**Title: Analyzing the efficiency of a lexicon dictionary obtained from weaponizedwords in interpreting the topic distributions of hate-related Tweets**

**Introduction:**

With the rapid growth of social networks, it has now become easier for people to communicate their ideas with other people, express their opinions publicly, and share the ideas they support. Hiding behind the shadow of the expression of free speeches, the generation and propagation of offensive languages and hate speeches are on the rise through excessive use of information diffusion or communication channels. For ensuring a safe cyberspace, the detection of offensive language or hate speeches and finding countermeasures to restrict the propagation of hate speech are of utmost importance.

Hate speech is generally any speech that may contain violent content, targeted actions, or intimidation towards a specific group or individual, discriminative language on the basis of gender, ethnicity, religion, race, nationality, sexual orientation, etc which can create an environment of hostility. In the detection process of hate speeches, the main issue that arises is the ambiguity of hate speeches across countries, languages, communities, etc. One speech that can be considered abusive in one language, may deem acceptable in another language. Several approaches for detecting hate speeches are discussed in the literature, which include word level matching, sentiment analysis, leveraging lexical and linguistic characteristics, multimodal information, knowledge-based features, classification techniques, etc.

While working with text documents, each text can be considered a collection of words. The overall sentiment or intention of the sentence (or, paragraph) can be analyzed by going through the words appearing in that text. A possible approach for detecting hate speech is the keyword matching technique, in which the words in a text will be contrasted against a dictionary of keywords related to hate speech and offensive or abusive languages.

The research questions for this work are provided below:

* How effective the lexicons from WeaponizedWords (WW) are in identifying the class or category of tweets in a hate speech Twitter dataset annotated by CloudFlower users compared to other supervised classification techniques?
* When the lexicons from WW are present in tweets, can the different orientations and attributes (e.g. gender, nationality, ethnicity, etc) of the lexicons be used to interpret the topic distributions of hate-related tweets obtained through Latent Dirichlet Allocation (LDA)?

In this work, we will work with a collection of documents (tweets/comments) from social media containing offensive language and hate speeches. Then, we will fetch the lists of keywords for different topics or categories from a trusted data source (e.g. <https://weaponizedword.org>). We aim to analyze if any of the documents contain any keywords for any topic. Performing this keyword-level analysis can help us examine if the text in a document is inclined toward certain topics of interest. If the text element in a document is aligned with the topics related to hate speech based on keyword matching, the document can be identified as hate speech. While conducting the keyword analysis, multiple other aspects of detecting hate speech can be explored which include understanding the topics that are mostly co-occurring in hate speech, the topics which contribute more to forming hate speeches, clustering the hate speeches, validating if the fetched keywords for different categories are helpful enough to identify a hate speech, etc.

**Literature Review:**

Rodríguez et al. combined social network graphs with sentiment or emotion detection and a clustering technique in order to identify highly discussed topics in Facebook posts containing hate speeches. Some Facebook pages known for promoting hate speeches are assumed to be “seed” nodes. Graph API is then used to form a 3-level graph containing all pages related to the seed nodes and Betweenness Centrality is then used to identify the 50 most influential pages in the graph. For each page, the latest 1000 posts are fetched along with the latest 1000 comments. As all the posts don’t contain hate speech, a dictionary-based approach is used to filter out only potential hate speeches containing swear or bad words and sensitive words which can trigger hatred. For further filtering, two sentiment analysis tools named VADER and JAMMIN are used to discard the posts which potentially don’t contain hate sentiments. The comments of the remaining posts are concatenated with the respective posts to create a collection of documents. The documents, each containing a Facebook post and its comments are converted into TF-IDF vectors after preprocessing of texts and then the documents are clustered into 20 groups using the K-means algorithm. The top 6 words for each cluster are identified to represent the most discussed topics in hate speeches. Although this work shows promising results, different assumptions about the initial seed nodes may yield completely different results. Because of the dynamicity of Facebook as a social media platform, the latest posts and comments change frequently which may alter the results of this work in unpredicted ways which question its performance for real-time data. Again, the replies to the comments can be added to each document to get a more comprehensive understanding of the topics.

