CE807 – Assignment 1 - Interim Practical Text Analytics and Report

Student ID: 2214000
ms22749@essex.ac.uk

Abstract

The classification or detection of offensive speech or anything related to it has been quite a common topic in the Natural Language Processing (NLP) community and researchers have taken this aspect as a serious problem and are coming up with various techniques to solve the problem. The objective of this report is to present a comprehensive exploration of the offensive speech classification and detection landscape by summarising a diverse set of papers by providing detailed analysis on the different approaches used by the authors and their alternative considerations, reflecting on the advantages and disadvantages of multiple approaches, and comparing state-of-the-art methods and limitations. Additionally, the report deep dives into the data characterisation, data split, and exploration of the identified offensive speech classification and detection dataset. Also, alternative approaches and the key lessons learnt is highlighted in the concluding section.

1. Materials

Please use the below links to access the resources related to this assignment.

- Code
- Google Drive Folder containing code, data, and saved outputs
- Zoom Presentation
- Option 1

2. Paper Summary

2.1. Appropriate coverage and contextualisation

In the paper (Sandaruwan et al., 2019), the authors explain why hate speech detection in Sinhala language is crucial to prevent the negative consequences on social media. The authors proposed lexicon-based and machine-learning-based approaches by using 2000 comments corpus and annotated them based on agreed definitions. Annotation is essential for categorising hate speech data. Then they pre-processed the text by removing non-Sinhala characters since the main focus was on the dialect, removing stop words since they do not add to the meaning and shallow stemming (Welgama, 2011) for creating a stemming dictionary since there is no stemmer for Sinhala. Unlike English, Sinhala is a low-resource language and there are no publically available dictionaries for Sinhala hate or offensive words, the authors decided to create their own through two approaches: dictionary-based and corpus-based. In the dictionary-based approach, the authors used a Google bad word list, translated the list using an online English-Sinhala dictionary and used two annotators to annotate 1128 translated words based on few agreements. Annotating helps to recognise the related words and categorize them to a which label best represents them. For the corpus-based approach, the authors used a 3000-comment corpus containing 1000 comments and classified them into hate and offensive categories. A seed word set from an online source was taken and their variations were identified by adding suffixes to the seed words, which signifies their adaptation in different contexts. Feature extraction from the pre-processed text was done using various techniques such as bag-of-words(BOW) because it maintains the frequency of occurrence of each word, word n-grams (unigram, bigram, trigram or combination) since it captures structure of the text unlike BOW which disregards the grammar, character n-grams for detecting patterns and Word skip-gram features where words are parsed to find its nearest neighbour in the text. The study uses two feature vectorization methods from Scikit-Learn's CountVectorizer and TfidfTransformer, for converting natural language text into numbers because statistical Machine Learning (ML) methods require input to be in numeric form.The ML approach leverages supervised ML algorithms such as Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Random Forest Decision Tree (RFDT) for classification. Five experiments were conducted with the generated lexicons and feature extraction methods along with classifiers and 300 comments balanced corpus, and various performance metrics such as accuracy, precision, recall, and F1-score were considered for evaluation of the models. Overall, the study identifies and shows that corpus-based lexicons are better than dictionary-based lexicons, Character trigram feature is the best among all other features, and the MNB classification model gave 92.33% accuracy and a recall of 0.84. This shows that a commendable effort was made by the authors to address the problem of hate speech in Sinhala.

In the paper (Abro et al., 2020), the authors highlight that in recent studies, attempts have been made to automatically detect hate speech on variety of datasets but there is no comparative study which tells which feature engineering or ML is the best . To achieve this goal, the authors used a standard publicly available tweets dataset shared by CrowdFlower and compared three feature engineering techniques and eight machine learning classifiers. Text pre-processing was done by removing noisy and irrelavant features lowercasing the tweets, removing URLs, hashtags etc.Tokenization and stemming was also performed. Three feature engineering techniques such as n-gram with TFIDF, Word2vec and Doc2vec were used. This is a crucial step in cleaning the data before training the model. Creating ngrams and representing them in a numeric form helps in identifying patterns in hate speech texts. They split the dataset in 80:20 ratio. They used eight ML algorithms such as SVM, RF, LR, NB, KNN, DT, AdaBoost, and MLP.They used Precision, Recall, F-Measure and Accuracy as metrics to assess the performance of classifiers. This helps in assessing the effectiveness of different techniques, models, or combinations that will work best for hate speech detection.The results show that bigram features with TFIDF representation along with SVM achieved the best performance with an accuracy of 79%. Overall, the authors have provided a baseline for future work through this study.

