Title: Multi-Class Classification for Hate Speech Detection in Social Media

1. Introduction

The surge of online hate speech on social media has raised significant concerns, manifesting in real-world instances of violence and discrimination based on attributes like religion, ethnicity, gender, and sexual orientation (Chetty et al., 2018). As social media platforms rapidly expand, the pervasiveness of online hate speech has become a critical challenge, prompting global efforts from governments and social media entities to control its dissemination and mitigate adverse effects (Laub et al., 2019). This urgent issue has spurred the research community to actively explore advanced methodologies, particularly leveraging Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms, to automatically identify and combat the proliferation of online hate speech.

Recognizing the detrimental impact of hate speech on user experiences and community well-being, major social media companies have implemented moderation policies. However, existing mechanisms, such as keyword filters, struggle to address nuanced expressions of hate, allowing harmful content to persist (Gao et al., 2017). Additionally, crowd-sourcing methods involving human moderators and user reporting face scalability challenges, leading to delays in detecting and removing hateful posts (Chen et al., 2019). This underscores the perceived importance of devising solutions that are both effective and scalable in addressing the issue of hate speech.

This study centers around a dataset collected by Davidson et al., (2017), comprising 24,783 tweets from the Twitter API. Human labelers categorized these tweets into three distinct groups: 0 for Hate speech, 1 for Offensive language only (not hate speech), and 2 for Neither. Serving as a valuable resource, this dataset is instrumental for training and evaluating hate speech detection models.

This research's main objective is to enhance the effectiveness of automatic hate speech detection systems by leveraging this foundational dataset. Our study focuses on developing and evaluating a multi-class classification model to discern between different categories of speech, contributing to ongoing efforts to create a safer and more inclusive digital environment.

2. Literature review

Hate speech detection has become a critical area of research due to the increasing prevalence of online hate. This literature review aims to provide a comprehensive overview of existing machine learning classifiers, their methodologies, and the challenges they face.

In the realm of hate speech detection, machine learning classifiers play a pivotal role, with content preprocessing and feature selection standing out as crucial components. Past research, exemplified by Neuman et al., (2013) underscores the significance of techniques like stemming and metaphor processing in enhancing the extraction of hate-indicative features. Classical machine learning approaches, including Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression, have been widely employed for text categorization Pedregosa et al., (2011). Davidson et al., (2017) proposed a feature-based model incorporating distributional TF-IDF features and linguistic features using SVM, emphasizing the importance of diverse feature sets.

In the pursuit of refining hate speech detection, recent studies have introduced innovative models. Zimmerman et al., (2018) proposed a neural ensemble approach, combining multiple convolutional neural networks, showcasing the exploration of ensemble techniques. Additionally, the multi-view SVM model, as outlined in the text (Zhao et al., 2017), leverages different feature types, underscoring the importance of capturing diverse aspects of hate speech. However, challenges and critiques loom over the effectiveness of existing models. Cross-dataset performance, as evidenced by various experiments, consistently reveals a significant drop in model performance. This observation raises questions about the generalization capabilities of state-of-the-art models, including recurrent neural networks (Davidson et al., 2017), CNN-GRU models (Zhang et al,. 2018), and LSTM networks (Yin et al., 2021). BERT and its variants have been established as the new state-of-the-art since its introduction (Devlin et al., 2019). Despite their success, challenges in generalization emerge in cross-dataset experiments, with noticeable drops in macro-averaged F1 scores (Gröndahl et al., 2018; Arango et al., 2022). Recent studies propose solutions, including masked language modeling pre-training on abusive corpora (Caselli et al., 2021) and integrating features from hate speech lexicons (Koufakou et al., 2020), aiming to address BERT's limitations in handling diverse datasets.

Moreover, the literature highlights a pervasive issue of overestimating model performance, emphasizing the need for caution when extrapolating results from within-dataset evaluations to real-world scenarios. This trend aligns with findings by Gröndahl et al. (2018) and Arango et al., (2022), suggesting that model evaluations on a particular dataset may not realistically represent the distribution of unseen data.

