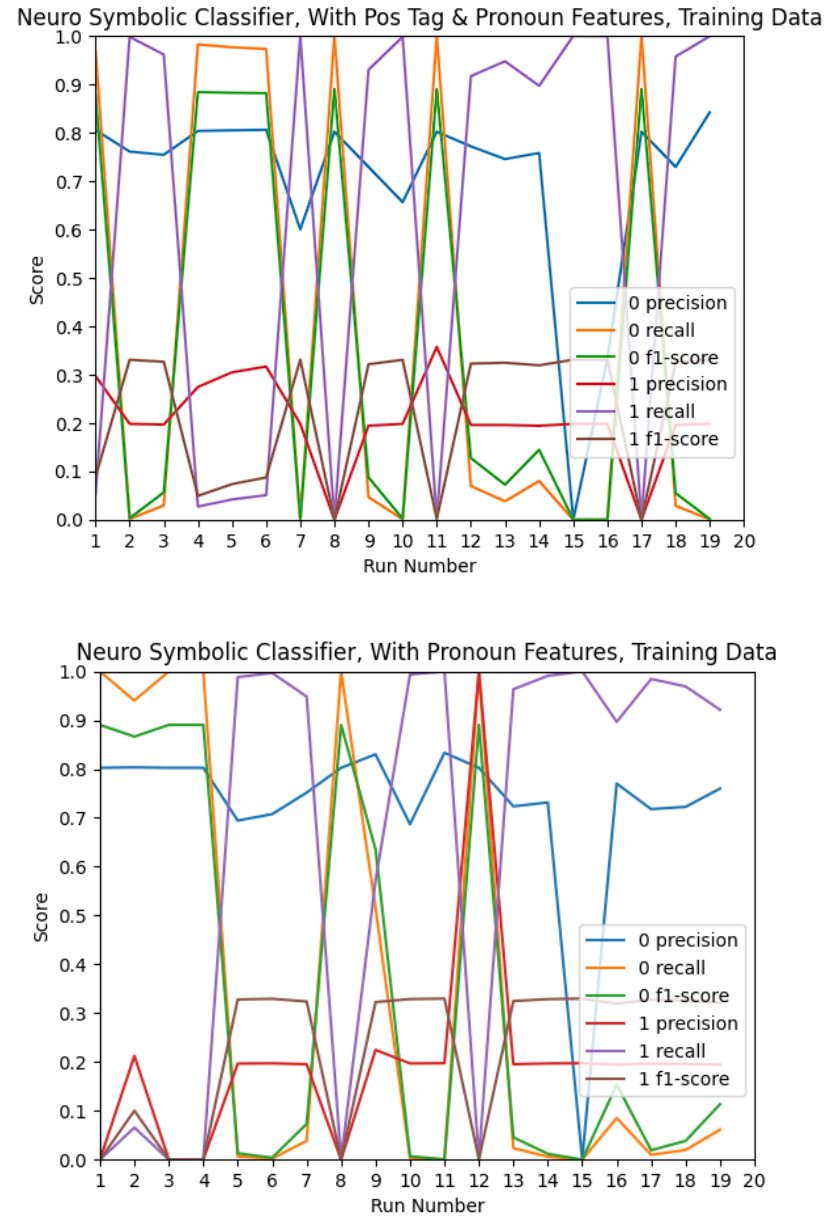


Figure 5.18: Neuro Symbolic Classifier Tested On Training Dataset

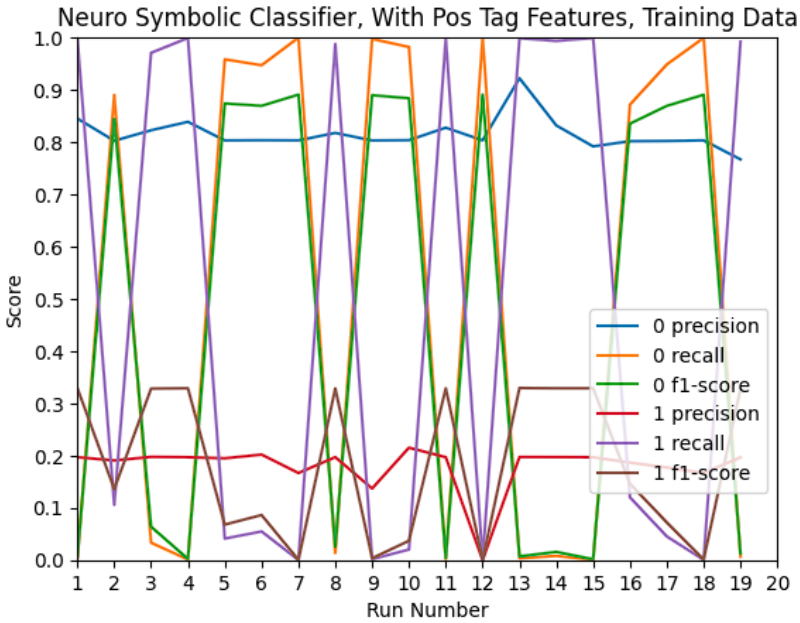


Figure 5.19: Neuro Symbolic Classifier Tested On Training Dataset

When tested on the training dataset, the neuro-symbolic classifier also kept delivering inconsistent results throughout all runs, suggesting that the main cause leading to the poor performance is not overfitting or underfitting on the training data.

We were also interested in whether the extra POS tag and pronoun features could improve the classification ability if compared to the text-only models. The results are as follows (**Next Page**):

