2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC) | 979-8-3503-7304-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/KHI-HTC60760.2024.10481686

Decoding Digital Hostility: Examining and Addressing Hate Speech on Social Media Platforms

Jawaid Ahmed Siddiqui[1]
Razak Faculty of Technology and Informatics
University Technology Malaysia (UTM)
Kuala Lumpur, Malaysia
jawaid@graduate.utm.my

Siti Sophiayati Yuhaniz[2]
Razak Faculty of Technology and Informatics
University Technology Malaysia (UTM)
Kuala Lumpur, Malaysia
sophia@utm.my

Zulfiqar Ali Memon[3]
Computer Science Department
National University of Computer and Emerging Sciences (NUCES-FAST)
Karachi, Pakistan
zulfiqar.memon@nu.edu.pk

***Abstract*—There is no denying that the widespread usage of social media and the sharing of information have greatly benefited humanity. However, a number of issues have also emerged as a result of this increase in online engagement, most notably the spread of hate speech. Recent research has addressed this growing problem on social media platforms by automatically detecting hate speech in a variety of datasets using a variety of feature engineering approaches and machine/deep learning algorithms. It is interesting, though, that many of these studies—to the best of our knowledge—resort to identifying hate speech messages using traditional feature engineering techniques, which leads to less-than-ideal classification results. This is explained by the shortcomings of the feature engineering techniques now in use, specifically their vulnerability to the word order and word context problems. Therefore, more advanced strategies are desperately needed to address these issues and improve the precision of hate speech identification on social media platforms. To create fundamental lexical benchmarks is the goal. As distinguishing characteristics, our methodology makes use of the power of n-grams, which cover both character and word levels, and skip-grams, which cover both character and word levels. Notably, we successfully identify posts in all three defined categories with an impressive 78% accuracy. The results highlight how difficult it is to discriminate between vulgarities and hate speech. We also conduct a thorough exploration of potential directions for future research projects.**

Keywords: Hate speech, machine learning, Classification, Categorization, n-grams, skip-grams.

I. INTRODUCTION

Enhancing social media security has seen a considerable uptick in research efforts over the past ten years. It is crucial to develop accurate methods for the detection and mitigation of abusive language on online platforms, such as blogs, microblogs, and social networks. There has been an increase in recent scholarly investigations into this field. For instance, whereas [15] focus on recognizing instances of hate speech [10] investigate the topic of detecting cyberbullying. [2] Thorough assessment highlights the fact that this subject has drawn a lot of interest and examination. The focus of [25] research is the identification of racist elements in user-generated content. Section 2 presents a thorough survey of relevant literature. The growing interest among researchers in the convergence of online safety, text analytics and cybersecurity is highlighted by significant events like the workshop on (TA-COS) Text Analytics for Cybersecurity and Online Safety conducted during LREC 2016 and the Abusive Language Workshop (AWL) during ACL 2017. This study sets out to explore the field of identifying hate speech by using a carefully controlled collection of English tweets. The dataset includes three unique classification groups for tweets: (1) offensive language free of hate speech (OFFENSIVE), (2) explicit hate speech (HATE), and (3) material free of problematic elements (OK). This study's major objective is to address the challenge of accurately detecting and responding to instances of hate speech.

The traditional methodology in the area of examining abusive language has primarily adhered to a straightforward binary classification, which encompasses only two distinct categories of speech, such as hate and non- hate as demonstrated by works like [14], [15], and [3]. To be able to effectively distinguish between these two discrete classes, models built from such datasets frequently rely on the predominance of unpleasant or socially undesirable utterances, as highlighted by [11] emphasize the potential for misdirection in classifiers when the absence of profanity or negativity is included in their study from 2011. Even when used alone, profanity may not always be a symptom of hate speech. Profanity is occasionally used without necessarily meaning to offend a specific individual for highlighting or creative resolutions. Secondly, hate speech does not exclusively depend on the use of profanity to express disparaging remarks or threats directed at particular people or groups. The central aim of this project is to construct a comprehensive lexical framework with the ability to differentiate between hate speech and profanity within a meticulously curated dataset. This thoughtfully chosen corpus serves as a crucial arena for evaluating the algorithm's efficacy in distinguishing hate speech from various forms of objectionable language. Through this foundational work, we gain insights into the intricacies of

979-8-3503-7304-2/24/$31.00 ©2024 IEEE

the project's complexity and identify the most challenging aspects that warrant attention and resolution.