This literature review synthesizes evidence from recent studies and models, providing a nuanced understanding of the current landscape of hate speech detection. The challenges identified underscore the need for continued research and innovation to develop models that generalize well across datasets and address real-world complexities in online hate speech detection

3. Problem Statement

Online hate speech plays a large role in the occurrence of violence and hate crimes based on specific characteristics such as religion, ethnicity, gender or sexual orientation (Chetty et al., 2018). Despite arguments that the basic human right of free speech should be upheld, governments around the world have enacted laws that require social media companies to step in and stop the spread of hate speech (Davidson et al., 2017). Automatic detection using NLP techniques combined with ML algorithms is an important task and is seen as crucial to stopping the spread of hate speech (Yin & Zubiaga, 2021)

4. Objectives

a. To conduct Exploratory Data Analysis (EDA) to increase understanding of the data and determine the appropriate preprocessing steps.
b. To apply supervised and deep learning algorithms in developing a multi-class classification model to classify a tweet as containing hate speech, offensive language only or neither.
c. To evaluate and compare the performance of the selected models.

5. Methodology

The dataset used in this study contains 24,783 tweets labelled as one of three classes, '0' - hate speech, '1' - offensive language but not hate speech, and '2' – neither and was made publicly available by Davidson et al., (2017). The dataset was created by first forming a hate speech lexicon then searching for tweets containing these words via the Twitter API. The entire timelines of all users returned in the search was then extracted to form a corpus of 85 million tweets. The final analysis dataset was compiled by taking a random sample of 24,783 tweets from the corpus then having them manually labelled by humans.

The analysis started with exploratory data analysis conducted using word clouds to visualize frequently used words in different categories and create a better understanding of the dataset as well as determine the preprocessing steps required.

In the preprocessing step, the tweets were cleaned and prepared for analysis by replacing URLs and mentions with placeholder terms "URLHERE" and "MENTIONHERE" respectively. The text was then lowercased, emojis were converted to text, punctuation was removed, and excess whitespace was normalized. Finally, the tweet was tokenized using nltk's TweetTokenizer. The dataset was then split into a training and testing set (80:20). TFIDF vectors of the tokens were then created as inputs for the modelling stage for the TFIDF models.

In developing the hate speech detection model, the study applied several machine learning algorithms for hate speech detection including Random Forest, Naive Bayes, Support Vector Machine (SVM) and Bidirectional Encoder Representations from Transformers (BERT). Due to the imbalanced nature of the dataset, Synthetic Minority Oversampling (SMOTE) was applied to the TFIDF models while Random Oversampling (ROS) was applied to the BERT model. ROS was chosen over SMOTE for the BERT model to reduce the complexity of the model and thus training time for the model. Oversampling was only applied to the training sets to avoid data leakage during model training.

Scikit-learn implementations with default hyperparameters were used for NB, RF and SVM models (Bi-Min Hsu (2020), Pedregosa et al., (2011)) while the Hugging Face implementation of the BERT (Jacob et.al, 2019) model in the Pytorch framework was used. The model was trained for 10 epochs with a learning rate of 1e-6, batch size of 8, optimized with AdamW optimizer function and cross entropy loss as the loss function.

Finally, the results of these models are compared using evaluation matrix namely accuracy, F1-score, precision, recall and ROC curve (AUC). Models were trained on the training data only then tested on the testing data to prevent data leakage.

6. Results & Discussion

a. Exploratory Data Analysis on Hate Speech Based on Twitter

This section provides an exploratory data analysis (EDA) conducted on a Twitter dataset to investigate patterns associated with hate speech. The analysis involved data exploratory, visualization, and initial steps for word embedding-based hate speech detection. The dataset used consist of 24,783 unique tweets crowdsourced from

Twitter which focused on hate speech and offensive language. This dataset is labelled based on classes label (0: hate speech, 1: offensive language, 2: neither), which were manually coded by coded by CrowdFlower (CF) workers (Davidson et.al, 2017). No missing data is found in the dataset. Table 1 below shows the head dataset consisting of hate speech classes.