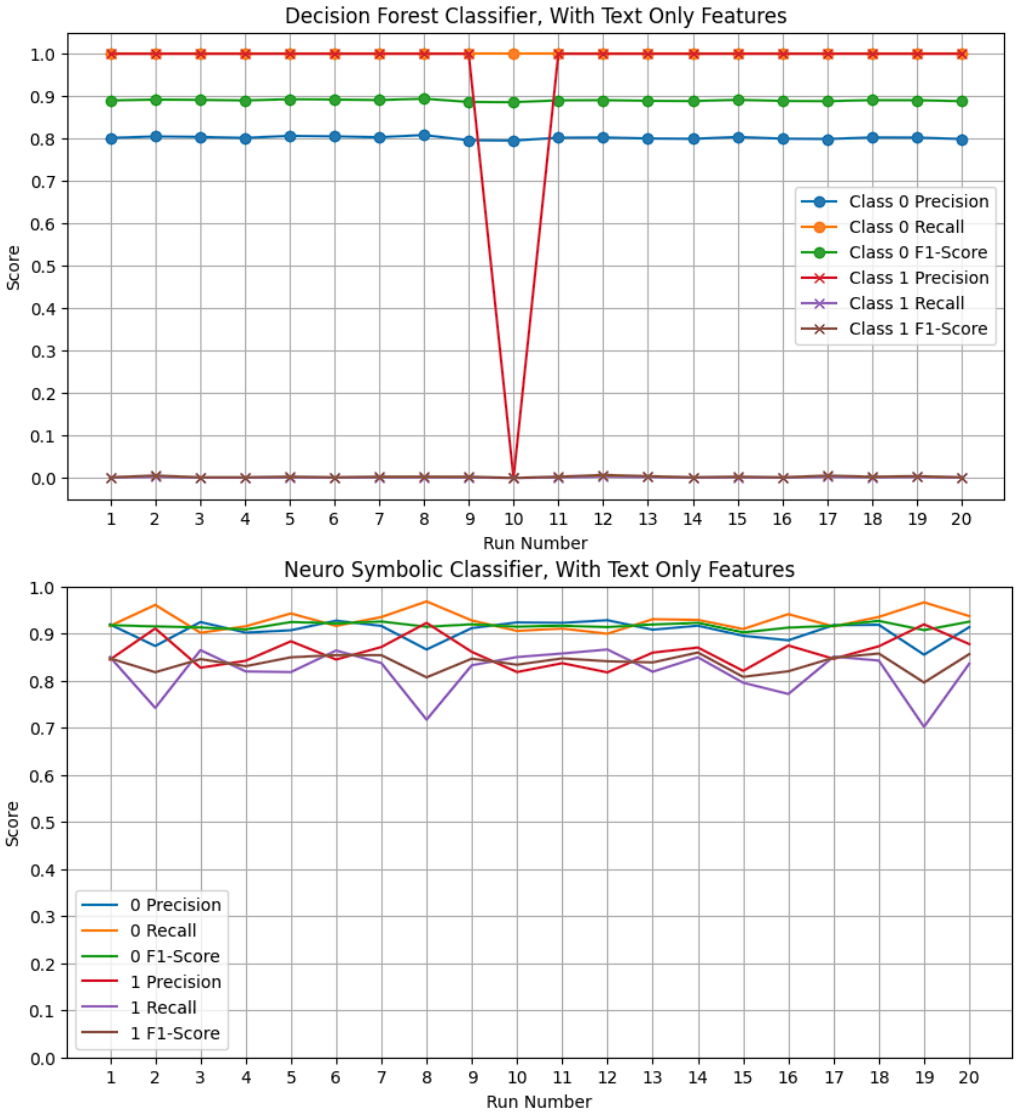


Figure 5.20: Decision Forest Classifier & Neuro Symbolic Classifier Trained With Text Only Features

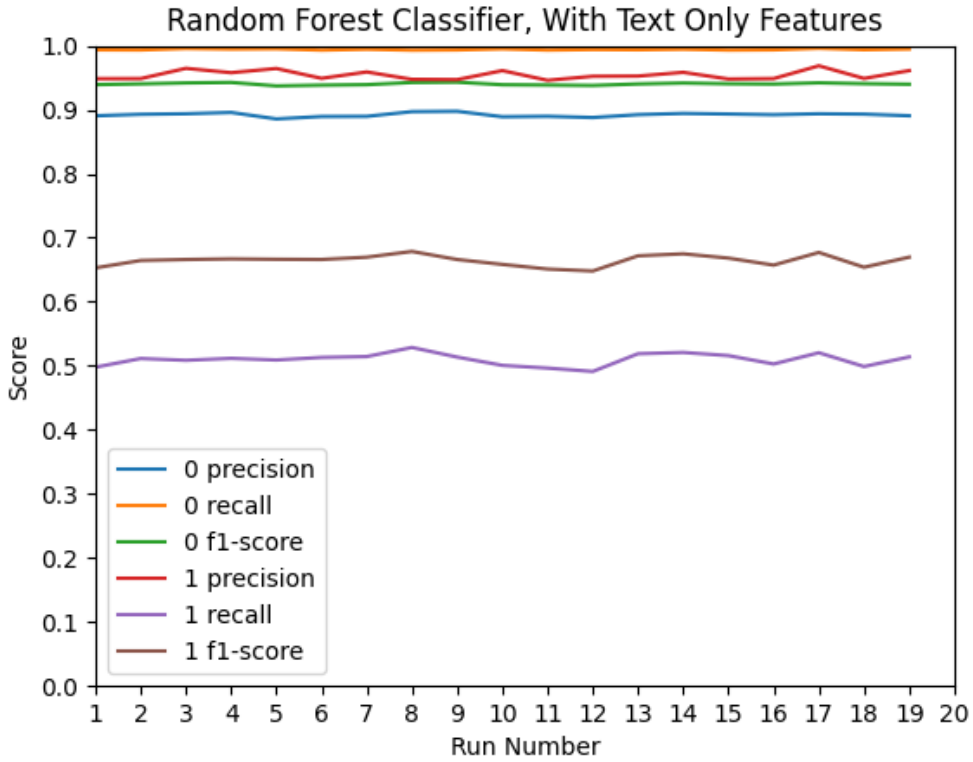


Figure 5.21: Random Forest Classifier Trained With Text Only Features

All classifiers when trained exclusively on comment text performed better when classifying non-hate speech. For the neuro-symbolic classifier, the stability was restored, and it delivered usable results. For the random forest classifier, the performances of classifying both classes increased by a great margin. However, we saw a near 0 value of class 1 recall in the decision forest classifier, suggesting its total incapability of identifying hate speech.

Aggregating these results, we conclude that the extra syntactic and lexical features do not improve the performance of the hate speech classifiers. In our testing scenarios, they brought down the recall and precision rates and yielded unstable and random outcomes. To achieve the highest performance, it is best to train it with the least amount of features. The result also denies the hypothesis that the linguistic features will slightly improve the performance of the machine learning models.

