Explaining Hate Speech Classification with Model-Agnostic Methods

Durgesh Nandini, Ute Schmid

19 September 2022

Content

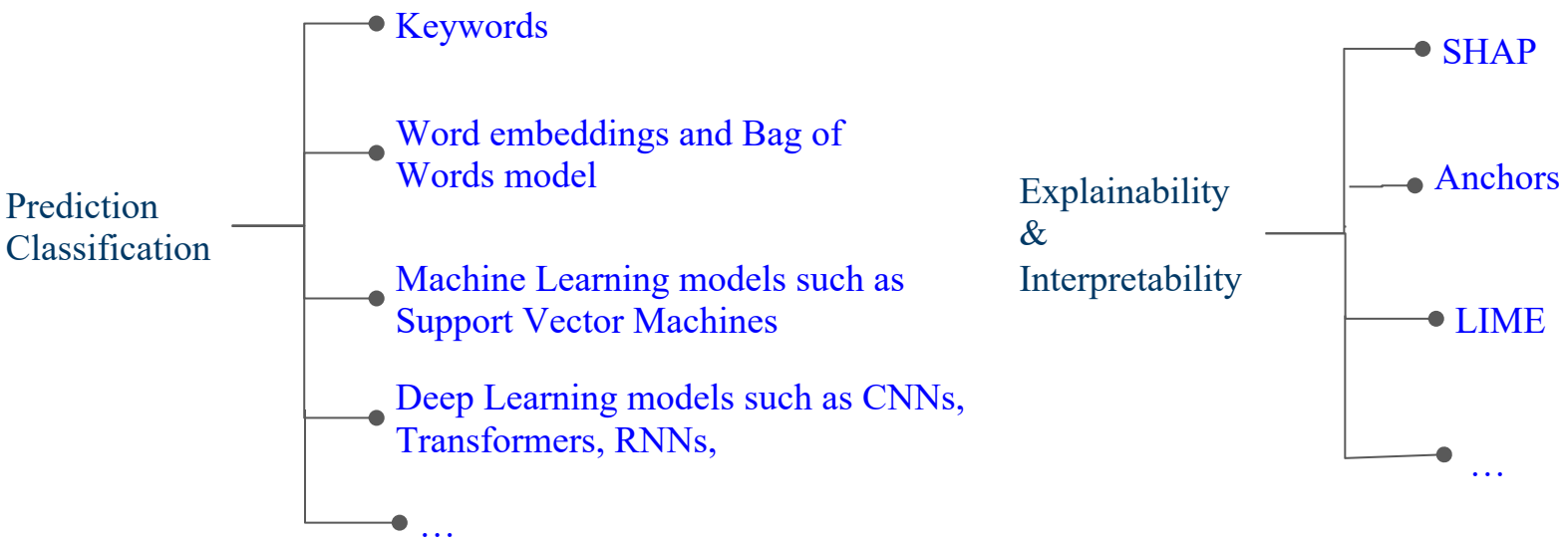
Introduction

- Hate speech detection in dialogues has been gaining popularity among NLP researchers with the increased use of social media.
- What can be defined as hate speech is that it is understood to be bias-motivated, hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics.
- Online hate speech is heterogeneous and dynamic.
- The characteristics that add to the dangers that hate speech poses are accessibility, diversity, instant reaction rates, anonymity and multiplicity.

Research Question

- This paper aims to predict and explain hate speech in tweets in the form of texts.
- The major goal of our work is to provide the basis for coherent, comprehensible, contextual, and realistic explanations with high local fidelity.