Index	Count	hate_ speech	offensive_ language	neither	class	Tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya 

Table 1: Data frame illustrates the first 5 rows of data.

In order to further explore the length of text in accordance to classes, a new column was inserted to calculate the text length for each tweet. From Figure 1 and 2 below show that hate speech (class 0) tweets tend to be slightly shorter than offensive language (class 1) or neutral (neither, class 2) tweets.

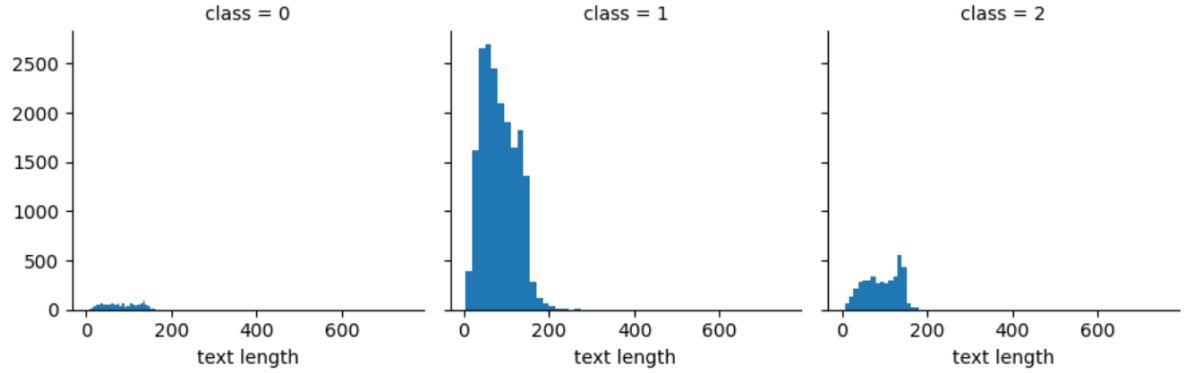

Figure 1: Histogram illustrates that the text length for class 0 is shorter than class 1 and 2

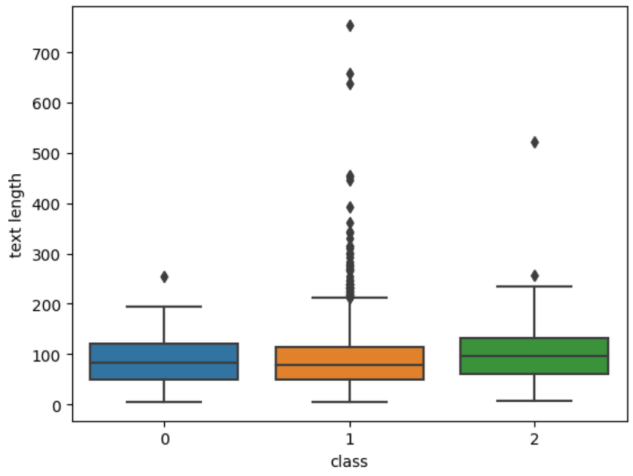

Figure 2: Boxplot illustrates that the text length for class 1 is more than class 0 and 2

Before we can further analyse the data, the dataset was thoroughly cleaned to remove irrelevant elements and normalize text. This included handling extra spaces, mentions, links, punctuation, numbers, capitalization, stop words, and stemming. The cleaned data is as shown in Table 2 below.

