ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

AN AUTOMATIC HATE SPEECH DETECTION IN SOCIAL MEDIA THROUGH COMPUTATIONAL LINGUISTICS: INFIDELITY VIDEOS IN FOCUS

Klein Mamayabay[1], Danilo G. Baradillo[2]
[1]PhD, Teacher Education Programs, St. Mary's College of Tagum, Inc., Tagum City, Philippines
[2]PhD, Program Chair, University of the Immaculate Conception, Davao City, Philippines

Article DOI: https://doi.org/10.36713/epra15377
DOI No: 10.36713/epra15377

ABSTRACT

The escalating prevalence of hate speeches, amplified by the misguided use of social media, introduces alarming challenges to the safeguarding of human rights and individual welfare. Motivated by this, the study explored the detection and classification of hate speech, specifically as observed in speeches and comments related to infidelity videos on YouTube Channel of Raffy Tulfo in Action. Further, the study utilized a computational linguistic algorithm through Long Short-Term Memory (LSTM). Additionally, the study sought to understand the distinctions in linguistic features between hate speech and non-hateful speech through LSTM. The researcher used 9,600,586 tokens for the analysis. To answer the first research question, the employment of LSTM helped identify hate speeches from non-hate speeches through effective data gathering through YouTube Application Programming Interface (API) and Whisper AI, text processing, labeling, coding, and algorithm deployment. Through that process, LSTM also classified them per target, including sex, quality, physical attributes, disability, religion, race, and class. Further, to answer the second research question, the study was able to identify 35 lexicons. Some samples include peenoise, U10, kokey, taitok, quibolok, squami, and shut@, which were used negatively. Lastly, to answer the last question, tokenization, embedding, sequential dependencies, padding, training-testing, and evaluating helped LSTM assess hate speech linguistic features. It is evident in the confusion matrix showing 46% true positives and 49% true negatives and its evaluation performance of 95% F1 score, affirming its high robustness and reliability.

KEYWORDS: *Applied linguistics, language, hates speeches, infidelity cases, computational linguistics, Long Short Term-Memory (LSTM), Philippines*

INTRODUCTION

Hate speech is a malicious expression that uses offensive language directed to a person or group of people based on the characteristics they are representing in areas including gender, relationships, politics, ethnicity, race, beliefs, etc. [1] and, sadly, it is now on the rise with the advent of social media [2], [3], [4], [5], [6]. In addition, the United Nations (UN) emphasized the dangers of hate speech to human rights and life [7], especially on the increasing cases of infidelity where studies found that people's comments and reactions toward their partners' cheating behavior could go up from verbal assaults to killing their unfaithful partner, thus, creating a very alarming human behavior [8] and [9].

In a study conducted in Germany, researchers found out that more than half of the participants indicated that they were more likely to commit cheating on their partners. Further, in the same country, another study found that 77.7% of the participants indicated that they had caught or suspected that their current or previous mates had been unfaithful, and during the data analysis, the results showed that some of their immediate comments and reactions about the issue were through violence, humiliating their partners, terminating their relationship in a harsh manner, using of psychological abuse, and hateful words and statements against their partners [10], [11].

Despite the relevance of this existing literature, there is only a little research utilizing computational linguistics to analyze and detect hate speech concerning cases of infidelity in the uploaded videos and comments on social media, particularly in the context of Philippines [12].

Additionally, with the advocacy of the United Nations in combatting hate speech, which can expose those targeted to discrimination, abuse, and violence, they heighten the necessity and priority in monitoring and analyzing hate speech through research. Motivated by these gaps, the researcher recognized the urgency to undertake this study in order to contribute to the prevention of any potential detrimental effects it may pose to our society [7]

This study aims to generate hate-contained datasets for application developers to combat online hate speech. The results can be used by educators and students in discussions and assessments. Additionally, infographics will be created to raise awareness of hate speech prevalence on social media. The study's algorithms will contribute to fostering an inclusive online community. Findings will be disseminated through

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

publications, lectures, and participation in research forums, seminars, and conferences at local, national, and international levels.

Purpose of the Study

The purpose of this quantitative corpus-driven study was to find out how computational linguistics techniques contribute to the identification, detection, prediction, classification, and analysis of hate speeches found in social media in relation to infidelity cases here in the Philippines. This study also utilized computational linguistic algorithms to distinguish, identify, and analyze the linguistic features of hate speeches from non-hateful speeches.

Research Questions

The following research questions were sought:

1. How is hate speech classified as found in speeches and comments relating to the infidelity videos through computational linguistics?
2. What are the linguistic features of hate speeches as identified through computational linguistics?
3. How do the linguistic features of hate speech differ from those of non-hate speech through the use of Long Short-Term Memory (LSTM)?

METHODOLOGY

This study is quantitative in nature, employing a corpus-driven approach. Quantitative research design is a research method that involves collecting and analyzing numerical data to answer research questions or test hypotheses and mentioned that this approach is characterized by its use of statistical methods and large sample sizes to identify patterns and relationships in the data [13].

In particular, quantitative corpus-driven analysis is a research approach that involves the use of large collections of texts or corpora and computational methods to analyze and interpret patterns and relationships in language use. This approach is quantitative in nature because it involves the use of statistical tools and techniques to identify patterns and relationships in the data. The corpus-based approach to language analysis is based on the idea that language use can be analyzed and understood by examining large collections of naturally occurring language data. By analyzing these collections of texts, researchers noted that patterns and relationships in language use that might not be apparent from individual texts could be identified [14], [15].