In the paper (Saini et al., 2020), the authors highlight how social media is being misused by a group of people to spread hatred and nasty content. The authors propose a methodology that is a combination of a topic modelling technique called Latent Dirichlet Allocation (LDA) and an unsupervised ML technique, a type of Artificial Neural network (ANN) called Self-Organising Maps (SOM) to detect hate speech on social media. LDA was used to extract topics from the dataset, while SOM was used to map the topics on a low-dimensional display. By applying LDA, important and dominant topics can be identified, while SOM helps in the visualisation of values in the data and their relationships.

The methodology involved steps such as raw or unclean data collection (Davidson et al., 2019), data preprocessing with manually implemented steps to clean the dataset, POS tagging: helps to identify different POS tags in the dataset, document-word matrix(BOW), application of LDA, application of SOM and application of K-means clustering: helps in identifying the distribution and patterns in the dataset. Ten topics with low perplexity and higher negative log-likelihood score was produced using the proposed LDA model. K means clustering gave 16% variance. Furthermore, the authors suggest to adapt their methodologies to non-English words , usage of large corpus, hypertuning and hate speech detection in images in the future.

2.2. Critical Discussion

The paper (Sandaruwan et al., 2019), acknowledges a study by (Dias et al., 2018) and mentions the limitations of the study, such as the usage of unbalanced and small training and testing sets and the narrow focus on racism-based comments. They justify how they considered vast range of characteristics to identify hate speech and used balanced 3000-comment corpus in their own study. Additionally, while acknowledging a work, (Welgama, 2011) the authors emphasise the shallow stemming method proposed in their study for Sinhala. The authors knew the main issue that the algorithm was unable to detect the stem of a particular word if the word's stem was missing from the document itself and still incorporated it into their own study by using the same concept. The authors did not mention what challenges the annotators might face while distinguishing between hate speech and free speech. For example, there might be a word that can be hate speech for someone and not for someone else. There is a lot of context involved, and criteria have to be identified. It will be helpful if they are native speakers from specific regions or have language familiarity. And the accuracy of the classifiers might be affected if the annotations are not done properly. The authors used SVM, MNB and RFDT for classification because in past researches (Malmasi and Zampieri, 2017) (Davidson et al., 2017) these algorithms have provided good results. They did not consider other classifiers and find which works best.

In the study,(Abro et al., 2020), the authors mention that in previous studies, the research has been done on hate speech recognition on other written languages but none of the studies provide a comparative study. Hence, this study can be used as a baseline study in future. The authors also mention about the 'no free lunch theorem' (NFLT)(Ho and Pepyne, 2002) which states that there is no one classifier that performs best on all the datasets and they suggest testing a variety of classifiers to see which one yields the best outcomes. Therefore, they used eight classifiers in their study. Two important limitations of their study are specified by the authors: 1) Inefficiency of real-time predictions 2) Inability to identify the severity of hate speech messages. But the authors do not discuss these limitations or potential sources of bias in the dataset or classification process. Also, the authors omit discussing the challenges of generalising the proposed approach to other languages or social media platforms. Additionally, the research used a publicly available hate speech dataset of 14509 tweets of which 16% belonged to hatespeech class. This shows that the non hateful content is oversized and the dataset is biased.

In this paper,(Saini et al., 2020), the authors mention briefly the application of K-means clustering with Euclidean distance to organise data into clusters based on the dominant topic assigned by LDA. However, there is no discussion on challenges like choosing the right distance, inconsistency is assigned topics, dealing with high dimensional data is mentioned. The paper only mentions that ten topics were outputted by the LDA model but does not explain on what basis these topics were selected or their relevance to hate speech detection. Another critical aspect is that the authors mention that LDA and SOM provide powerful analysis, but they did not consider alternative approaches like LDA and SOM with other supervised techniques and did not mention the limitations of their proposed methodology. The authors considered 2400+ tweets for their study. It is a challenge to build efficient systems with smaller data sets. Information on the performance of unsupervised learning, and more research into the effectiveness of different unsupervised learning algorithms in the identification of hate speech is required.

2.3. Discussion of advantages and disadvantages

In general,one advantage common to all papers is that all three papers address the problem of hate speech detection by exploring different techniques and approaches. While the paper (Sandaruwan et al., 2019) focuses on discussing how Sinhala hate speech is detected, considering the issues and representation of the language, the paper (Abro et al., 2020) compares various techniques and algorithms, and the paper (Saini et al., 2020) introduces a unique methodology combining LDA and SOM. A few disadvantages common among the papers are limited comprehensive evaluations, and representation and the generalizability of the results to other languages and size of the datasets. While the paper (Sandaruwan et al., 2019) analyses various approaches and feature extraction methods but lacks details on experimental setups and does not consider foreign language in the texts like Singhlish,(Abro et al., 2020)compares feature engineering techniques and ML classifiers but lacks discussions on overcoming limitations and faces issues with training data, and the paper (Saini et al., 2020) employs LDA and SOM for topic modelling and visualisation but has limited evaluation metrics and discussions on biases and generalizability.