Related Work



Methodology

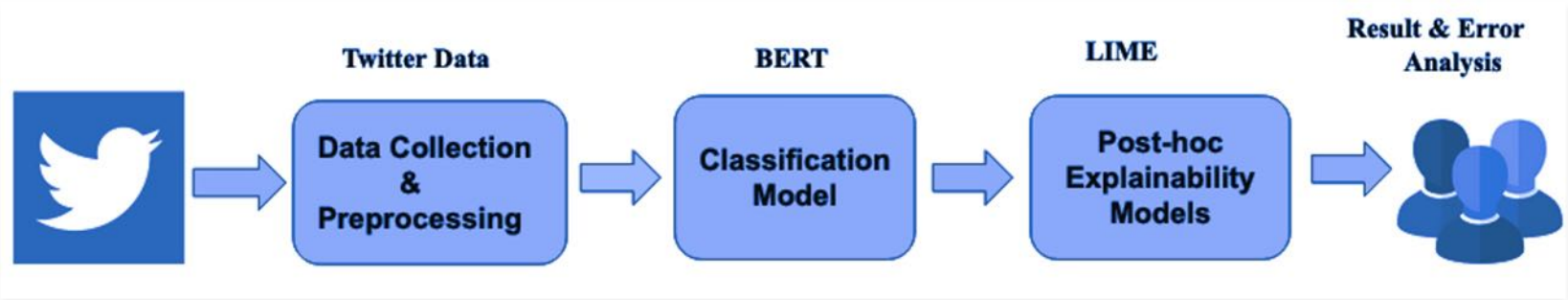


Figure: A Pipeline of the Proposed Methodology

Data

- We used a Twitter dataset for the case study in this work.

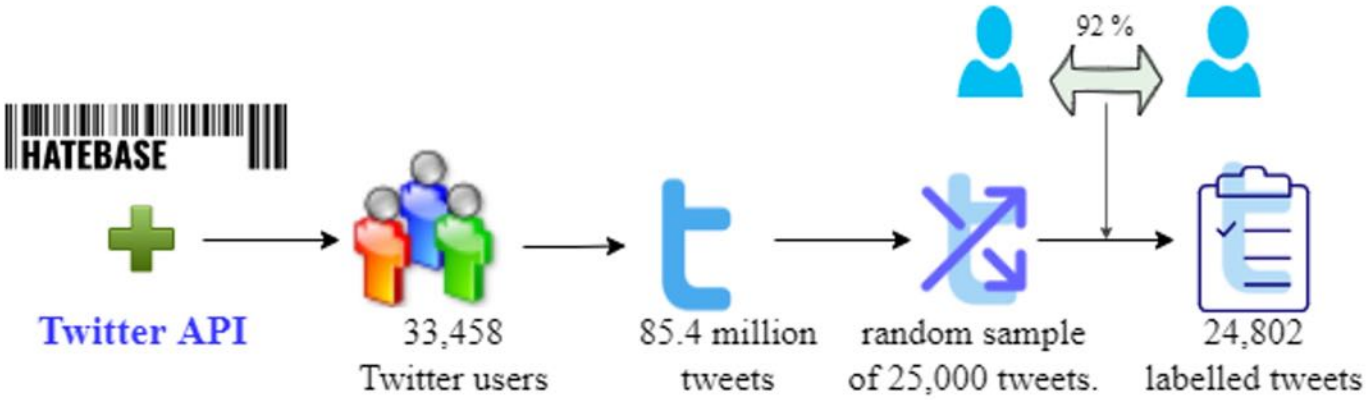


Figure: Data Collection

Data

- We have used three classes from the dataset: 'Hate', 'Offensive' and 'None'.
- Hate: Describing negative attributes or deficiencies to groups of individuals because they are members of a group. There are hateful comments toward groups because of race, political opinion, sexual orientation, gender, social status, health condition, or similar. An example of this category would be "all poor people are stupid".
- Offensive: Posts which are degrading, dehumanising, insulting an individual, or threatening with violent acts are categorised into this category. An example for this category would be "f**king forget that b***h".
- None: Posts that do not belong to any of the above categories are categorised in this set.