Further, one of the specific approaches of a quantitative corpus-driven analysis is computational linguistics. This approach seeks to scrutinize extensive text datasets to discern language usage patterns and trends. This technique is well-suited for linguistic data because of its capacity to examine linguistic features, employ natural language processing methods, make use of machine learning models, and efficiently manage substantial data volumes [16].

In this study, the researcher found this method useful in shaping the results toward detection, identification, prediction, and classification of hate speeches found in the infidelity video comments and video transcripts from the infidelity cases found in the Youtube Channel of Raffy Tulfo.

Research Material

The research materials for this study were the transcripts from the selected infidelity videos from various social media platforms, including the comments written by the netizens. Moreover, Devopedia, known for natural language processing methodologies, suggested that a good text corpus should be at least half a million words or 500,000. It is to ensure that low-frequency words are also adequately represented. Specifically, to get better results, the researcher utilized 9,600,586 tokens for the data analysis and scraped them out from the transcribed audio-visual files and the extracted comments on the infidelity videos from YouTube [17].

Furthermore, the selected materials were chosen through the following inclusion criteria: their popularity, infidelity-related topics, and the varied hate or aggressive speeches present. In the first criterion, the term media popularity refers to the number of engagements, which includes the reactions, comments, and number of views. The second criterion is about the YouTube videos that center on infidelity topics from Raffy Tulfo in Action YouTube Channel. The last criterion is the presence of varied hate speeches among the comments of the netizens on the video. These speeches are the main concern of this study and were used to analyze the selected deep-learning techniques under computational linguistics.

Data Analysis

To aid me in answering the research questions in my study, relevant steps were performed in conducting this quantitative corpus-based research.

To answer the research questions from the previous chapter, frameworks under the computational linguistics concept [18], the concepts of hate speech identification [19], and the concept of hate speech targets [20].

Specifically, the steps below helped answer the first research question. In order to classify the hate speeches found among the infidelity videos through computational linguistics, the researcher utilized a deep learning model through Long Short-Term Memory (LSTM) using Phyton. Phyton is commonly used for developing websites and software, task automation, data analysis, and data visualization. On the other hand, LSTM is a special type of neural network which is designed to work with Phyton. It has sequences of a data set, and a long-term dependency exists. LSTM is useful when one needs a network to remember information for a longer period. This feature makes LSTM suitable for processing textual data.

Further, LSTM is a collection of similar cells, where each cell processes the input in a specific approach. Using the forget gate, information to be forgotten is identified from a prior time step. It has an input gate and tanh, where new information is sought to update the cell state. The information from the two gates below is used to update the cell state. Lastly, the output gate and the squashing operation provide useful information. This arrangement of cells facilitates LSTM to remember earlier information for a longer time [21].

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

To proceed with the classification of hate speeches, the researcher collected a dataset from YouTube, which is an open-source platform. These are the 72 videos from Raffy Tulfo in Action YouTube page where the researcher scraped the comments using YouTube Application Programming Interface (API) while the video transcriptions were taken using Whisper AI implemented through Google Colaboratory – Whisper is an automatic speech recognition system by Open AI that allows an efficient transcription of audios from videos [22]; meanwhile, Google Colaboratory is a cloud-based platform provided by Google that allows users to write and execute Python code in a collaborative environment.

The researcher combined the gathered YouTube comments and video transcriptions to serve as the main dataset to be used during the deep learning analysis in order to attain appropriate results. From these data sources, the researcher was able to gather 9,600,586 tokens prior to data pre-processing.

During the pre-processing, the researcher performed lowercasing of all tokens and then proceeded to remove the following: a.) stopwords in English, Tagalog, and Bisaya languages, as well as the removal of special names. Stopwords provide no meaningful information, especially if we are building a text classification model. Therefore, the researcher removed stopwords from the dataset. The researcher also removed punctuations and transformed multi-spaced words into single spaces, including the uniform resource locator (URLs), emoticon Unicode, dates, and other special characters.

Using Phyton, the researcher also proceeded to tokenization using the Natural Language Toolkit (NLTK) library. Through this process, whitespaces were added before and/or after preserved special character strings (wherever necessary) in order for the tokenizer to recognize them as individual tokens. This allows for the analysis of linguistic features at the word level, such as identifying the presence of specific words or patterns associated with hate speech. This also includes word embedding, padding, and sequential dependencies to capture linguistic relationships. These processes play a role in enhancing LSTM's ability to understand, generalize, and make predictions based on sequential input data.

To identify comments as hate speeches from non-hate speeches, the concept of hate speech identification by Waseem and Hovy is used [19]. Specifically, they conceptualized rules and indicators on how a certain remark was to be categorized under hate and non-hate speeches. The researcher also utilized Silva et al. classification schemes of hate speeches to identify which targets these hate speeches belong [20].

To answer the second research question, the researcher still used the computational linguistics concept [18] through the help of Long Short-Term Memory (LSTM). Here, the researcher obtained the unique lexicons related to the targets of hate speeches and also provided its other spelling variants.

Lastly, to answer the third research question, the same deep learning model was used in order to identify and analyze the linguistic features of hate speeches and compare them with the features of non-hate speeches. It is done by allowing the dataset under processes such as tokenizing, sequencing, embedding, and padding. Further, it is also done by getting the performance evaluation of LSTM, such as the recall, precision, accuracy, and F1 score. It also included the confusion matrix of the hate and non-hate features.