Vijayaraghavan et al. proposed a multimodal approach for hate speech detection in tweets leveraging the user demographics and socio-cultural information along with the semantics of the text content. The dataset in this work is accumulated from several publicly available hate speech datasets created using hate keywords or lexicons and manual annotation. Using the data provided by Southern Poverty Law Center (SPLC) on extremist groups on Twitter, around 3k associated user accounts are identified. These 3k accounts are used as seed nodes to create a social graph of Twitter accounts and the top 10k accounts are obtained from the graph. A deep learning-based model with a self-attention layer is used for classifying hate speeches where the tweets, authors' information, and social graphs of users are provided as inputs. The graph of users is used to extract the social context features representing the association of users with any of the extremist groups. Previous work of the author named multimodal demographic classifier is utilized to identify the demographic features of users. A late fusion technique is used to train the deep learning model with both text only and text along with socio-cultural features to understand the influence of each feature in the hate speech classification task. The learned weighted features are used as embeddings to cluster the tweets into 5 categories and the top 5 words from each category represent the latent topic of that category. As there’s no ground truth for the categories, human workers are assigned to validate the accuracy of the clusterings in both approaches. As the final results, the text-only approach obtains a purity score of 0.52 in clustering the hate topics where using social-cultural or graph information along with text yield a purity score of 0.76. This work goes beyond the analysis of only texts by incorporating social and cultural information to categorize hate speeches and understand the importance of features other than texts in hate speech detection.

Abro et al. provided a comparative analysis of different combinations of feature engineering and machine learning techniques for automatic hate speech detection in the publicly available Twitter dataset. Three feature selection techniques namely bigram with TF-IDF, word2vec, and doc2vec are used with eight machine learning algorithms including Support Vector Machine, Random Forest, Naive Bayes, K Nearest Neighbors, etc. The 14509 tweets in the dataset are labeled by CloudFlower into three classes: “hate speech”, “offensive but not hate speech”, and “neither hate nor offensive speech”. The experimental results show that the combination of support vector machine and bigram with TF-IDF yields the best result with 79% accuracy. Although this work can provide a comparison framework for future hate speech detection models, it’s inefficient for real-time predictions. This work classifies the tweets into three hate-related classes but is incapable of identifying the severity of hate speeches. Again, lexicon-based approaches for feature selection weren’t included in the comparative performance analysis.

Davidson et al. conducted experiments to address potential racial biases in five different Twitter datasets annotated for hate speech or abusive language detection. Logistic regression classifiers are trained on the datasets and the predicted outcomes of the hate speech detection task are compared for the tweets written in African-American English (AAE) with the tweets in Standard-American English (SAE). The experiments show evidence of racial biases in all five datasets. When the classifiers are trained on these datasets to predict hate speech, the tweets written in African-American English are labeled as abusive language at significantly higher rates. This bias even persists when hate lexicons or keywords from Hatebase are used to identify hate speeches. In Hatebase lexicons, the terms related to African-American English are over-represented as abusive words. The identification of hate speech using classifiers trained with these annotated datasets and also the presence of hate lexicons can have discriminatory negative impacts on social media users using African-American English. Though this study aimed to detect one dimension of racial bias, further research can be conducted to consider other protected groups. Again, other types of biases in the datasets related to gender, ethnicities, regions, etc can be potentially present which is not examined. Further investigation of hate speech lexicons can be conducted considering the commonly used linguistic terms in protected minority groups that are marked as abusive.