II. Literature Review

The use of computer approaches to identify objectionable language has become the subject of an increase in recent research papers. For instance, by extracting pertinent subjects from tweets, [10] employ sentiment analysis to find instances of bullying by utilizing Latent Dirichlet Allocation (LDA) topic models as defined by [6]. Academics are clearly passionate in studying hate speech recognition. As seen by the distinction between hate speech and non-hate speech that was previously discussed, each of these projects, to the best of our knowledge, employs a binary classification system.

Work done by [3], [14], [15], and [9] shows significant contributions to this field. The predominance of English data continues to be a common feature in studies of abusive language, a pattern that is also present in our own research and is primarily caused by the accessibility of pertinent corpora. However, little focus has been placed on investigating abusive language detection in languages other than English. The systems released by [8] for the Arabic context and [7] for the identification and reformulation of profanity in Chinese represent noteworthy breakthroughs in this technique. The introduction of new annotated datasets designed for hate speech and abusive language analysis is a promising development that makes it easier to explore non-English languages within this research environment. Notably, such resources are now available in languages like [25], [1] offering a more convenient avenue for study in this field.

III. Research Methodology

Presented herewith is the dataset employed for identifying hate speech, forming the bedrock of our analytical exploration. For this endeavor, a linear Support Vector Machine (SVM) classifier is coupled with three distinct sets of extracted features: surface n-grams, word skip-grams, and Brown clusters. This section offers a thorough examination of the classifier's characteristics and abilities as well as a detailed explanation of our testing procedures. It is crucial to stress that the assessments made by our testing system are based on the Hate Speech Detection dataset, kindly made available through CrowdFlower. The dataset consists of 14,509 English tweets that have been thoroughly annotated by at least three different annotators. Each tweet had to be assigned to one of three categories by these assessors: those that contain hate speech, those that contain provocative language but do not contain hate speech, or those that are deemed acceptable (OK). Each element in the dataset, along with its matching Tweet5, has one of these three labels attached to it. Table 1 provides a visual representation of how the material is distributed among these three groups to aid in comprehension.

Table I. Tweets and class labels in hate speech

Class	Texts
Hate	2,399
Offensive	4,836
Ok	7,274
Total	**14,509**

Detection dataset are broken down. Tokens are preprocessed to lowercase and URLs and emoticons are stripped from all messages.

IV. Classifier

The classifier of choice for the multi-class classification in our experimental study is the linear Support Vector Machine (SVM). We make use of the powerful LIBLINEAR software's capabilities to complete this particular text categorization assignment [17]. It is significant to highlight that the effectiveness of the LIBLINEAR SVM implementation has been validated for a number of tasks, including language classification, temporal text classification, and native language identification [18]; [19].

V. Features

Our experimental strategy incorporates two different categories of surface features, as follows: Character n-grams, which range in order from (2 to 8) and word n-grams which range i.e. order from (1 to 3), are included in our basic surface n-grams. Character n-grams are extracted even when they cross word borders because the extraction process for n-grams is carried out after all tokens have been converted to lowercase. In addition, we extract bigrams with 1, 2, and 3 skips that have characteristics with the aforementioned ones. These particular properties were purposefully selected to act as approximations for such complicated interactions between words because bigrams alone are unable to capture longer distance dependencies.