Further, I noted that before utilizing the selected computational linguistic model, I first checked the data set since it might become highly imbalanced. This needs to be considered in order to avoid biased results. Next, the researcher divided the data set into training and testing. I divided the modeling dataset into training and testing sets by assigning two-thirds of the data points to the training set and one-third to the testing set, or a 70:30 ratio – a process needed to avoid overfitting. Consequently, I trained the model on the training set before applying it to the test set. This allows us to assess the efficacy of our model [21].

The following performance evaluation criteria were obtained, which include the values of accuracy, precision, recall, and F1 score. First, accuracy is the ratio of the total number of entries correctly classified to the total number of observations. For a balanced dataset, accuracy is the metric by which algorithm performance can be compared. Second, precision. It is the proportion of total positive entries that correspond to entries that have been correctly forecasted as positive. Moreover, a greater value for precision indicates a lower rate of false positives. The third is recall. It is the ratio of the total number of positive entries to the number of positive entries that are correctly predicted. This ratio is expressed as a percentage, and it basically indicates the fraction of positive observations that were categorized correctly. Lastly is the F1 score. It is the weighted average of precision and recall. It takes both false negatives and false positives into account. For a problem in which the classes are unbalanced, the F1 score is a more reliable metric than accuracy [21].

Also, I set aside all my prejudgments by analyzing the text through a quantitative corpus-driven analysis. I am dedicated to engaging in a thorough and unbiased examination of facts, theories, and evidence. This process was necessary to keep a balance between subjectivity and objectivity.

I also emphasized the need for thorough expert debriefing and analysis of quantitative-driven corpora. This was done to secure the accuracy and reliability of the data. The results that emerged from the experts' review were discussed and interpreted objectively. I also employed the help of an expert debriefer to confirm my analysis, interpretation, and discussion.

Lastly, the gathering and analysis of data started in July 2023 and ended in October 2023. More so, I believe that this time duration is sufficient to get this study done on time.

RESULTS AND DISCUSSION

Classification of Hate Speeches Found in Infidelity Videos Through Computational Linguistics

The table below presents the general difference between the number of hate speeches and non-hate speeches present among

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

the infidelity videos from Raffy Tulfo in Action YouTube channel through the computational tool named Long Short-Term Memory (LSTM). Here, we can see that there are 471,714 or 72% of hate-contained speeches while only 180,777 or 28% for non-hate speeches. These data were generated when LSTM was categorized, as depicted in the table below. Further, under its classification per target, hate speeches pertaining to sex garnered the largest dataset, totaling 328,836 or 50%. It is then followed by hate speeches targeting or based on a quality that does not fall under any of the other targets, totaling 320,116 or 49%. Physical attributes received 75,866 or 12% hate speeches while targeting individuals with disabilities received 33,202 or 5% hate speeches. Further, hate speeches targeting someone's religion received 19,290 or 3% hate speeches, while hate speeches based on racial attributes received 6,170 or 0.9% hate speeches. Lastly, class-based hate speeches, likely related to socioeconomic status or social class, received the lowest number of hate speeches, amounting to 1,538 or 0.2%.

Table 1.
Hate and Non-Hate Dataset Composition

Label	Number of Comments & Transcriptions	Sample Lines
Hate Speeches	**471714**	
Race	6170	*buhok pa lang nung lalaki pang bisakol na e magpapabuntis ka pa dyan hahahaha* **-Data No. 153634** (Even the man's hair looks like a Bisakol, and you're still trying to get pregnant there, hahahaha.)
Sex	328836	*babaeng paiyot mababang uri mababa ang lipad dapat dyan pinapakulong.* **- Data No. 2293** (Flirty woman, low class, has low morals, she should be imprisoned.)
Physical	75866	*ang ganda ng original tapos ang pangit ng kabeet mukhang tae, yawa.* **-Data No. 104895** (The original is beautiful, but the mistress is ugly, looks like feces, damn.)
Disability	33202	*..cge mn ug panilap.. murag halas mn.. wa pa cguro ka tugpa.. mongoloid mn ang laki..* **-Data No. 102030** (You keep on licking… acting wild… maybe you are not yet cured. You look like a Mongoloid.)
Religion	19290	*i think this girl is a muslim …sorry but i really hate muslims,,,poor baby ..stay strong kuya,,,makapal mukha ng mga cheater.* **-Data No. 4380** (I think this girl is a Muslim… sorry, but I really hate Muslims. Poor baby, stay strong, brother. Such thick-faced cheaters.)
Class	1538	*ipakulung yan oi pistii, pobre raba kaayu unya gapangabit ewwww.* **-Data No. 322809** (Put that person in jail, damn it, they are very poor and yet they're having an affair. Ewwww.)
Quality	320116	*17ongoli te.. sana mamatay ka nalang.* **-Data No. 29808** (You're overreacting… I wish you would just die.)
Non-hate Speech	180777	*bakit ganito ang mga comments? Parang mga 17ongoli naagawan ng kendi…* **-Data No. 174** (Why are the comments like this? It's like children who had their candy taken away…)
Total	**652491**	

The table also mentioned the presence of 180,777 non-hate speeches. It is also imperative to share sample data that represent the mentioned category. In this study, non-hate speeches are categorized and denoted by code 0 during the deep learning analysis using LSTM. Further, no classification of non-hate speeches was made in this study since it was not presented in the adopted framework of analysis.