Some of the advantages of the paper (Sandaruwan et al., 2019) are: comprehensive analysis of lexicon-based and multiple supervised machine learning approaches, creating Sinhala hate or offensive words dictionary, comparison of feature extraction methods like word n-grams and character n-grams, which provide probability of words, given the history of the context(as seen in lecture 6). The disadvantages are: the primary focus is on detecting hate speech in Sinhala, and combined scripts such as Singlish (Sinhala + English) are not considered. And for training the machine learning models, a larger volume of labelled data may be required. The study used a small corpus.

Some of the advantages of the paper (Abro et al., 2020)are that it provides insights into the performance of machine learning models by comparing three feature engineering techniques and eight ML classifiers for hate speech detection, and in comparison to word2vec and doc2vec, the usage of bigram features with TFIDF representation produces better results. This highlights how n-gram can be highly predictive when combined with others. The disadvantages are: lack of discussion on overcoming limitations of feature engineering techniques and machine learning algorithms, poor performance of the MLP classifier due to a lack of training data.

Some of the advantages of the paper (Saini et al., 2020) are: LDA is a widely used topic modelling technique. It can help in uncovering dominant topics and patterns in the data and provide insights into the performance of different methods, such as TFIDF, Naive Bayes, and K-means clustering, for hate speech detection. Effective visualisation and converting complex statistical relationships between high-dimensional data items into pure linear relationships can be obtained by SOM, and organisation of data into meaningful subgroups based on similarity can be achieved by K-means clustering. The disadvantages are: it fails to provide limitations of LDA, such as the need for manually provide the number of topics. No discussion about bias in the dataset used. In k-means disparate initial partitions can result in disparate final clusters.

2.4. Contextualisation and discussion of state of the art

In the paper (Sandaruwan et al., 2019) the authors proposed a lexicon, feature extraction techniques, and machine learning classification models, which contribute to the advancement of the field. The paper suggests future research directions, including the detection of hate speech in other languages. The study also suggests that unsupervised learning techniques and micro-features such as patterns in the speeches can also be considered to identify Sinhala hate speeches. Recent studies (Fernando et al., 2022) have used deep learning models and BOW, Tfidf, Word2Vec, and FastText feature extraction methods and have provided good results.

The paper (Abro et al., 2020), cites previous studies that used different techniques for hate speech detection, such as (Burnap and Williams, 2016), (Gitari et al., 2015), (Tulkens et al., 2016),(Waseem and Hovy, 2016),(Nobata et al., 2016), and (Malmasi and Zampieri, 2017) and compares its findings with related research. The authors acknowledge that the proposed ML model has limitations in terms of real-time prediction accuracy and the ability to identify the severity of hate speech messages. Although the paper does not provide major discussions of the specific state-of-the-art methods, it positions the proposed study as a baseline that can be used to compare future research. However, the methods of hate speech continue to evolve and there may be more approaches that can achieve higher accuracies. In a recent study (Kumar et al., 2023) an intelligent system named 'HateDetector-a recursive system' using LSTM-CNN model was created to monitor and generate alerts on hate speech text.

The paper (Saini et al., 2020) does not provide a comprehensive review of existing state-of-the-art methods and their effectiveness. A more detailed contextualization and comparison with other approaches is needed to fully evaluate the proposed approach. The paper suggests that via parametric tuning, expanding the database to include more offensive and racist comments, and exploring the detection of hate speech in images, the results can be improved. However, in a recent study (Nagar et al., 2023), advancements in topic modeling is shown by introducing a framework called Variational Graph Auto-encoder (VGAE), which can be used as a plug-in to obtain the textual features of the data.