Koufakou et al. have presented the applicability of using hate lexicons to improve the detection of aggressive language. In this approach, multilingual hate lexicons from HurtLex are used to improve the pre-trained word embeddings for English, Hindi, and Bengali language from FastText. The retrofitting technique is used to adjust the vectors representing words in the embedding. For each lexicon fetched from Hurtlex, the unique category of that lexicon is identified and then a set of words from the same category is accumulated which is used for retrofitting or adjusting the embeddings. To achieve the two subtasks: identifying aggressive language and identifying misogynistic aggression, the retrofitted embeddings are used with Long Short Term Memory (LSTM) classifier to predict the labels for both subtasks. The LSTM solution is implemented with the Keras library in Python, and also an additional support vector machine classifier with unigrams is used for Bengali for comparison. The experimental results show that combining the training datasets with an additional dataset named TARC-1 always improves predictions. The impact of retrofitting using HurtLex categories boosts the performance for English and Hindi, but it’s ineffective for Bengali which might be caused by the smaller coverage of Bengali lexicons in HurtLex. The multilingual approach to hate speech detection by improving the word embeddings is the novelty of this work. The results point out that this approach of retrofitting can be more beneficial for predicting tasks in a narrower scope (e.g. misogynistic aggression) compared to the generalized identification of hate speech.

Guo et al. explore two different computer-assisted methods: a dictionary-based approach and an unsupervised topic modeling named LDA for analyzing social big data related to journalism and mass communication. This work aimed to understand the qualitative aspects and proportions of topics discussed in the tweets covering Barack Obama and Mitt Romney during the 2012 presidential election using the two above-mentioned approaches. As the tweets are generally short texts and LDA can yield better results for longer documents, data augmentation is performed to combine four consecutive tweets in a single document. The experimental results show that the topic based on foreign affairs is dominant in the tweets about Obama and the tweets about Romney discuss the issue of taxation the most. To examine the reliability of the results from two different methods, the opinions of human coders are taken into account for randomly selected 100 tweets for getting qualitative insights which show that LDA based approach performs better than the dictionary approach in identifying the topics about Obama and Romney. This work compared the topic distributions of tweets obtained from two different computer-assisted approaches and performed validity tests using human coders. The reliability and validity of the results in this work can be further analyzed using more test samples.

The literature in hate speech research covers a range of methods including dictionary analysis, unsupervised topic modeling, supervised and unsupervised learning algorithms, and even deep learning algorithms like LSTM for hate speech detection. Different sets of textual features of the text contents such as the presence of hate or abusive keywords, frequency of words, semantic meaning, etc are considered for better predicting hate content in texts. Some works in the literature also attempted to leverage the social graph of users in social media and the demographic information of the users for better identifying hate speeches and relevant topics present in the hate-related text. In our work, we aim to analyze the efficiency of using a lexicon dictionary obtained from weaponizedwords.org in identifying hate or abusive text contents and understand if any lexicons present in tweets can interpret the topic distributions of tweets obtained by LDA unsupervised topic modeling.

**Methods:**

1. **Data Source and Data Harvesting Techniques**

In order to examine the efficiency of a lexicon-based approach to analyze hate speech, we can use a dictionary containing lexicons related to offensive or negative languages. **Weaponizedword** site contains up-to-date lexicons for hate speech detection and analyzing negative language patterns. A dictionary of negative and abusive words can be created for our work by fetching the lexicon lists from **weaponizedwords.org** by making API requests. Weaponizedword provides additional attributes for each lexicon including information on the topic genres of the lexicon, the severity of offensiveness, meaning in different contexts, etc. The columns or attributes of the lexicons we are currently most interested in for our analysis are - gender, nationalities, orientation, ethnicities, severity, etc. Till now, a dictionary of 19700 lexicons is created by fetching the terms from weaponizedwords.

As the dataset, a pre-existing dataset of tweets will be used which is annotated by CloudFlower (CF) users into three categories: Hate, Offensive, and Neither. Each tweet is coded by at least three CF users, more if there’s ambiguity in the tweet. The final class label of a particular tweet is decided by taking the highest-voted category. The dataset contains a collection of 25296 tweets, voting count for each of three categories by CloudFlower users, and the derived class with maximum votes.

1. **Data Cleaning and Linking Sources**

We are using two data sources here for hate speech investigation: weaponizedwords for hate lexicons and a Twitter dataset annotated by CloudFlower. With proper authentication, weaponizedwords can respond with 100 hate lexicons per API request. When a list of hate lexicons is fetched to create the lexicon dictionary, we observe that multiple responses can contain the same lexicon information. We create a python data frame for storing the lexicons where the columns contain various attributes of the lexicons including the relation to different topics, offensiveness level, etc. Then, the rows containing the duplicate lexicons are preserved to analyze the characteristics of the same lexicon in different contexts.