These findings corroborated the study's concept of hate speech classification by Waseem and Hovy. They provided criteria for identifying hate speech and non-hate speeches derived from the Critical Race Theory by Bell in 1970. Specifically, they conceptualized rules and indicators for categorizing a particular remark under hate and non-hate speeches [19]. The findings also agreed with the classification schemes of hate speeches of Silva et al. in identifying the targets of hate speeches present in the corpora. They postulated that hate speeches can be grouped under the following targets: race, sex, physical, disability, religion, class, and quality [20].

Further, the results of this study also corroborate with its foundation study from Silva et al.'s framework that categorized

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

hate speeches into seven distinct targets — race, sex, physical attributes, disability, religion, class, and quality. The findings of this study reveal that sex and quality were the most prevalent categories, comprising 328,836 and 320,116 instances, respectively. Subsequently, fewer instances were observed in the categories of physical, disability, religion, race, and class, which are ranked in descending order [20].

Use of Computational Linguistics to Identify Linguistic Features of Hate Speeches

In this section, the researcher utilized LSTM to generate and locate hate-containing words as well as other words that are directly related to their respective targets. This is further done by obtaining their numerical frequencies, which means that this is the number of times being used among the nine million tokens that were gathered.

With these, the researcher provided tables from Table 2.1 to Table 2.7, which contained the varied lexicons per target, corresponding spelling variants, and their respective sample comments. Note that some lexicons per target may overlap in terms of their meaning, but they were still chosen by the researcher due to their unique use and the presence of other linguistic functions they portrayed within the comments and transcriptions.

Table 2.1 below indicates the present linguistic features of the ten selected lexicons that are related to the target of race, especially in terms of their spelling variants. These are the lexicons that express comments towards or on the basis of race, ethnicity, or nationality. The first sample lexicon is “peenoise”.

Table 2.1
Linguistic Features of the Lexicons Targeting Race

Lexicon	**f**	**Orthography** (Spelling Variants)	**Sample Lines**
negraaa	341	Negraaaa Negraha Negrang Negrat Negrita Negritaaaaaa Negritang Negro negron	*ganda ni missis pero si kabit ay negraaa* **-Data No. 297868** (Missis is beautiful, but the mistress is a Negraaa.)
mongoloid	39	Mongoid Mongoiloid Mongol Mongol10 Mongolayd Mongolepsy 18Ongolia 18Ongolian Mongolloid Mongoliod Mongoloyd Mongolyd Mongolyod	*anoh bah kc nakita mo sa mongoloid na 2.saksakan ng pangit.* **-Data No. 511374** (What did you see in that Mongoloid? He is very ugly.)
moklo	3	Muklo	*moklo man diay na, dabdabi na.* **-Data No. 6849** (She is a “moklo”. Burn her!)
peenoise	2	Pinoy Pinoys Penoy Filipinoy	*kabit 18ongol ang tapang pa. peenoise talaga.* **-Data No. 500641** (The mistress is wise. Despite being a mistress, she acts confidently. Truly, a ‘peenoise.’)
bisakol	2	Bisakul	*nung lalaki pang bisakol na e* **-Data No. 153634** (By just looking at his hair, you would know he is a ‘Bisakol.)

The next linguistic analysis focuses on lexicons related to sex and sexuality, examining their linguistic features. These terms may encompass lexicons that express opinions, attitudes, or derogatory comments about individuals based on their sexual

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

identity or behavior. Table 2.2 presents the ten selected sex-related lexicons that are selected because of their hate-containing characteristics or that they are directly related to the said target.

Table 2.2
Linguistic Features of the Lexicons Targeting Sex

Lexicon	**f**	**Orthography** (Spelling Variants)	**Sample Lines**
kipay	54	kipey kipks kipkip kipwang kipyas	*ipakulong nalang yan…libog ni kuya at kati ng kipay ng kabit* **-Data No. 37391** (Just put that person in jail. Brother is horny, and the vagina (kipay) of the mistress is itchy.)
palautog	26	palauttug parautog pautog utogan utoggg	*palautog man mura mag iro ning kigwa.* **-Data No. 100289** (This idiot is horny (palautog) like a dog.)
jer2	6	jejerjer jer2x jer jerbaks	*ahahaha kabibo nila ui hahaha jer2 pa more.* **-Data No. 472516** (They are having fun, that what you get for having “jer2”.)
Paeut	3	paeutin paiyots paiyut	*kapal ng mukha mo. Ikaw nga nag paeut sa may asawa* **-Data No. 370305** (You have the nerve. You’re the one who had sex (paeut) with someone else’s spouse.)
u10	3	uten otien otin tttt ttttiiiii ttttttt tite titeng titi tt	*utak gamitin at wag ang u10* **-Data No.227708** (Use your brain and not your dick (U10).

The next tabular data of Table 2.3 which is presented above explored the selected lexicons that express hate or criticism based on physical characteristics. These terms are designed to insult or demean individuals by targeting their appearance.

Table 2.3
Linguistic Features of the Lexicons Targeting Physical Attributes

Lexicon	**f**	**Orthography** (Spelling Variants)	**Sample Lines**
kokey	559	kokey kokeey kokeeee koke kokeng kokeyyyyyyy kokeyyyy koki kokie kokkey	*parang kokey mukha ng kabit* **-Data No. 569667** (Her face is like ‘Kokey.)