Data

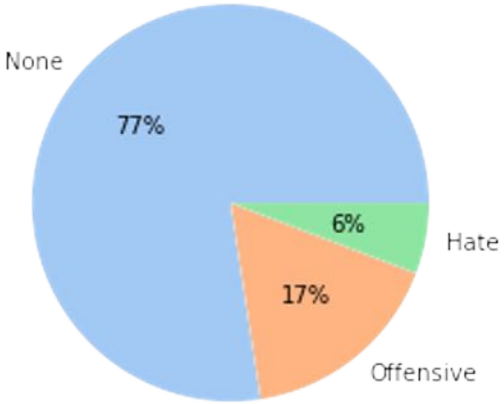


Figure: Data Distribution

Figure: Word Cloud

Data Preprocessing

- For our pipeline, we have preprocessed the data to remove smileys, emojis and any other symbol that may be present.
- In addition to that, we have also eliminated stopwords as the model performed almost the same when with or without the stopwords.
- The hashtags were not eliminated because we observed that the hashtags contributed to the meaning of sentences and would often encapsulate the emotions of sentences

Classification Model: BERT

- The data is trained using the Bidirectional Encoder Representations from Transformers (BERT) model
- BERT is a state-of-the-art NLP model that applies bidirectional training of attention mechanism to language modelling tasks.
- The bidirectional flow of training provides a deeper insight into the language context.
- In vanilla form, BERT is composed of an encoder that reads the text input, which may then be integrated with a classification model to predict a task.
- Unlike directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder in BERT reads the entire sequence of words simultaneously.
- This characteristic allows the model to learn the context of a word based on all of its surroundings.

Interpretability model: LIME

- LIME is a local surrogate approach that approximates any black-box machine learning model with a local, interpretable model to explain individual prediction.
- The model specifies the importance of each feature to an individual prediction.
- The model works by tweaking the inputs slightly and observing the changes in prediction.
- The tweaked data points are weighed as a function of their proximity to the original data points, then fitting a surrogate model such as linear regression on the dataset with variations using those sample weights.
- Each original data point can then be explained with the newly trained explanation model.
- The learned model generates a local prediction model while it may or may not provide a precise global approximation.
- Since LIME models treat the machine learning models as a black box, these are model agnostic.

Error Analysis

- ML models often face evaluation challenges regarding performance, accuracy, and reliability.
- In practice, there might be a possibility that the model accuracy may not be uniform across subgroups of data and that input conditions might exist for which the model fails.
- We analyse the results obtained from the experiments and draw meaningful conclusions from the results obtained. To achieve this, we perform an error analysis to evaluate the performance of the classification and the explanation model.

Experimental Setup

- We have used the NLTK library for the removal of stopwords.
- The preprocessed data is then split into a training set, a test set and a validation set in a standard ratio of 70:20:10.
- For the supervised learning classification, we use a state-of-the-art BERT model.
- For the explainability paradigm, we use a LIME model.
- We perform all our experiments using the Python language.

Results

- We run the classification model for four epochs as the performance stabilises after the four epochs. The result presents the precision, recall, F1 and accuracy scores rounded to the third decimal digit.

Epoch	Precision	Recall	Accuracy	F1 Score
1	0.819	0.824	0.819	0.820
2	0.818	0.817	0.815	0.817
3	0.824	0.826	0.823	0.826
4	0.832	0.814	0.826	0.828

Results

- Then we add the LIME architecture to the pipeline and evaluate the LIME results.