6

Discussion

6.1. Summary Of Main Findings

In this research, we explored the variations of regional English hate speech. We performed lexical, semantic, and syntactic analysis. Through these analyses, we discovered disparities in the usage of pronouns, sentence length, vocabularies, part-of-speech tags, and phrasal transitional probability. We identified the changes from their average speech to hate speech, revealing regional characteristics. These comprehensive analyses provide new insights into the hate speech study.

We used these results to train machine learning classifiers and build an automated hate speech detection system. There are three types of classifiers being trained, each with three versions (text-only, combined, weighted). The linguistic features derived from the first part of the research are used in the training. The test showed varying levels of effectiveness in identifying non-hate and hate speech, dependent on the model types. It revealed that the models trained with complex linguistic features do not always outperform simpler models with less features.

We recognized the limitation in the data scope and the classifier tuning. The experiment suggested that the current integration of prior knowledge is ineffective, calling for a better approach for transforming linguistic statistics into usable features in natural language processing tasks. Moreover, it is necessary to expand the dataset with new data from more sources and reduce the inherent bias of certain online forums.

6.2. Implications

The research provides detailed statistics on the hate speech text that potentially contributes to online moderation, social policy-making, and academia. It provides some useful datasets for further research of regional English. They are beneficial to the studies of computer science, linguistics, law, and sociology. The linguistic feature analyses can help policymakers build thorough regulations while respecting cultural diversities. Finally, the moderation tools can be tailored to the regional specificities, and help avoid abrupt filtering. The tuning approach can be shared across online platforms and user groups.

6.3. Limitations

This research, which encompasses the aim of drawing the panorama of the hate speech differences in regional English variants, is constrained by the insufficient amount of data and the lack of coverage. Chiefly, the dataset is gathered from only one online forum, it contains only tens of thousands of comments, which are far from sufficient to conclude the overall linguistic characteristics.

Meanwhile, the insufficiency of the data leads to inaccurate statistics. A clear example is the occurrence of extreme numbers in the lexical and syntactic analysis results, such as a -100% decrease in the usage of certain parts of speech tags, or an extremely small average occurrence of a certain pronoun. A much larger data set is needed to minimize the randomness.

Moreover, the statistical analysis results in an enormous number of items (Occurrences & Percentages), complicating the selection of useful features. Explaining these numbers is more challenging, as we are not able to tell the meaning of all changes by checking each comment text (e.g., what each change suggests about the topic). Thus we have to interpret these findings in a way that leads to multiple possibilities in the manner that accounts for the uncertainty, or even put aside some of the results as there is no, which takes much more interdisciplinary knowledge such as psychology and sociology than we presently know.

When it comes to classifier training, one of the major limitations is how the linguistic differences from the first part of the research should be transformed into usable features. Currently, we take the differences between average speech and hate speech as weights, expecting that the classifiers learn about them. What is infeasible about this training mode is the creation of a new dataset in which extra features are added to the original dataset, prolonging the time for data preparation, expanding the need for hardware resources, and frequent updating of these weights due to the ever-changing nature of hate speech. Nevertheless, the training results have proved this methodology ineffective, as it significantly lowers performance.

Another major limitation is the number of models selected. There are only three types of classifiers trained, while other cutting-edge models are not used, for example, large language models. Therefore, we are unable to conclude the general effectiveness of our methodology on a wider range. For the three classifiers we have now, the result is heavily restrained by our ability to optimize them. Our works do not focus on modifying the parameters or the design, and their impact on performance remains unknown. For example, models like Long-Short-Term Memory can be tuned more by changing the design of the network layers.

7

Future Work

We list out the future work here for the convenience of future researchers. It is divided into three parts: acquiring larger datasets, improving labeling accuracy, and improving the integration of linguistic analysis and classifier training.

7.1. Acquiring Larger Datasets

As mentioned in the methodology section, there are several more available sources: News websites, YouTube comment sets, and Twitter datasets. Making use of them will expand the scope of the dataset and reduce the bias.

7.2. Improving Labelling Accuracy

Our current labeling is done through an automated process. However, there is no way to verify the accuracy, unless the classifier is trained on the same dataset. In any further research, crowd-sourcing work shall be introduced to manually label the new datasets. This approach, however, will massively increase the cost by thousands of Euros, more or less based on the actual size of the dataset. Researchers shall be considerate of the trade-off between accuracy and their finances.