Table 2: Table shows data set before and after cleaning process

Tweets before cleaning	Tweets after cleaning
0 !!! RT @mayasolovely: As a woman you shouldn't...	0 woman complain clean hous amp man alway take t...
1 !!!!! RT @mleew17: boy dats cold...tyga dwn ba...	1 boy dat cold tyga dwn bad cuffin dat hoe st place
2 !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	2 dawg ever fuck bitch start cri confus shit
3 !!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	3 look like tranni
4 !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	4 shit hear might true might faker bitch told ya
5 !!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just...	5 shit blow claim faith somebodi still fuck hoe
6 !!!!!!"@__BrighterDays: I can not just sit up ...	6 sit hate anoth bitch got much shit go
7 !!!!“@selfiequeenbri: cause I'm tired of...	7 caus tire big bitch come us skinni girl
8 " & you might not get ya bitch back & ...	8 amp might get ya bitch back amp that
9 " @rhythmixx :hobbies include: fighting Maria...	9 hobbi includ fight mariam bitch

In further exploration of the dataset, word cloud is applied to reveal the distinct word patterns across classes. Based on Figure 4 to 5 below, it shows that hate speech tweets frequently feature words expressing aggression or negativity which are considered as hate speech and defensive. Whereas Figure 6 shows words which are commonly used and not considered as offensive or reflecting negativity.

Figure 3: Word cloud illustrates the words that are most commonly used in the twitter dataset

Figure 4: Word cloud illustrates the words that are most commonly used for hatred speech

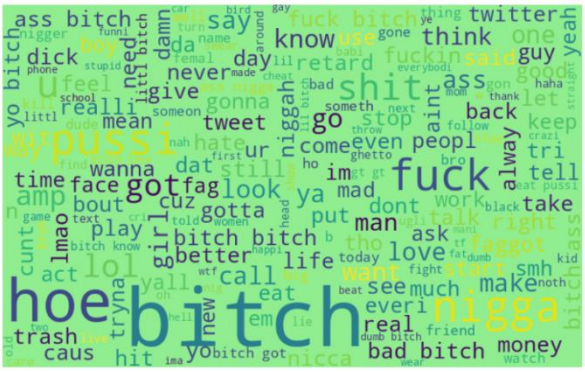

Figure 5: Word cloud illustrates the words that are most commonly used for offensive speech

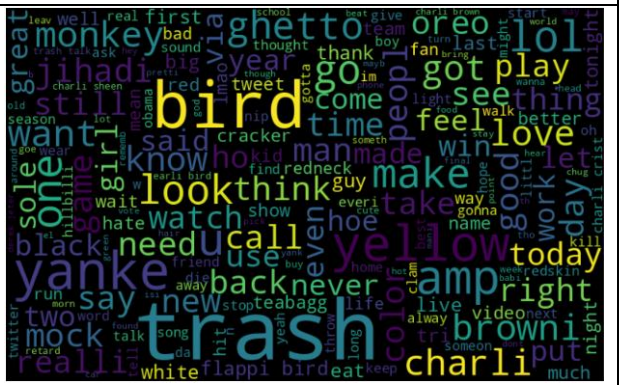

Figure 6: Word cloud illustrates the words that are most commonly used for neutral speech