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

shokoy	18	Shokey	*mukhang shokoy ang pinagaagawan* **-Data No. 600988** (They're fighting over someone who looks like a 'shokoy'.)
yobab	5	yobabs baboyy baboyyyyyyyy	*ang sarap sipain nung babaeng yobab!* **-Data No. 11095** (The 'yobab' girl is so enjoyable to kick!)
jungit	4	junget	*jusko jungit naman asawa mo teh!* **-Data No. 503304** (Oh my, your spouse is really 'jungit'!)
balmond	3	balmon balmont	*baka kamukha 20ongolian20 ang maging anak.* **-Data No. 271689** (Their child might look like 'Balmond.)

The next table under Table 2.4 explains the unique lexicons targeting disability. These are the lexicons that express hate or non-hate opinions towards or on the basis of a health condition, including but not limited to a physical, mental, sensory, or emotional disability or impairment.

Table 2.4
Linguistic Features of the Lexicons Targeting Disability

Lexicon	**f**	**Orthography** (Spelling Variants)	**Sample Lines**
buang	1882	buang2 buanga buanget buangit buangon buanq	*buang ka babae ka mukha kang titi* **-Data No. 12521** *(You're an idiot(buang), woman. Your face looks like a penis.)*
pakno	6	none	*walang ecip ang pakno.* **-Data No. 23523** *(The pakno has no brain.)*
boduy	6	ambudoy bodoy budoy budoybudoy budoyyy boboe obob	*parang si boduy ang pangit ng tawa* **-Data No. 132235** *(He is like 'boduy,' the laughter is ugly.)*
otistic	6	autism autistic	*otistic c boy at monggoloid c kabit.* **-Data No. 109399** *(The boy is otistic and the girl is monggoloid.)*
taitok	2	none	*naa man guro kay taitok gurl* **-Data No.** 84801 (Gurl, you might have a "taitok".)

Table 2.5 discusses the linguistic features of the selected lexicon targeting religion. These selected words may or may not directly express hate towards or on the basis of religious affiliation or belief.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

Table 2.5
Linguistic Features of the Lexicons Targeting Religion

Lexicon	f	Orthography (Spelling Variants)	Sample Lines
demonyo	2977	dedemonyohin demonyoan demonyoca demonyoha demonyohan demonyohin demonyoka demonyolica demonyooooooooooo	*angelica, pero demonyo ka.* **-Data No. 640775** (Angelica (is the name), but you're a demon.)
y*waa	44	yawa2 yawaa yawaaa yawaaaa yawaaaaa yawaaaaaa yawaaaaaaa yawaag yawoo yawooo	*kalamiay kulatahon ning y*waa nga feeling guapo.* **-Data No. 259616** (It's disgusting to entertain this demon who thinks he's handsome.)
taning	29	satanas	*ng aantay na sa kanila c taning sa impyerno* **-Data No. 209978** (Taning is waiting for them in hell.)
Jablo	3	Jablu Dyablo Dyablos Dyabyo Diyablo Diyablong	*ate na kapatid ni jablo dapat di kana sumabat pa.* -Data No. 183702 (You should not interfere lady who is a sibling of "jablo.)
quibolok	1	quiboloy quibs quibuloy	*kampon din yata to ni quibolok c ate* **-Data No. 541242** (I think she is also a member of quibolok.)

Next in line are selected lexicons targeting or relating to class which are found in Table 2.6. These are the lexicons that express hate towards or on the basis of social class or socioeconomic status.

Table 2.6
Linguistic Features of the Lexicons Targeting Class

Lexicon	f	Orthography (Spelling Variants)	Sample Lines
kafal	24	ankafal Kafalll Kafallll kafalmuks	*kafal nung lalaki ah* **-Data No. 72454** (The guy is shameless.)
abogaga	16	Abogago abogagao	*un nanay parang abogaga ng anak nya….* **-Data No. 72007** (Her mom acts like an "abogag" for her child.)
squami	6	squammy squamy Skwammy Skwamy	*hairstyle at diy na braces palang ng lalake ,halatang walang kwentang squami.* **-Data No. 65291** (With that hairstyle and DIY braces, the guy seems a worthless "squami".)
poorita	4	Poorito Poorpes	ewww 5,000, poorita! **-Data No. 256395**

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

		poorever	(Ewww, 5,000, 'poorita!)
felingera	4	Felingero Felingon Felings Feling felengrera	*felingera kang kabit ka.* **-Data No. 139961** (You are a pretentious mistress.)

The last table discussing the linguistic features of the selected lexicons is the words targeting quality, which can be found in Table 2.7. These are the lexicons that either express hate or non-hate but are still related to the target or based on a quality that does not fall under any of the previously mentioned targets.

Table 2.7
Linguistic Features of the Lexicons Targeting Quality

Lexicon	**f**	**Orthography** (Spelling Variants)	**Sample Lines**
ogag	450	Ogagg Ogaggg Ogago Ogags Ogak Gago gago qaq0 qaqi qaqu qaqo	*saya ng mga demonyo, mga ogag!* **-Data No. 266209** *(The demons are happy, you fools!)*
fishtea	62	Fishte Fishteaaaa Fishteng fishty	*mukha naman bakla amfota* **-Data No. 170054** (Looks like a gay guy, the fuck!.)
amputek	58	Amputa Amputaa Amputaaaaaa Amputaaaka Amputah Amputahh Amputahhh Amputang amputangina amputha	*nauutal utal pa, amputek!* **-Data No. 590147** (Having stuttering, amputek!)
animels	11	Animal Animels Animelsss Animelz animl	*animels! ipkulong nlng yan o putulan ng hootenn.* **-Data No. 35589** (Animals! Just lock them up or cut off their dicks.)
shut@	2	Shutaa Shutaaaa Shutaaaaaaa Shutaina Shutaena Shutacca Shutaca Shutainamez Shutanamels Shutanamers Shutanames Shutaness	*hay naku kuya shut@ ka* **-Data No.35410** (Oh, brother, you shit!)