VI. Evaluation

The effectiveness of our approaches is evaluated using ten-fold cross-validation. To ensure that each division has an equal number of classes, we employ stratified cross-validation to build the folds [16]. We express our conclusions in terms of precision, a crucial performance parameter. By comparing the performance of our methods to both an Oracle classifier and a majority-class baseline, we can determine how effective they are. Notably, the oracle classifier's final judgment is influenced by the cumulative efforts of all the classifiers described in Table 2. This means that the output of the classifier that produced the right label

for a given instance is taken into consideration. This methodology is used to determine the dataset's theoretical maximum performance. It's important to remember that similar methods have been used in earlier studies like in [18] Native Language Identification and studies focused on identifying languages and language varieties [4], where an oracle classifier analysis is used to determine the theoretical upper bound of performance for shared task datasets.

VII. Results And Analysis

We conduct an analysis of their performance to determine the characteristics' applicability in this situation. Training a single classifier using all of the available characteristics is the first step in our process. We next move on to training a composite model that combines all of our unique characteristics into a single framework. We use both the oracle classifier and the majority class baseline to assess these models. Table 2 succinctly presents the findings of these analyses.

Table II. Classification results under 10-fold cross validation

Feature	Accuracy (%)
Majority Class Baseline	50.1
Oracle	91.6
Character bigrams	73.6
Character trigrams	77.2
Character 4-grams	**78.0**
Character 5-grams	77.9
Character 6-grams	77.2
Character 7-grams	76.5
Character 8-grams	75.8
Word unigrams	77.5
Word bigrams	73.8
Word trigrams	67.4
1-skip Word bigrams	74.0
2-skip Word bigrams	73.8
3-skip Word bigrams	73.9
All features combined	**77.5**

Owing to the significant class imbalance within the dataset, the majority class baseline exhibits notably elevated results. With an accuracy of 91.6%, the oracle demonstrates that none of our features can reliably label a sizable fraction of our data. Here, we observe the superior performance of character n-grams, with 4-grams providing the best overall performance. Unigrams of words also fare well, although performance drops off with longer sequences of letters (such as bigrams, trigrams, or skip-grams). While the other feature types may be better at catching short-term dependencies, the skip-grams may be better at capturing long-term dependencies.

According to research, skip-grams are particularly successful at capturing information that is tightly linked with syntactic dependencies in tasks that depend on stylistic details [18]. Surprisingly, despite combining 5.5 million characteristics altogether, the amalgamation performs no better than a character 4-grams model. Given that character n-gram models outnumber their word-based equivalents, it is unclear if this model will be able to comprehensively include the varied information contained in the three different feature categories. Our focus then shifts to evaluating the learning trajectory displayed by these qualities. Figure 1 clearly illustrates the classifier's learning curve that produced the best overall performance. Notably, the foundation of this particular classifier was the use of 4-gram character characteristics.

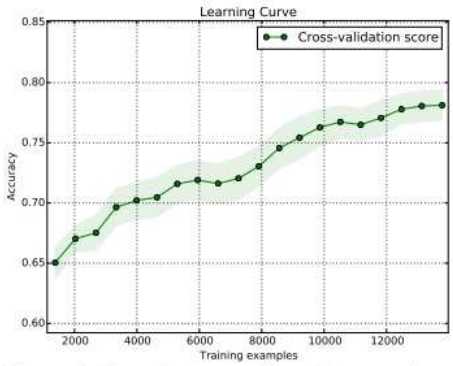

Figure 1: Learning curve for a character 4-gram model, with standard deviation highlighted. Accuracy does not plateau with the maximal data size.

With the expansion of the training dataset, there was a corresponding increase in accuracy, accompanied by a reduction in the standard deviation observed across the cross-validation folds. As a result, it appears that using more training data will result in even higher accuracy. However, after 15,000 training examples, the pace of improvement in accuracy slows significantly.

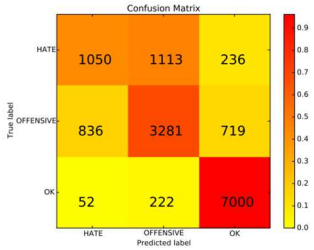

Figure 2: Confusion matrix of the character 4-gram model for our 3 classes. The heatmap represents the proportion of correctly classified examples in each class (this is normalized as the data distribution is imbalanced). The raw numbers are also reported within each cell. We note that the HATE class is the hardest to classify and is highly confused with the OFFENSIVE class.