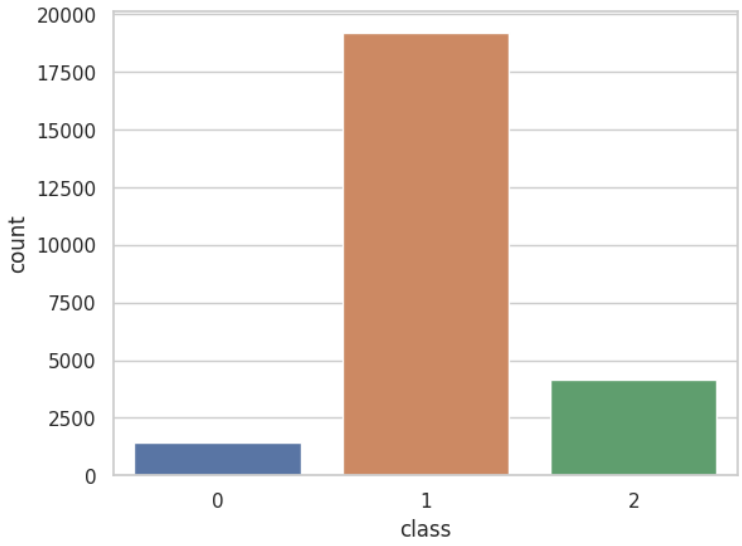

Figure 7: Count plot showing target variable distribution

b. Multi-Class Classification Model

Table 3: Evaluation metrics

Model	F1-Score (%)	Accuracy (%)	Precision (%)	Recall (%)	ROC AUC (%)
TFIDF-NB	50.92	68.93	51.20	54.22	66.85
TFIDF-RF	73.00	88.70	72.72	73.69	92.02
TFIDF-SVM	67.86	84.57	65.86	71.18	90.10
BERT	77.62	91.76	77.55	77.73	93.94
BERT-ROS	76.30	89.57	74.45	78.66	92.27

As seen in Figure 7, the target variable of the dataset is highly imbalanced, with class 0 (hate speech) making up just 5.77% of the total number of tweets. In such a situation, an accuracy score would not provide a proper representation of model performance as a classifier that is only good at predicting class 1 but predicted class 0 and 2 wrongly would still have a high accuracy score. Similarly, the ROC AUC score does not a comprehensive representation of a model's performance as it is heavily affected by true negatives. Hence, F1-score was chosen as the metric by which we judge and compare the performance of the various models. F1-score is defined as the harmonic mean of the Precision and Recall scores and provides a balanced representation of the model's ability to classify both the majority and minority classes.

In this context, a high precision score means that when the model predicts a tweet as hate speech, it is likely to be correct. However, if precision is high but recall is low, the model is likely missing a lot of actual hate speech instances, resulting in a high number of false negatives. Conversely, a high recall score indicates that the model performs well when identifying hate speech instances, but if recall is high and precision is low, the model may be over-predicting hate speech, including many instances that are not actually hate speech (false positives). In the task of hate speech identification, the balance between both scenarios that F1-score provides would be appropriate because a social media platform would ideally like to keep as much hate speech as possible off the platform while not turning users away from platform with overzealous policing of user content.

When comparing TFIDF based models, the RF model was the best performer with an F1-score of 73.00% and accuracy of 88.70% while both RF and SVM models outperformed the baseline NB model. The NB model likely performed the worst due to the complexity of the classification task and the underlying assumption of the NB model where features are assumed to be independent of one another. This would affect the performance of the model because in natural language, words used together carry semantic meaning and are not independent of one another. The RF model likely performed the best due to being an ensemble learning method that builds multiple decision trees and aggregates them to output the final prediction. Such an approach has the advantage of being more robust to overfitting while also being able to learn the complex features in natural language. Lastly, the poorer performance of the SVM model could be attributed to lack of hyperparameter tuning to select the kernel function as well as the regularization parameter. SVM model performance could also be further improved by scaling the input vectors.

The Bidirectional Encoder Representations from Transformers (BERT) model was also applied to this task and produced the highest F1-score of 77.62%. This is expected because BERT was pre-trained on an extremely large corpus of 3.3 billion words from Wikipedia and Google's BookCorpus in an unsupervised manner and is specialized in Natural Language Processing (NLP) tasks. The model applied here 'bert-base-uncased' learns 110 million parameters enabling it to learn context, nuances and interdependencies of natural language (Hugging Face, 2024). When applied to tweets, the BERT model generates embeddings from entire tweets, enabling it to capture context and the order of words which are not captured in TFIDF vectorization. The BERT-ROS model with random oversampling applied performed slightly worse than the default BERT model. This could be due to the BERT-ROS model experiencing a slightly higher degree of overfitting when trained on the duplicated minority instances created when applying the ROS method.