The tweets in the annotated Twitter dataset need to be processed for further analysis. The tweets can contain mentions, hashtags, and other junk information which are not important for the text analysis for hate speech detection. For textual analysis, as the preprocessing steps, the tweets are then tokenized, lowercasing is applied, and then stopwords are removed from the tweets. The Twitter handles, mentions and hashtags are identified using regular expressions and added as separate columns for each tweet. After this preprocessing and removals, the list of filtered tokens is obtained for each tweet which can be utilized for further text analysis. A document term matrix is created for the tweets where each row in the matrix represents a tweet and the columns contain the frequency of different terms in the tweets.

1. **Data Analysis Plan**

In this step, we have a dataset of cleaned tweets containing potential hate speech or abusive language. The associated Twitter handles and hashtags are separated from the tweets and stored as other column values. For analyzing the efficiency of lexicons from WW in identifying the abusive texts in tweets, we will perform a keyword-matching technique first to examine the presence of hate lexicons in the tweets. As the target class label is already provided in the dataset, we will measure how well the dictionary of lexicons is doing in recognizing hate speeches. The assumption is that only using lexicons won’t be able to detect all the tweets with hate or abusive language. We will also sort out the tweets which are labeled as hate or offensive by CloudFlower users but couldn’t be captured by matching the lexicons. The analysis of these false negative tweets can provide a better understanding of the capability of the lexicon dictionary in recognizing hate speech content. The accuracy or F1 score obtained from this keyword-matching approach can be compared with the results of other supervised learning algorithms to address the comparative effectiveness of using a lexicon dictionary.

To address the second research question, we aim to understand if the presence of hate lexicons in a tweet can represent the overall genre or topic discussed in that tweet. The lexicons from weaponizedwords come with attributes related to gender, ethnicity, religion, etc. So, if a lexicon related to religion is present in a hate or abusive tweet, the assumption is that the tweet most probably contains religious content. To cross-match and verify this assumption, we will create the topic distribution of the tweets containing hate speech contents (Hate and Offensive class labels). We will leverage a topic modeling technique named Latent Dirichlet Allocation (LDA) to identify and understand mostly discussed topics in each tweet. The topic distributions of the tweets will be contrasted against the genre/attribute information of the lexicons provided by weaponizedwords to understand if the seed lexicons’ attributes (e.g. gender, ethnicity, nationality, religion, etc) for each tweet would be able to interpret the resulting topics yielded by LDA for that tweet. For a large dataset, it can be difficult to manually compare the results from LDA topic distribution with attributes of respective lexicons in a tweet. For initial understanding, we can sample 100 tweets from the dataset to observe the topic distributions of the tweets and analyze the effectiveness of the lexicons from WW in interpreting the topic distributions of the tweets obtained by LDA.

**References:**

1. Rodriguez, A., Argueta, C., & Chen, Y. L. (2019, February). **Automatic detection of hate speech on facebook using sentiment and emotion analysis.** In *2019 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 169-174). IEEE.
2. Vijayaraghavan, P., Larochelle, H., & Roy, D. (2021). **Interpretable multi-modal hate speech detection.** *arXiv preprint arXiv:2103.01616*.
3. Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). **Automatic hate speech detection using machine learning: A comparative study.** *International Journal of Advanced Computer Science and Applications*, *11*(8).
4. Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). **Thirty years of research into hate speech: topics of interest and their evolution.** *Scientometrics*, *126*, 157-179.
5. Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. **"Racial bias in hate speech and abusive language detection datasets."** *arXiv preprint arXiv:1905.12516* (2019).
6. Koufakou, Anna, Valerio Basile, and Viviana Patti. **"Florunito@ trac-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection."** *Proceedings of the second workshop on trolling, aggression and Cyberbullying*. 2020.
7. Guo, Lei, et al. **"Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling."** *Journalism & Mass Communication Quarterly* 93.2 (2016): 332-359.