Last but not least, we have a look at a confusion matrix for the 4-gram character model (Figure 2). This shows that people are more likely to misinterpret hate speech for other types of objectionable content, and that this occurs more

often with hate speech than with other types of offensive content. Many examples of offensive material are incorrectly categorized as neutral. Classifying the vast majority of samples correctly, the non-offensive category emerges victorious.

VIII. Conclusion

This study employs text categorization techniques to differentiate hate speech from various textual expressions, including profanity. In our initial approach, we integrated traditional lexical attributes with a linear support vector machine (SVM) classifier. Remarkably, the character 4-gram model emerged as a standout performer, achieving an impressive accuracy rate of 78%. These results underscore the intricate nature of distinguishing between profanity and hate speech. Strikingly, this study stands as one of the pioneering attempts to identify hate speech within the dynamic landscape of social media, where diverse forms of vulgarity coexist. Importantly, it marks one of the first endeavors to address the formidable challenge of recognizing hate speech on social media while also acknowledging the prevalence of other forms of profane language. A significant portion of notable publications up to now have concentrated on the utilization of binary classification algorithms to differentiate between hate speech and socially acceptable content, as exemplified by works like [15] and [14]. According to [11], classifiers have effectively demonstrated their proficiency in discerning between hate speech and socially acceptable material within the realm of binary classification, leveraging the frequency of offensive phrases.

We recognize several potential avenues for extending the scope of this study. First, the incorporation of more robust ensemble classifiers holds the promise of enhancing overall performance. Second, delving into a linguistic exploration of the most informative attributes could yield deeper insights. Lastly, a comprehensive error analysis focusing on instances that were misclassified may offer valuable insights for refinement. A comprehensive elucidation of these facets is detailed in the forthcoming section.

IX. FUTURE WORK

Our future endeavors encompass a thorough and detailed exploration of classifier ensembles and metalearning, aiming to comprehensively assess their effectiveness within the scope of this project. These methodologies have demonstrated their efficacy within the realm of competitive team tasks centered on text classification. Notable recent instances include the automatic detection of lexical complexity, dialect identification, and native language recognition, as exemplified by works such as [12][13]. Another topic ideal for investigation is revealed by a detailed analysis of the key characteristics for each class in this dataset. Explicit and harsh language carry major value for both HATE and OFFENSIVE classifications, according to our preliminary examination of the most informative word unigrams and bigrams, which presents some uncertainty and hurdles for the classifiers. Regarding the HATE classification, we noticed a rise in the frequency of phrases that refer to certain groups or cultures. Notably, we observed the predominance of grammatical terms among the most impactful bigrams in the OK category, which is a notable finding that calls for further examination. A thorough investigation of these characteristics has the potential to improve feature engineering tactics. A thorough grasp of the difficulties present can be assisted by an error analysis. Given the complicated nature of annotation in the dataset of Hate Speech Detection, as was highlighted by [1] researching this topic could provide significant insights into the performance of the classifiers and offer insights into potential difficulties in the annotation process. As can be seen in Figure 2, most people have trouble differentiating between HATE and OFFENSIVE texts, which are to be expected. However, we also recognize that a great deal of offensive material is incorrectly labeled as being inoffensive. Insights on this can be gleaned through the aforementioned mistake analysis.

References

[1] Femi Emmanuel Ayo a, Olusegun Folorunso, Friday Thomas Ibharalu , Idowu Ademola Osinuga. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. In Elsevier. Computer Science Review 38 (2020) 100311

[2] Bochum, Germany. Anna Schmidt and Michael Wiegand. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics. Valencia, Spain, pages 1–10.

[3] Aggarwal, P. H. (2019). BERT and Two-Vote Classification for Categorizing Offensiveness. 13th International Workshop on Semantic Evaluation. Presented at the SemEval 2019, Association for Computational Linguistics, 678–682.

[4] Castelle, M. (2018). The Linguistic Ideologies of Deep Abusive Language Classification. 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics, 160–170.