3. Dataset Characterization and Exploration

3.1. Appropriate coverage of 5W-Qs questions

Who produced this language originally?
- The language was originally produced by the authors of the paper (Mandl et al., 2019) which was published in 2019 at the 11th Forum for Information Retrieval Evaluation (FIRE).
Who collected this text data originally?
- The text data was originally collected by the authors for the HASOC track for Hate speech detection, sampled from Twitter and partially from Facebook for Indo-European languages such as English, German, and Hindi. The data collection process involved using hashtags from Twitter tweets and keywords related to offensive content. To increase the variety, the last posts from authors were collected and the annotation of the data was performed by several juniors.
Who has added to it or modified it since?
- The paper does not provide information about subsequent modifications or additions to the dataset.
Are they alright with you using it?
- The permission to use the data indicates that one can make copies for personal or classroom use without paying a fee, provided certain conditions are met. However, it is advised to obtain permission from the author(s) or owner(s) of the work for any other use.
What is in the data?
- The data consists of samples collected from Twitter and Facebook for Hindi, German, and English with tweets classified into two classes (Sub-task A): hateful and offensive (HOF) and non-hateful and offensive (NOT). Sub-task B classifies hate speech and offensive posts into three categories: hate speech, offensive (OFFN), and profane (PRFN). Sub-task C considers the type of offence, with two categories: targeted insult (TIN) and untargeted insult (UNT).
– Topic: Hate speech and offensive online content.
- Genre: Social media posts from Twitter and Facebook.
- Quantity: In total, 7005 posts.
- Organization: Authors of the paper and the organizers of the HASOC track at FIRE 2019.
- Language[s]: German, English, and Hindi.
How much of it is relevant to your specific use? Is it what you need for your work?
- Yes, it is relevant since the task involves hate speech detection on English dataset. Data splitting and exploration are done on this dataset, which is discussed in detail in the next section.
Where was the language data produced?
– Online because the data was produced by sampling Twitter and partially from Facebook for Hindi, German, and English.
Where are the actual files? Can you get to them?
- It is believed to be hosted and made available through GitHub, which contains three categories of datasets: English dataset, Hindi dataset, and German dataset. Researchers and practitioners can access and download the dataset from GitHub.
Why was it produced? Can you trust it?
- The HASOC Dataset was created to facilitate research and development in the field of hate speech and offensive content detection on social media.
- Trustworthiness is generally considered reliable, but may have limitations or biases that need to be considered. If needed, the authors or organisers can be contacted via email.
When was it produced?
- The 1st edition of HASOC in Indo-European Languages was organized by FIRE organizers in 2019 in Kolkata, India.
Was it edited at any point? Do you have an accurate timeline of where this data has been and what's been done to it?
- The data is available on GitHub since 2019. No accurate information of what has been done to the data is provided. But every year the organizers announce call for participation and the interested participants can participate in the Shared Task on the Identification of Offensive Content for Indo-European Languages.
What was it collected for?
- It is inspired by two previous evaluation forums, GermanEval and OffensEval, and aims to develop data and evaluate resources across different languages to stimulate research and find out the quality of hate speech detection technology and specifically collected for the purpose of hate speech detection in German, English, and Hindi languages.

3.2. Data Split Details

For this assignment, I have considered the English dataset. The dataset contains labelled English text samples with columns for 'text_id', 'text', 'task_1', 'task_2', and 'task_3'.The 'text' column represents the text samples, while 'task_1', 'task_2', and 'task_3' denote the labels for different tasks. I have considered only task_1 (subtask A) for the exploration and splitting.The original dataset contains 7k plus records, so to sub-sample the dataset to 4k samples, the sample() function from pandas was used by merging original and test sets. 4k samples were randomly selected for the sub-sample, and file was saved. To split the sub-sampled dataset into train/validation/test sets in the ratio of 70/10/20, stratified sampling was incorporated. The train_test_split() function was called twice, and the resulting train, validation, and test sets were saved as 'train.csv', 'val.csv', and 'test.csv," files respectively. The details of the dataset is shown in Table 1.

Dataset	Total	% NOT	% HOF
Original	5852	61.36	38.63
Train	2800	63.75	36.25
Valid	400	63.75	36.25
Test	800	63.75	36.25

Table 1: Dataset Details

3.3. Data Exploration

Basic and advanced text pre-processing on the 'text' column of the dataset was performed using various techniques, such as converting the text to lowercase, removing URLs, removing punctuation, expanding contractions, sentiment analysis, tokenizing the text, removing stopwords,text representation, lemmatization, stemming, n-gram, word embeddings, topic modelling (LDA) and clustering are performed and inferences are drawn.

4. Summary

4.1. Discussion of alternative approaches

Alternative approaches to paper (Sandaruwan et al., 2019) are using Deep learning models such as Recurrent Neural Networks (RNNs) and transformers helps to capture contextual information in text and process the input sequentially. Also, pre-trained Language models(LM) such as Bidirectional Encoder Representations from Transformers (BERT) or Generative Pretrained Transformer (GPT) can be fed into the LMs for predicting words at each steps and patterns specific to the Sinhala language. They are used for predicting text and have to be trained on large corpus to estimate. Using weighted loss (cross entropy) to improve recall, playing around with probability threshold and all evaluation metrics and volume of labelled data with diverse range of hate speech categories can be increased to tackle hate and offensive linguistic content on social media platforms. During pre-processing, the study could have identified the foreign language in the text and removed it like stopwords. But, this is also a challenge because the words can be cluttered in the text and lead to loss of context.