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

In summary, the linguistic features of these lexicons under this research question illustrated the dynamic, context-dependent nature of hate speech and expressive elements to convey derogatory beliefs about targeted individuals or groups across various targets of hate, including race, sex, physical attributes, disability, religion, class, and quality.

The researcher's results agree with a number of study findings. They postulated that LSTM models can also be an effective approach for identifying spelling variations within a dataset. LSTM models excel in discerning patterns within sequential data, particularly in text. This capability makes them well-suited for applications such as spelling detection and normalization. Within this framework, spelling normalization is conceptualized as a task involving character-based sequence labeling, and the appropriateness of employing a deep bi-directional LSTM model is investigated. It is crucial to recognize that the efficacy of these models is contingent upon the quality and representativeness of the training data, as well as the specific characteristics of the text data and the nature of the problem at hand [23], [24], [25].

This study also corroborate with the findings of another study where highlighted that participants extensively utilized virtual spellings like 'bz' for 'busy,' 'wid' for 'with,' and 'u' for 'you.' These novel communication practices have swiftly emerged, contributing to innovative orthographic features within English words. The study concluded by suggesting that the permanence of these orthographic changes in English orthography would be determined over time [26].

Lastly, a study findings using the Keyword in Context (KWIC) approach employing Mozdeh's concordance to analyze the words used by people in social media resulted that there were numerous laugh variants (e.g., hahahahahahahaha, hahaha), along with abbreviations like 'lmao' and 'lol.' The researcher also discovered terms with numerous spelling variations, such as 'this' (88 variants), 'screaming' (84), 'slayed' (76), 'ahmazing' (74), 'sick' (69), 'bomb' (69), 'preach' (66), 'gorg' (55), 'lush' (54), 'omfg' (53), 'poppin' (50), and 'lit' (49). The diverse ways these words are spelled offer insights into sentiment analysis and online communication culture [27].

Difference in terms of the Linguistic Features of Hate Speech from Non- Hateful Speech Through the Use of Long Short-Term Memory (LSTM)

In this study, the researcher employed the Long Short Term Memory (LSTM), a type of deep learning model known for its capability to retain information over extended sequences, a crucial feature in language-related tasks. LSTM network was able to classify, process, and make predictions based on time series data. This successfully analyzed the text and speeches since LSTM was properly trained to identify patterns and structures in the text that are indicative of hate speech. Given its proficiency in handling sequential data, which in this study's case are tokens, the model processes individual words through the tokenization process, embedding and padding, retains them in its memory, and recalls information acquired from preceding words. To explain further, tokenization helps the model understand the structure of the input text by representing it as a sequence of discrete tokens, thus, capturing linguistic features associated with hate speech. In another hand, embedding allows the model to understand the contextual and semantic relationships between words. Lastly, padding ensures that each input sequence has the same length, allowing the model to process multiple sequences simultaneously. In all, these steps collectively contribute to the model's ability to differentiate linguistic features of hate speech during training and testing phases.

Table 3 shows the restructuring of the dataset, especially considering imbalanced class distributions and downsampling of the majority class. By addressing biases and ensuring the representation of specific hate speech categories, the dataset preparation can facilitate a more accurate and balanced learning process for the model. Further, the dataset was split into 70% for training and 30% for testing, similar to our studies that required training and testing.

Table 3.
Hate and Non-Hate Classification After Downsampling

Label	Number of Comments & Transcriptions
Hate	**180777**
Race	2331
Sex	125929
Physical	29115
Disability	12674
Religion	7443
Class	586
Quality	122514
Non-hate	**180777**
Total	**261554**

The data splitting was vital when using a deep learning, as it helps avoid overfitting, thus obtaining the best result. On the contrary, when the data is overfitted, which means the training is lower than the testing, the model cannot generalize and fits too closely to the training dataset instead. It is understood that typically, a 70-30 percent data split yields optimal results.

The LSTM operates by initially training on a labeled dataset, using 70% of the collected data for this purpose. The training involves adjusting the model's parameters to minimize the difference between predicted and true labels. Subsequently, the model's accuracy is assessed using the remaining 30% of the

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

dataset as a testing set. The training dataset consists of labeled tokens, indicating instances of hate speech or non-hate speech, with additional labels specifying the specific targets of hate for instances categorized as hate speech.

Table 4.
LSTM Hyperparameters

Hyperparameter	Value
Number of Nodes	256
Batch Size	256
Maximum Number of Words	4906
Maximum Sequence Length	818
Epoch Size	5
Learning Rate	0.01
Loss Function	binary_crossentropy
Activation Function	ReLu
Optimizer	Adam
Dropout Rate	50%

During the training phase, the model adjusts its internal parameters to minimize the difference between predicted and actual labels. This iterative process enables the LSTM to recognize patterns and linguistic features associated with hate speech. Once trained, the LSTM applies its learned patterns during the testing phase to classify new texts as either hate speech or non-hate speech based on the acquired knowledge.

Since the original dataset composition, shown in Table 1, is highly imbalanced, in order to remove bias, the majority class was down-sampled, resulting in the new dataset shown in Table 3. Afterward, the dataset was split accordingly, and it was later employed using the LSTM architecture with adjusted hyperparameters, as shown in Table 4.