7.3. Improving The Integration Of Linguistic Analysis And Classifier Training

Based on the prior findings, the most critical task is to explore advanced machine learning architectures for better incorporating the lexical, semantic, and syntactic findings into the models, such as adopting other machine learning models and modifying the model structure. Furthermore, it is beneficial to evaluate the classifier performance with more metrics, such as the accuracy and recall rate over different comment topics to get deeper insight into the capability. Lastly, it is necessary to involve more interdisciplinary knowledge in interpreting linguistic statistics.

Bibliography

[1] English pronoun types. URL https://twp.duke.edu/sites/twp.duke.edu/files/file-attachments/pronouns.original.pdf.

[2] Every publicly available reddit comment. URL https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/.

[3] Twitter api documentation. https://developer.x.com/en/docs/twitter-api,.

[4] Twitter usage statistics, . URL https://www.internetlivestats.com/twitter-statistics/.

[5] Lexical. URL https://www.merriam-webster.com/dictionary/lexical. Accessed: June 25, 2024.

[6] Natural language toolkit. https://www.nltk.org/. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

[7] Syntax. URL https://www.merriam-webster.com/dictionary/syntax. Accessed: June 25, 2024.

[8] Youtube discussions tab dataset (245.3 million comments). https://www.reddit.com/r/DataHoarder/comments/xz0e02/youtube_discussions_tab_dataset_2453_million/, 2022.

[9] Andrew Altman. The Harm in Hate Speech By Jeremy Waldron. *Analysis*, 75(1):177–179, 11 2014. ISSN 0003-2638. doi: 10.1093/analys/anu098. URL https://doi.org/10.1093/analys/anu098.

[10] Felix Ameka. Interjections: The universal yet neglected part of speech. *Journal of Pragmatics*, 18(2):101–118, 1992. ISSN 0378-2166. doi: https://doi.org/10.1016/0378-2166(92)90048-G. URL https://www.sciencedirect.com/science/article/pii/037821669290048G.

[11] Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. Resources for multilingual hate speech detection. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.12. URL https://aclanthology.org/2022.woah-1.12.

[12] Oluseun Omotola Aremu, David Hyland-Wood, and Peter Ross McAree. A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data. *Reliability Engineering System Safety*, 195:106706, 2020. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2019.106706. URL https://www.sciencedirect.com/science/article/pii/S0951832019304752.

[13] Australian Department of Home Affairs. English—our national language. Website, Accessed 2024. URL https://www.homeaffairs.gov.au/about-us/our-portfolios/social-cohesion/english-our-national-language.

[14] Steven Bird, Ewan Klein, and Edward Loper. Categorizing and tagging words, 2023. URL https://www.nltk.org/book/ch05.html. This chapter of the NLTK Book discusses techniques for categorizing and tagging words, introducing the basics of part-of-speech tagging and the use of various NLP tools within Python.

[15] Burton-Roberts and N. *Analysing Sentences: An Introduction to English Syntax*. Learning about language. Longman, 2011. ISBN 9781408233740. URL https://books.google.nl/books?id=DBdrRQAACAAJ.

[16] Cambridge Dictionary. Conjunctions. URL https://dictionary.cambridge.org/grammar/british-grammar/conjunctions.

[17] Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352, 2022. ISSN 1866-9964. doi: 10.1007/s12559-021-09862-5. URL https://doi.org/10.1007/s12559-021-09862-5.

[18] Richard Delgado. Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard Civil Rights-Civil Liberties Law Review*, 17:133, 1982. Seattle University School of Law Research Paper, 51 Pages Posted: 8 Feb 2012.

[19] Neha Deshpande, Nicholas Farris, and Vidhur Kumar. Highly generalizable models for multilingual hate speech detection, 2022. URL https://arxiv.org/abs/2201.11294.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[21] Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12), 2020. ISSN 2076-3417. doi: 10.3390/app10124180. URL https://www.mdpi.com/2076-3417/10/12/4180.

[22] Simona Frenda, Bilal Ghanem, Manuel Montes y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.*, 36:4743–4752, 2019. URL https://doi.org/10.3233/JIFS-179023.