Alternative approaches to paper (Abro et al., 2020) are using ensemble methods that combine multiple classifiers to predict instead of relying on single classifier like SVM and ADABoost. Focus on the relevant parts of the input can be achieved through Attention mechanism (as seen in Lecture 8) can be used.Character-level n-grams or skip-grams can help to capture additional linguistic information. Different RNN variants like Long Short-Term Memory (LSTM),Gated recurrent Units (GRUs) can be explored since they are more powerful than NNs and help in sequential processing of inputs.

Alternative approaches to paper (Saini et al., 2020) by fine-tuning pre-trained language models like GPT-3 or RoBERTa that have a good understanding of language semantics to explicitly target hate speech detection. Deep Learning Models and ensemble learning can be incorporated as discussed earlier. Images containing texts can also be included for analysis because they also depict hate speech. However, the approaches have to be experimented to find their effectiveness.

4.2. Lessons Learned

By reading and summarising papers I learnt the importance of hate speech detection, its applications, its significance on social media platforms, the importance of providing proper context in a research paper, and how to discuss the state-of-the-art or related work. By understanding the hate speech problem and the existing knowledge in the field, a specific gap can be addressed. I learnt that if dataset is undersampled it can lead to errors and bias in the model. Analysing if data is sufficient and equitable for modelling and training is crucial.Annotation might be difficult when the text shows properties of multiple classes. Hence data must be carefully annotated. I learnt about new approaches and processes employed in their studies, the complications of misclassifying and biassing the models, presenting results and findings, drawing conclusions from the experiments, and talking about limitations. Introducing new approaches can broaden our knowledge in the field, and citing and referencing can help us explore the topic more.

As part of Data Exploration, I understood answering the background questions, examining the dimensions of the dataset (number of rows and columns), exploring the the columns that are included in the dataset, their names, and the kinds of data they hold, checking for any missing values in the dataset, examining the distribution of classes for each label helps understand the balance or imbalance between different classes and potential biases in the dataset, sub-sampling helps reduce processing time and by using a stratified sampling approach distribution of classes can be maintained, splitting the dataset into train, validation, and test sets is essential for model evaluation and performance estimation, creating histograms, bar plots, or pie charts, helps to gain a better understanding of the distribution or relationships between variables and performing techniques such as word frequency counts, sentiment analysis helps to gain insights into the textual data present in the dataset. I learned from the text preprocessing steps that it helps in transforming raw text into a format that is more suitable for analysis or modelling, such as standardising text, Unicode removal, converting to lowercase, removing URLs and punctuation, splitting text into individual tokens or words, stopword removal reduces noise and focus on the essential content, lemmatization reduces words to their base or root form, known as the lemma making text more interpretable and readable.

Overall, this assignment has provided inspiration in the field.

5. Bibliographical References

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., and Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2019). Hate speech and offensive language. *Github*.

Dias, D. S., Welikala, M. D., and Dias, N. G. (2018). Identifying racist social media comments in sinhala language using text analytics models with machine learning. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 1–6. IEEE.

Fernando, W., Weerasinghe, R., and Bandara, E. (2022). Sinhala hate speech detection in social media using machine learning and deep learning. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 166–171. IEEE.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Ho, Y.-C. and Pepyne, D. L. (2002). Simple explanation of the no-free-lunch theorem and its implications. *Journal of optimization theory and applications*, 115:549–570.

Kumar, A., Kumar, S., and Tyagi, V. (2023). Automatic detection and monitoring of hate speech in online multi-social media. In *Advanced Communication and Intelligent Systems: First International Conference, ICACIS 2022, Virtual Event, October 20-21, 2022, Revised Selected Papers*, pages 605–612. Springer.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Nagar, S., Barbhuiya, F. A., and Dey, K. (2023). Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining*, 13(1):47.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Saini, Y., Bachchas, V., Kumar, Y., and Kumar, S. (2020). Abusive text examination using latent dirichlet allocation, self organizing maps and k means clustering. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1233–1238. IEEE.

Sandaruwan, H., Lorensuhewa, S., and Kalyani, M. (2019). Sinhala hate speech detection in social media using text mining and machine learning. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, volume 250, pages 1–8. IEEE.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Welgama, V. (2011). Evaluation of a shallow stemming algorithm for sinhala. In *Poster in Sri Lanka Student Workshop on Computer Science*. Faculty of Science, University of Jaffna.