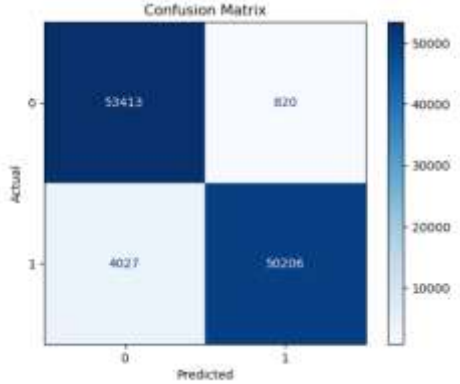
Every hyperparameter plays a crucial role in fine-tuning the model. The value of nodes indicated above can enable the model to learn more complex patterns in the data, which can be beneficial for capturing intricate linguistic features. Batch size represents the number of samples processed in each iteration during training. The mentioned batch size can lead to more stable updates of the model's weights but may require more memory. It can impact how the model generalizes linguistic features from the training data. The specified maximum number of words and maximum sequence length are determined by characteristics inherent to the dataset utilized in this study. A larger vocabulary allows the model to potentially capture a more diverse range of linguistic features while the specified sequence length helps the model handle varying lengths of text.

Concurrently, the epoch size and dropout rate mirror those used by other researchers having related tokens of the study. The epoch size refers to the number of times the algorithm worked through the entire training dataset, meaning that the model went through the training data 5 times during the training phase. Meanwhile, the dropout rate is a regularization technique used to prevent overfitting. Furthermore, the choice of the loss function, activation function, and optimizer is also employed LSTM for hate speech detection in social media. These features are crucial since they help the model effectively learn and generalize the dataset. Having the right combination of these parameters can help the model to converge to an optimal solution and make accurate predictions on new, unseen data.

At the same time, the remaining values were chosen by the researcher and are set to their default values. After the implementation of the LSTM Model, the performance evaluation and the confusion matrix were computed. These metrics are needed to see how well the LSTM model performs hate speech detection. Each evaluation metric – Precision, Recall, F1 Score, and Accuracy provides a perspective on the model's strengths and weaknesses.

Also, it avoids biased assessments and, instead, provides a more nuanced understanding regarding the model. Additionally, the confusion matrix helps clearly visualize the model's predictions into true positives, true negatives, false positives, and false negatives. The result of the confusion matrix is found in Figure 1 below.

Figure 1.
Confusion Matrix for LSTM



ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

For this study, 50206 were identified as true positives, meaning that these are the number of hate speeches that were correctly identified as hate speech; 53413 were true negatives, meaning that these are the non-hate speeches correctly identified as non-hate speech; 4027 were false negatives meaning that these are the number of classified texts as non-hate speech but are actually hate speech and 820 were false positives meaning that it identified texts as hate speech but are actually non-hate speech, as depicted in Figure 2. Through this confusion matrix produced by LSTM, we can see how the linguistic features of hate speeches and non-hate speeches were processed, differentiated, and categorized. More so, it can be observed that in this study, only a few mistakes were made by the LSTM, as observed in the huge discrepancies of values between true positive and false positive and between true negative and false negative, which means that LSTM has performed well during the process.

Furthermore, this matrix also allows for the performance metrics to be derived, as presented in Table 5.

Table 5.
Performance Evaluation of LSTM

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.9553	0.9839	0.9257	0.9540

In assessing the performance of the LSTM model in this study, various metrics were employed to measure its efficacy in detecting hate speech. The model's accuracy, reaching 95.53%, signifies the overall correctness of its predictions, encompassing both true positive and true negative instances. Precision, representing the ratio of correctly predicted positive instances to the total predicted positives, yielded a substantial value of 98.39%. This implies that when the model designates a text as hate speech, its accuracy exceeds 98%. The recall, gauging the model's capacity to capture all genuine instances of hate speech, recorded a value of 92.57%. The F1 score, a harmonic blend of precision and recall, was computed at 95.40%. These metrics – all above 90%, collectively underscore the model's robust performance in accurately discerning both hate speech and non-hate speech instances, maintaining equilibrium in minimizing false positives and false negatives. The high precision indicates low false-positive rates, and the relatively high recall suggests effective capture of actual instances of hate speech. Overall, the combination of these metrics reflects the robustness and reliability of the LSTM model in hate speech detection.

The result agrees with the algorithm model of Long Short-Term Memory (LSTM) that through the process of tokenization, embedding, padding, sequential dependencies, training, testing, evaluating its performance, and other important recurrent processes, it helps LSTM assess hate speech and non-hate speech linguistic feature [28].

The study's result also approved the contention that LSTM networks can be trained to recognize linguistic features of hate speech. They also added that a confusion matrix can help assess the linguistic features of hate speeches and non-hate speeches. A confusion matrix is a table that is used to evaluate the performance of a classification model. It shows the number of true positives, true negatives, false positives, and false negatives [29].

The mentioned results above are also supported by various researchers' findings utilizing various machine learning or deep learning methods. The study's overarching conclusion affirms the LSTM model's robustness and reliability in accurately discerning hate speech and non-hate speech instances based on linguistic features. This finding holds significant implications for real-world applications, particularly in content moderation across various platforms where hate speech detection is a critical concern [21].

Further, the results of this study are also corroborated by the study conducted which demonstrates the efficacy of the LSTM network classifier in achieving a notable accuracy of 86%. The implementation of an early stopping criterion based on the loss function during training enhances the model's performance. The findings underscore the potential of LSTM networks in effectively discerning hate speech, thereby offering a valuable contribution to the ongoing efforts to mitigate the proliferation of toxic content in online spaces [30].