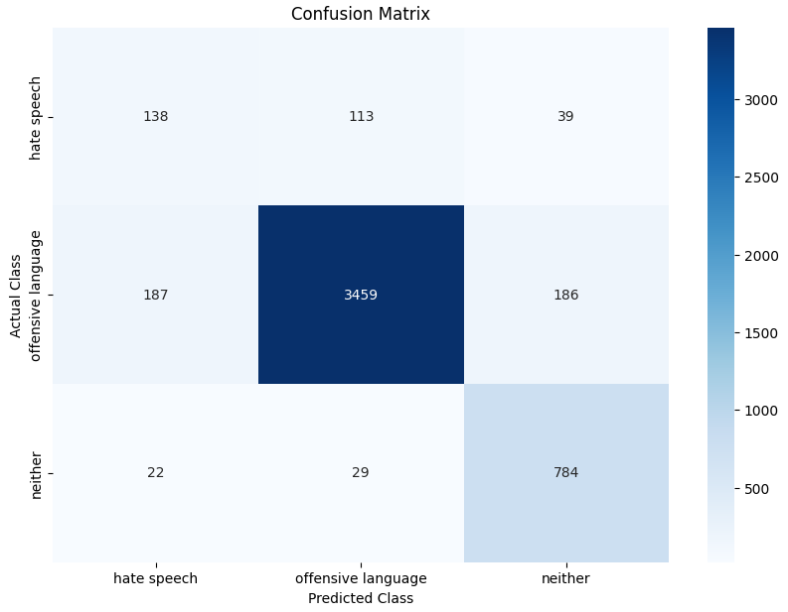

Figure 8: Confusion matrix plot on test data (BERT)

Figure 8 shows the confusion matrix plot produced by the BERT model on the test data. From the chart, we see that the model struggles the most at correctly classifying class 0 (hate speech) with a relatively high number of false negatives (187) and false positives (113 and 39) while performing better when classifying class 1 and 2. This is due to the imbalanced nature of the training dataset. It can also be seen that the model struggles more at differentiating between hate speech and offensive language which could be due to the complex nature of the task, where even manual human classification would face issues of bias and subjectivity. The following are two examples of tweets that the model classified as hate speech but were actually offensive language *"it gets me mad when a mexicans go against mexico like wtf support your team b**ch"* and *"shut the fuck up you f**king c**t !"*. Confusion of the model in the former is understandable as the tweet could be misconstrued as hate speech against a particular country while more context such as preceding tweets would be required to classify the second example as hate speech. Similarly, the tweet *"that was my f**got rant for the night"* was labelled as hate speech but predicted as offensive language by the model, but it would be hard to determine if it was meant in a derogatory manner without additional context.

The performance of the various models could potentially be improved by using a more complex oversampling technique that is targeted for multi-class imbalanced scenarios such as Mahalanobis Distance-based Oversampling

(MDO) (Abdi & Hashemi, 2016) or ADASYN (He et al., 2008). More advanced large language models such as Llama 2 by Meta or GPT-4 by OpenAI could also be applied to increase the effectiveness of the model, both of which have been trained on larger datasets and are built with neural networks that learn a much higher number of parameters. Alternatively, a more balanced dataset could be created by taking multiple random samples, labelling them then taking a stratified sample.

7. Conclusion

Hate speech is currently one of the critical social issues due to the expansion of social media in society. This project underscores the urge to cease this hate speech issue. To cut down the negative impacts caused by hate speech, it is important to have functional model in detecting hate speech as the first step in stopping this society sickness.

In this project, EDA had been conducted on the Twitter dataset to identify the suitable preprocessing step, and necessary preprocessing is conducted to clean the dataset. In the modelling part, the selected dataset is being trained and tested based on supervised learning (TFIDF) and deep learning (BERT). In overall, BERT based model is performing better as compared to TFIDF which the evaluation had been further discussed under the sub-topic Results and Discussion of this report.

One of the challenges encountered in this project is the detection of imbalance data. To resolve this problem, oversampling technique is applied to balance the class distribution after preprocessing. However, different models required different sampling techniques. Not to forget, preprocessing is slightly different whereas normalizing whitespace and removing punctuation is not required for BERT based model. Hence, distinct approach required on the data cleaning and oversampling.