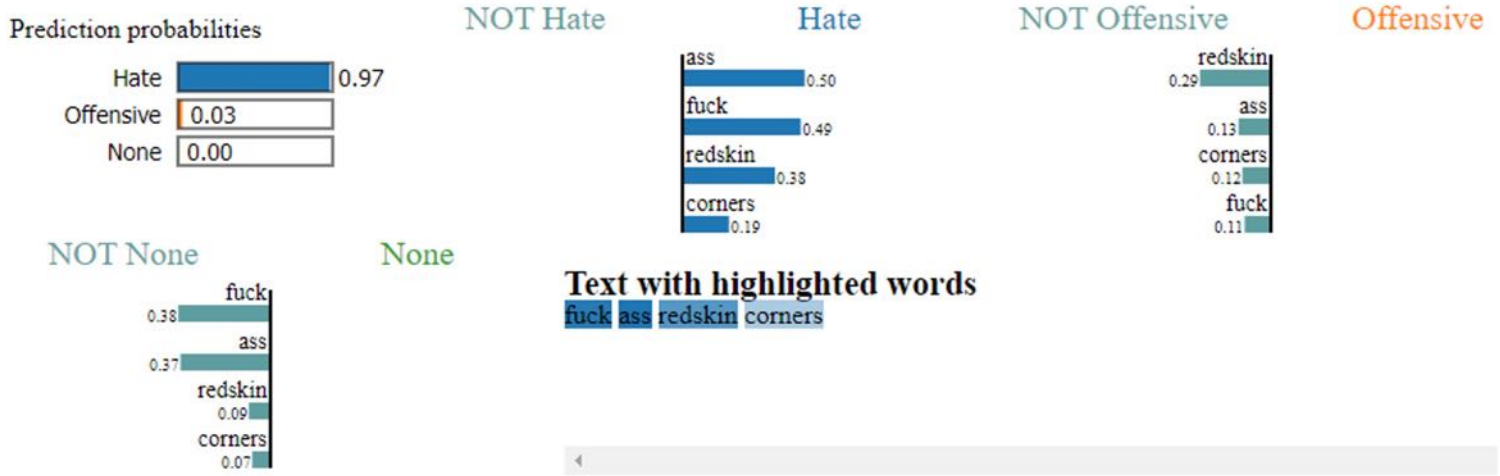


Figure: An Example Tweet and Results

Result and Model Evaluation

- To evaluate errors in the model, we start by observing 150 tweets, taking 50 random tweets from each class.
- Overall, we identified a 21% error rate of our 150 tweet texts, where 5% were predicted as false positives, and 16% were predicted as false negatives, where false negatives in our case would be an incorrect classification of tweets.
- In the observed number of tweets, a mere 0.6% tweets had an exactly equal probability of 'hate' and 'offensive'.

None	2	1	42
Offensive	9	37	1
Hate	39	12	7
Predicted / True	**Hate**	**Offensive**	**None**

- In diagnosing the predicted tweet texts and their classes, we identified a few words that always caused the results to fall into a particular category. We also observed that certain words had a higher frequency of occurrence in each class.