[5] D arja Fiser, Toma ˇ z Erjavec, and Nikola Ljube ˇ si˘ c. (2017). ´ Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In Proceedings of the Workshop Workshop on Abusive Language Online (ALW). Vancouver, Canada.

[6] NANLIR SALLAU MULLAH ,AND WAN MOHD NAZMEE WAN ZAINON. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review . in IEEE Access. Digital Object Identifier 10.1109/ACCESS.2021.3089515.

[7] Hamdy Mubarak, Darwish Kareem, and Magdy Walid. (2017). Abusive Language Detection on Arabic Social Media. In Proceedings of the Workshop Workshop on Abusive Language Online (ALW). Vancouver, Canada

[8] Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. (2017). Rephrasing Profanity in Chinese Text. In Proceedings of the Workshop Workshop on Abusive Language Online (ALW). Vancouver, Canada.

[9] Matta Chenna Rao ,Kalyan Chakravarti Yelavarti, AND Nakka Pavan Kalyan. (2023). A Framework for Hate Speech Detection using Different ML Algorithms. in Proceedings of the 7th International Conference on Trends in Electronics and Informatics (ICOEI 2023). ISBN: 979-8-3503-9728-4.

[10] Garima Koushik ,Prof. K. Rajeswari, AND Suresh Kannan Muthusamy. (2019). Automated Hate Speech Detection on Twitter. IEEE Access. ISBN: 978-1-7281-4042-1.

[11] Mahamat Saleh Adoum Sanoussi ,Chen Xiaohua,George K. Agordzo,Mahamed Lamine Guindo,Abdullah MMA Al Omari,and Boukhari Mahamat Issa. (2022). Detection of Hate Speech Texts Using Machine Learning Algorithm. IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC).

[12] Marcos Zampieri, Shervin Malmasi, and Mark Dras. (2016a). Modeling language change in historical corpora: the case of Portuguese. In Proceedings of Language Resources and Evaluation (LREC).

[13] Marcos Zampieri, Shervin Malmasi, Octavia-Maria Sulea, and Liviu P Dinu. (2016b). A Computational Approach to the Study of Portuguese Newspapers Published in Macau. In Proceedings of Workshop on Natural Language Processing Meets Journalism (NLPMJ). pages 47–51.

[14] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. (2015). Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pages 29–30.

[15] Pete Burnap and Matthew L Williams. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet 7(2):223–242.

[16] Vildan Mercan ,Akhtar Jamil, Alaa Ali Hameed,Abdullah MMA Al Omari,Irfan Ahmed Magsi, Sibghatullah Bazai, and Syed Attique Shah. (2021). Hate Speech and Offensive Language Detection from Social Media. International Conference on Computing, Electronic and Electrical Engineering (ICE Cube).978-1-6654-0154

[17] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, XiangRui Wang, and Chih-Jen Lin. (2008). LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9:1871–1874.

[18] Shervin Malmasi and Aoife Cahill. (2015). Measuring Feature Diversity in Native Language Identification. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Denver, Colorado.

[19] Shervin Malmasi and Marcos Zampieri. (2017). German Dialect Identification in Interview Transcriptions. In Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial).

[20] Shervin Malmasi and Mark Dras. (2015). Large-scale Native Language Identification with Cross-Corpus Evaluation. In Proceedings of NAACL-HLT 2015. Association for Computational Linguistics, Denver, Colorado.

[21] Shervin Malmasi and Mark Dras. (2017). Native Language Identification using Stacked Generalization. arXiv preprint arXiv:1703.06541 .

[22] Rawat, T. K. (2019). Feature Engineering (FE) Tools and Techniques for Better Classification Performance.

[23] Schneider, J. R. (2018). Towards the Automatic Classification of Offensive Language and Related Phenomena in German Tweets.

[24] Sharma, S. A. (2018). Degree based Classification of Harmful Speech using Twitter Data. First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Association for Computational Linguistics, 106–112.

[25] Nikolov, A. R. (2019). Offensive Tweet Classification with BERT and Ensembles. 13th International Workshop on Semantic Evaluation. Presented at the SemEval 2019, Association for Computational Linguistics, 691–695.