The implications of our findings are relatively great. The performance of the BERT-based model suggests a more promising method for developing more effective hate speech detection tools, which are crucial for curbing the negative impacts of hate speech in society. Furthermore, our study highlights the importance of tailoring preprocessing methods to specific models, a consideration that can vastly improve model performance.

However, we acknowledge the limitations of our study, particularly in the scope of the dataset and the complexity of hate speech nuances that may not be fully captured. Future research should aim to include more diverse datasets and explore the integration of other advanced machine learning techniques. By continuing to refine our approach to hate speech detection, we can make significant strides towards creating a more respectful and inclusive online environment.

Link to Video Presentation: https://youtu.be/uMRSwok8fQY

Links to Analysis:

1. EDA: https://colab.research.google.com/drive/1yBq0nZA5EhrhOaciu4EEoxNKtj_nRuja#scrollTo=F3QHMlPnDviL
2. TDIDF_Models: https://colab.research.google.com/drive/148oaFPZ2nvUl72_C-YI50jeictLtnSBI
3. BERT: https://colab.research.google.com/drive/13Z6FvATJB5V3J4Gu99vmIgOZEc6xHUvS
4. ROS_BERT: https://colab.research.google.com/drive/131qkF3e_zV6_onAe3CURSWhym8gVjYVA#scrollTo=NC8XbaTDHB1k

References:

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. Aggression and violent behavior, 40, 108-118, https://doi.org/10.1016/j.avb.2018.05.003

Laub, Z. (2019). Hate speech on social media: Global comparisons. Council on foreign relations, 7, https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. arXiv preprint arXiv:1710.07394, https://doi.org/10.48550/arXiv.1710.07394

Chen, H., McKeever, S., & Delany, S. J. (2019, July). The use of deep learning distributed representations in the identification of abusive text. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, pp. 125-133), https://doi.org/10.1609/icwsm.v13i01.3215

Neuman Y, Assaf D, Cohen Y, Last M, Argamon S, Howard N, et al. Metaphor Identification in Large Texts Corpora. PLoS ONE. 2013; 8(4), https://doi.org/10.1371/journal.pone.0062343

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. JMLR. 2011; 12:2825–2830, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM. 2017, https://doi.org/10.1609/icwsm.v11i1.14955

Zimmerman S, Kruschwitz U, Fox C. Improving Hate Speech Detection with Deep Learning Ensembles. In: LREC; 2018, https://aclanthology.org/L18-1404.pdf

Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: Recent progress and new challenges. Information Fusion. 2017; https://doi.org/10.1016/j.inffus.2017.02.007

Zhang Z, Robinson D, Tepper J. Detecting hate speech on twitter using a convolution-gru based deep neural network. In: European Semantic Web Conference. Springer; 2018. p. 745–760, https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7, e598, https://doi.org/10.7717/peerj-cs.598

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM workshop on artificial intelligence and security (pp. 2-12).https://doi.org/10.1145/3270101.3270103

Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). Information Systems, 105, 101584.https://doi.org/10.1145/3331184.3331262

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805., https://doi.org/10.48550/arXiv.1810.04805

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. arXiv preprint arXiv:2010.12472.,https://doi.org/10.48550/arXiv.2010.12472

Koufakou, A., Pamungkas, E. W., Basile, V., & Patti, V. (2020). HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In Proceedings of the fourth workshop on online abuse and harms (pp. 34-43). Association for Computational Linguistics., https://iris.unito.it/handle/2318/1769037

Bi-Min Hsu, Comparison of Supervised Classification Models on Textual Data, Mathematics 2020, 8(5), 851; https://doi.org/10.3390/math8050851, 24 May 2020

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs.CL], or arXiv:1810.04805v2 [cs.CL], https://doi.org/10.48550/arXiv.1810.04805, 2019, https://huggingface.co/docs/transformers/model_doc/bert

BERT 101 - state of the art NLP model explained. (n.d.). https://huggingface.co/blog/bert-101#1-what-is-bert-used-for

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)