Result and Model Evaluation

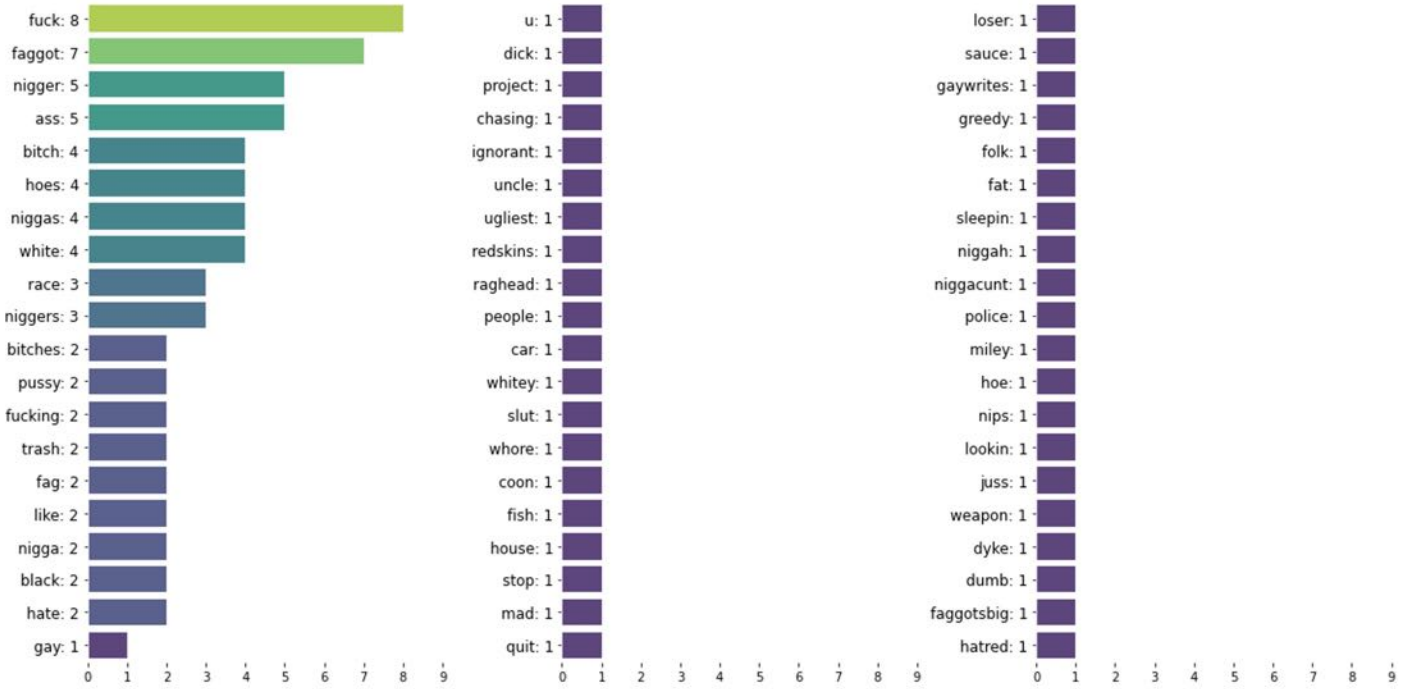


Figure: Word Frequency Class Hate

Result and Model Evaluation

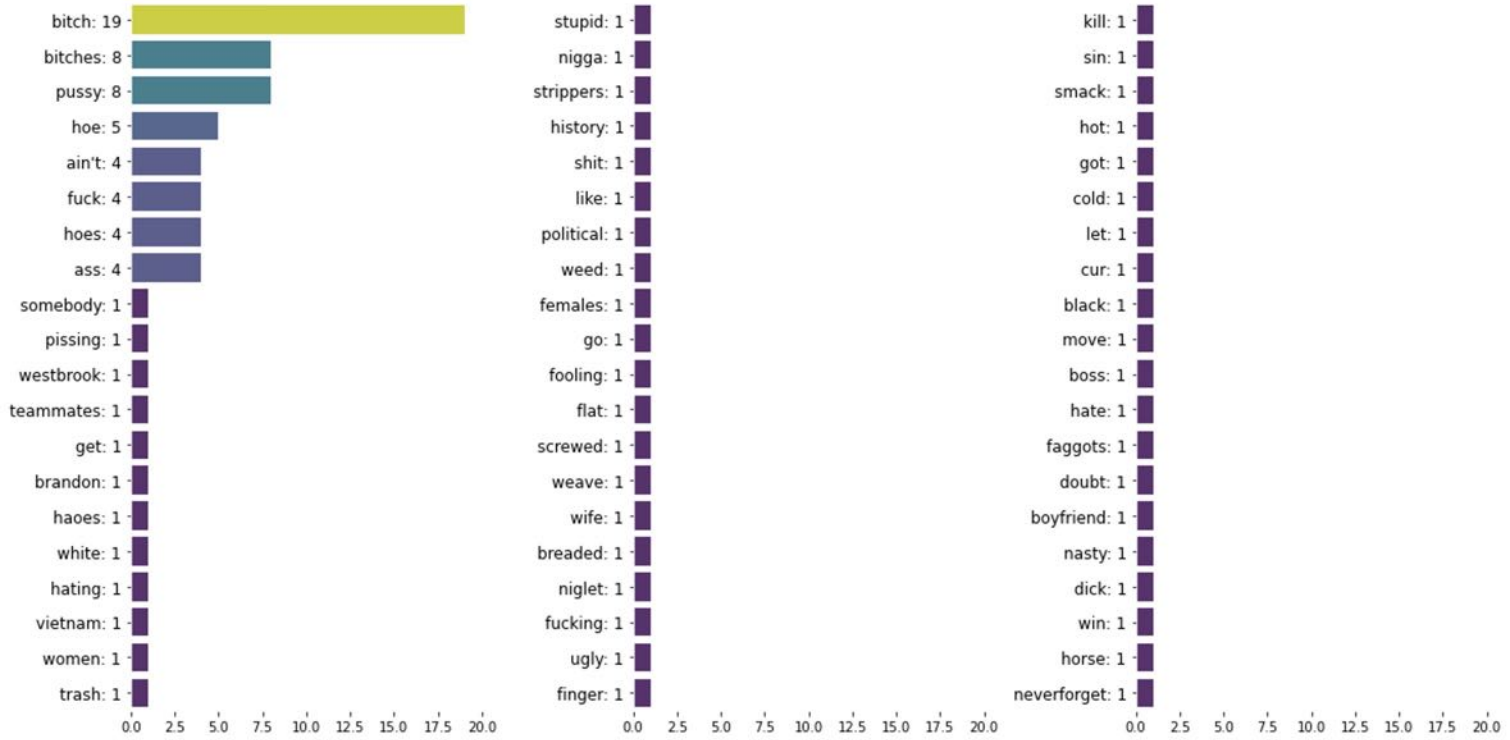


Figure: Word Frequency Class Offensive

Result and Model Evaluation

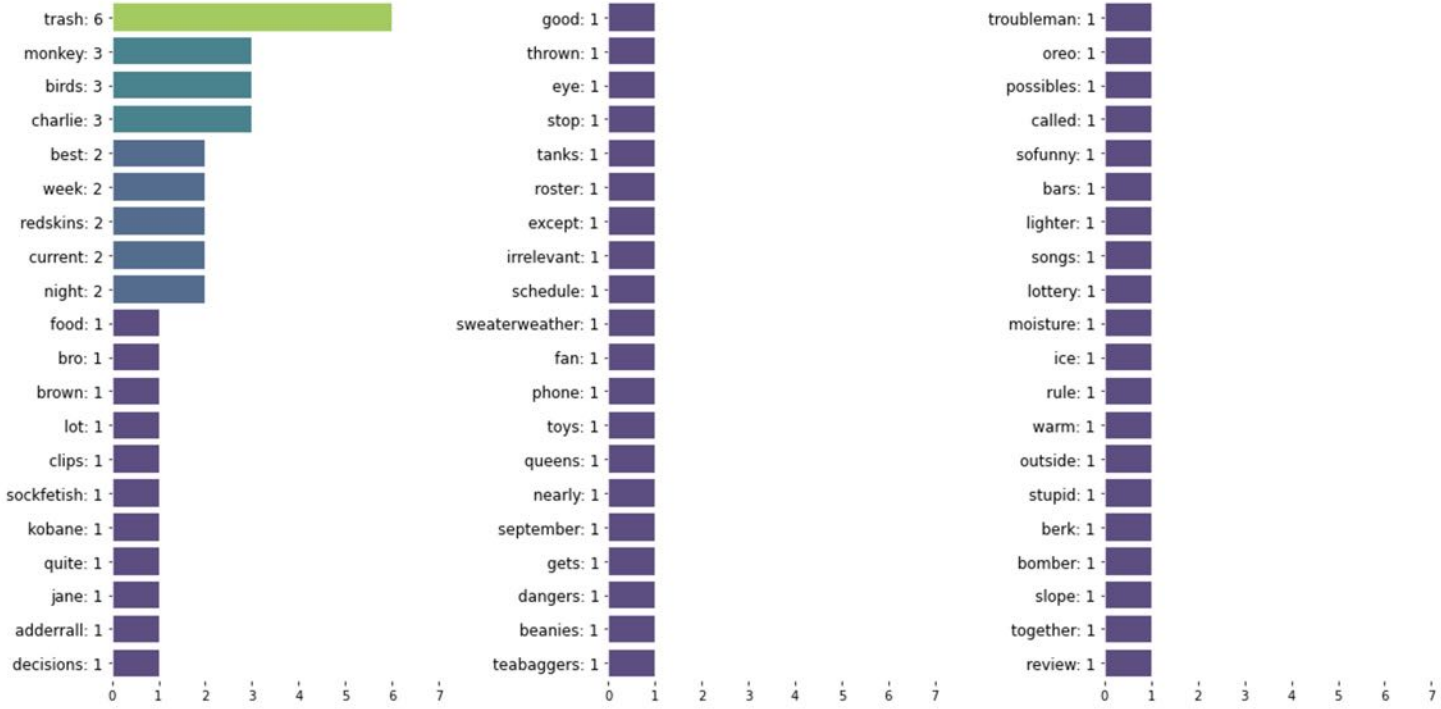


Figure: Word Frequency Class None

Result and Model Evaluation

- While evaluating the results, we observed that while there may be a feature overlap among classes, the model decision was the effect of a combined factor of hate words, nouns, and pronouns.
- For example, if the features were directly using the words like 'you are + hate word (from the shown hate word list, for example)', then the model decision was in favour of the class Hate.
- At the same time, the tweet text was classified as offensive if there were an indirect reference or no direct pointing of objects.
- Additionally, if the tweet used an offensive word but discussed abstract universal concepts, then the tweet was classified as 'offensive' as well.
- If the tweet had no offensive or hate words, then the model classified them as 'none'. For instance, the words like "f**k", "b***h", "black", are frequently occurring in Hate, None, as well as Offensive classes. Here the model checks other combinatorial words and then decides whether a tweet is Hate or Offensive.