[23] Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. EPIC: Multi-perspective annotation of a corpus of irony. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.774. URL https://aclanthology.org/2023.acl-long.774.

[24] Lei Gao, Alexis Kuppersmith, and Ruihong Huang. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach, 2018. URL https://arxiv.org/abs/1710.07394.

[25] Michael Greenacre. Measures of distance between samples: Euclidean. URL https://econ.upf.edu/~michael/stanford/maeb4.pdf.

[26] A. F. Gupta and K. (ed.) Fischer. Epistemic modalities and the discourse particles of singapore. In *Approaches to Discourse Particles*, pages 244–263. Elsevier, Amsterdam, 2011. URL URLoftheoriginaldocument. Archived from the original (DOC) on 5 February 2011. Retrieved 2 February 2011.

[27] Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis, 2024.

[28] Rijul Magu, Kshitij Joshi, and Jiebo Luo. Detecting the hate code on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):608–611, May 2017. doi: 10.1609/icwsm.v11i1.14921. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14921.

[29] Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Henriques. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66, 2018. doi: 10.1109/BRACIS.2018.00019.

[30] Michele Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8:e859, 2022. ISSN 2376-5992. doi: 10.7717/peerj-cs.859. URL https://doi.org/10.7717/peerj-cs.859.

[31] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. 2021.

[32] Merriam-Webster Dictionary. URL https://www.merriam-webster.com/dictionary/lexicon. Definition 1: A book containing an alphabetical arrangement of the words in a language and their definitions, e.g., a French lexicon. Definition 2a: The vocabulary of a language, an individual speaker or group of speakers, or a subject. Definition 2b: The total stock of morphemes in a language. Definition 3: Repertoire, inventory.

[33] Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. Emojis as anchors to detect arabic offensive language and hate speech, 2022. URL https://arxiv.org/abs/2201.06723.

[34] Francimaria R. S. Nascimento, George D. C. Cavalcanti, and Márjory Da Costa-Abreu. Exploring automatic hate speech detection on social media: A focus on content-based analysis. *Sage Open*, 13(2):21582440231181311, 2023. doi: 10.1177/21582440231181311. URL https://doi.org/10.1177/21582440231181311.

[35] Cecil L. Nelson, Zoya G. Proshina, and Daniel R. Davis, editors. *The Handbook of World Englishes*. John Wiley Sons, 2019. ISBN 9781119164210. doi: 10.1002/9781119147282. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119147282.

[36] James A. Piazza. Politician hate speech and domestic terrorism. *International Interactions*, 46(3):431–453, 2020. doi: 10.1080/03050629.2020.1739033. URL https://doi.org/10.1080/03050629.2020.1739033.

[37] Rajbhasha. Constitutional provisions - official language related part-17 of the constitution of india. Website, Accessed 2024. URL https://rajbhasha.gov.in/en/constitutional-provisions.

[38] Nick Riemer. *Introducing Semantics*. Cambridge Introductions to Language and Linguistics. Cambridge University Press, 2010.

[39] Statista. Most common languages on the internet. Website, January 2024. URL https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/. Release date: January 2024. Region: Worldwide. Survey time period: January 2024. Special properties: Share of websites featuring content in each language, compared with the share of the global population that speaks each language. Supplementary notes: Languages include sub-languages (e.g., "Chinese" includes Mandarin, Yue, etc.). Figures are based on a traffic analysis for the top ten million websites, as ranked by Alexa Internet.

[40] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445092. URL https://doi.org/10.1145/3411764.3445092.

[41] Robert D. van Valin, Jr. *An Introduction to Syntax*. Cambridge University Press, 2001.

[42] Jinghua Xu and Zarah Weiss. How much hate with china? a preliminary analysis on china-related hateful tweets two years after the covid pandemic began, 2022.

[43] Wanwan Zheng and Mingzhe Jin. The effects of class imbalance and training data size on classifier learning: An empirical study. *SN Comput. Sci.*, 1(2), feb 2020. doi: 10.1007/s42979-020-0074-0. URL https://doi.org/10.1007/s42979-020-0074-0.

[44] Miguel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. Authorship and aggressiveness analysis in mexican spanish tweets. *Language Technologies Lab., Computational Sciences Department, Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Mexico*, 2020.

- END -