Concluding Remarks

In conclusion, this study has been a complex and challenging exploration of the intricate world of hate speech within the context of infidelity videos through the help of computational linguistics. The researcher's investigation into the linguistic features of hate speeches, employing advanced computational linguistics techniques, aimed not only to uncover patterns within these expressions but also to contribute meaningfully to fostering a safer online environment. As we advance, the intersection of language study, computer learning, and ethical considerations becomes increasingly important. Exploring hate speech involves more than just using algorithms; it requires a careful balance between technological advancements and ethical responsibility. The researcher aims to foster a digital space where diverse voices can coexist, engage in meaningful dialogue, and contribute to a safer, more inclusive online community.

REFERENCES

1. *Imaduddin, F. (2018). Ujaran Kebencian.*
2. *Azman, N. F., & Zamri, N. A. K. (2022, September). Conscious or Unconscious: The Intention of Hate Speech in Cyberworld – A Conceptual Paper. In Proceedings (Vol. 82, No. 1, p. 29). MDPI.*
3. *Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). Countering Online Hate Speech. UNESCO Series on Internet Freedom. Paris: UNESCO.*
4. *Hardaker, C., and McGlashan, M. (2016). 'Real men don't hate women': Twitter rape threats and group identity. J. Pragmat. 91, 80–93. Doi: 10.1016/j.pragma.2015.11.005*
5. *Baider, F. H. (2019). Le discours de haine dissimulée: le*

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 10| Issue: 1| January 2024|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2023: 8.224 || ISI Value: 1.188

mépris pour humilier. Dév. Soc. 43:359. doi: 10.3917/ds.433.0359

6. *Matamoros-Fernández, A. (2017). Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Inform. Commun. Soc. 20, 930–946. Doi: 10.1080/1369118X.2017.1293130*
7. *United Nations, (2020). Hate speech on Facebook poses 'acute challenges to human dignity' – UN expert. https://news .un. org/en/story/2020/12/1080832*
8. *Fincham, F. D., & May, R. W. (2017). Infidelity in romantic relationships. Current opinion in psychology, 13, 70-74.*
9. *Buss, D. M. (2021). When men behave badly: The hidden roots of sexual deception, harassment, and assault. Little, Brown Sparks.*
10. *Apostolou, M. (2019). Why Greek-Cypriots cheat? The evolutionary origins of the big-five of infidelity. Evolutionary Behavioral Sciences, 13, 71– 83. https://doi.org/10.1037/ebs0000140*
11. *Apostolou, M., Constantinou, C., & Zalaf, A. (2022). How people react to their Partners' infidelity: An explorative study. Personal Relationships.*
12. *Biere, S., Bhulai, S., & Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. Master Business AnalyticsDepartment of Mathematics Faculty of Science.*
13. *Creswell, J. W. (2013). Steps in conducting a scholarly mixed methods study.*
14. *Biber, D., Conrad, S., & Reppen, R. (2002). Corpus linguistics: Investigating language structure and use. Cambridge University Press.*
15. *McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press. *
16. *Bender, E. M., & Langendoen, D. T. (2010). Computational linguistics in support of linguistic theory. Linguistic Issues in Language Technology, 3.*
17. *Devopedia. 2020. "Text Corpus for NLP." Version 5, December 20. https://devopedia.org/text-corpus-for-nlp*
18. *Hays, D. G. (1967). Introduction to computational linguistics. https://www.amazon.com/Introduction-computational-linguistics-David-Hays/dp/B0000CNJX0*
19. *Waseem, Z. (2016, November). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science (pp. 138-142).*
20. *Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 10, No. 1, pp. 687-690).*
21. *Chayan, P., & Bora, P. (2021). Detecting hate speech using deep learning techniques. International Journal of Advanced Computer Science and Applications, 12(2).*
22. *Richards, R. (2023, September). How to Use Whisper AI: The Only Guide You Need. http://surl.li/ohcmq*
23. *Bollmann, M., & Søgaard, A. (2016). Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. arXiv preprint arXiv:1610.07844.*
24. *Bhashkar, K. (2019). Spelling Correction Using Deep Learning: How Bi-Directional LSTM with Attention Flow works in... https://bhashkarkunal.medium.com/spelling-correction-using-deep-learning-how-bi-directional-lstm-with-attention-flow-works-in-366fabcc7a2f*
25. *Zaky, D., & Romadhony, A. (2019, September). An LSTM-based spell checker for 26ndonesian text. In 2019 international conference of advanced informatics: concepts, theory and applications (ICAICTA) (pp. 1-6). IEEE.*
26. *Aslam, R. F. M., Ahmad, A., & Sajid, M. A. (2011). A Study of Orthographic Features of Instant Messaging in Pakistan An Empirical Study. Language In India, 11(1).*
27. *Thelwall, M.A. (2021). ! Identifying New Sentiment Slang through Orthographic Pleonasm Online : Yasss Slay Gorg Queen.*
28. *Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(1735-1780).*
29. *Enriquez, R. C. K., & Estuar, M. R. J. E. (2023, March). Determining Linguistic Features of Hate Speech from 2016 Philippine Election-Related Tweets. In 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD) (pp. 1-6). IEEE.*
30. *Bisht, A., Singh, A., Bhadauria, H. S., Virmani, J., & Kriti. (2020). Detection of hate speech and offensive language in twitter data using lstm model. Recent trends in image and signal processing in computer vision, 243-264.*