- So, if the word directly refers to a Noun, Pronoun, for example, "I'd f**k a dog before I f**k you fish black p***y", then the tweet was classified as Hate, while "aye yo black car is superior" was classified as None.
- Another conclusion from the analysis was that tweets were almost always categorised as Hate when they were racist (use of stem words such as "black", "n***a", "f****t", "white", etc.), while they were almost always classified as Offensive when they were sexist (use of stem words such as "cunt", "b***h", "hoe", "p*****g wife", "p***y", etc.).

Conclusion & Future Work

- In this work, we shed light on the rising effects of hate speech on social media and the dangers that they pose due to various factors.
- To solve this issue, we propose a methodology to detect hate speech on social media platforms and provide an explanation for the same using feature vectors. We have worked on the Twitter dataset for experimental purposes.
- This work opens the prospects for numerous future works, such as enriching the architecture with rule-based learnings using named entity recognition (NER) in association with relational features.
- The architecture can also be extended to various dimensions of data, for example, using image data, spatial relations in textual or image data, or both. Moreover, to evaluate how human users evaluate the model, a survey can be conducted to evaluate the prediction outcomes or the explanations.

References

- Adadi, A.; Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access 6/, pp. 52138–52160, 2018.
- Brown, A.: What is so special about online (as compared to offline) hate speech? Ethnicities 18/3, pp. 297–326, 2018.
- Christoph, M.: Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2019, 2020.
- Davidson, T.; Warmsley, D.; Macy, M.; Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media. Vol. 11. 1, pp. 512–515, 2017.
- Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805/, 2018.
- Ribeiro, M. T.; Singh, S.; Guestrin, C.: Why should i trust you?Ëxplaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Pp. 1135– 1144, 2016.
- Ribeiro, M. T.; Singh, S.; Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 32. 1, 2018.
- Lundberg, S. M.; Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems 30/, 2017.
- Whillock, R. K.; Slayden, D.: Hate speech. ERIC, 1995.

Thank you!