CE807 – Assignment 1 - Interim Practical Text Analytics and Report

Student Id: 2202592

Abstract

The papers explore different methods for text classification using machine learning. One paper presents a model that uses natural language processing to classify news articles with 86 percentage accuracy, while another paper introduces a framework combining feature selection, parameter tuning, and ensemble learning for better text classification results. A third paper combines machine learning algorithms with optimization techniques for automatic text classification, and a fourth paper suggests a text classification method using WordNet hypernyms. The fifth paper proposes an SVM-based algorithm that uses genetic algorithms for high-accuracy text classification on multiple datasets. These papers offer insights and frameworks for achieving accurate and efficient text classification using machine learning.

1 Review of Generic Text Classification Methods (Task 1)

The five papers in this study tell us the various methods for text classification by using machine learning techniques. first paper [6] displays us the text classification model that uses the natural language processing (NLP) technique/approach, with pre-processing techniques applied to it like tokenization, stemming and stop word removals. This model also uses feature selection (chi-squared test integrated with classification on a support vector machine (SVM) algorithm), performing well around the accuracy rate of 86 percentage on research journal articles.

The second paper [3] describes us the framework for efficient English words and text classification using various machine learning techniques. This framework integrates many of the techniques like feature selection, parameter tuning, and ensemble learning algorithm like Random forest, etc to achieve well-performing rate accuracy in text classification, outperforming traditional machine learning models in terms of accuracy and efficiency.

A third paper [2] tells us about an automatic text classification approach that combines multiple machine learning algorithms, optimization techniques, utilizing feature selection with the chi-squared test and use of classification machine learning algorithms like i,e decision trees, SVM, and k-nearest neighbors (KNN). The method was evaluated on various datasets which resulted in better results compared to traditional machine learning models.

The fourth paper [8] suggests a text classification method that uses (WordNet hypernyms), where the text is presented using WordNet hypernyms which then is used for naive Bayes classifier for applying the classify of text into predefined categories. The paper tell us that they evaluated the method on a dataset of Reuters journal and compared its performance with other traditional machine learning models i,e (DT, etc), showing that the WordNet-based method performed well among them in terms of accuracy.

The fifth paper [9] proposes an optimal SVM-based text classification algorithms that has achieved high accuracy through parameter optimization with the help of a genetic algorithm and its mutation and combine. The proposed algorithm involves feature selection with the chi-squared test and classification using SVM with a radial basis function kernel. It was evaluated on multiple datasets and outperformed other traditional machine learning models in terms of accuracy.

1.1 Critical Discussion (Task 1)

Determining the best technique and methos used in the papers among the following ones is challenging without specific problem criteria. However, they differ in their techniques and methods used:

[6] applies NLP techniques such as tokenization, stemming, and stop word removal to perform feature selection with classification using the SVM algorithm.

[3] combines various techniques including pre-processing, feature selection, and multiple machine learning algorithms such as decision tree, SVM, and naive Bayes, and assesses the effectiveness of each approach.

[2] suggests a text classification method using machine learning and optimization algorithms like PSO and FFA to optimize feature selection and classification.

[8] investigates the use of WordNet hypernyms to categorize text, which involves assigning words to their broader categories.

[9] proposes an optimized SVM-based text classification algorithm that selects the best features and fine-tunes the SVM parameters.

On contrary, these papers have unique techniques and objectives. Therefore, choosing the most suitable approach depends on the context and specific goals of the text classification task, but each has used different methods and different techniques which gave a different performance of the algorithm which all in all is useful for different problems and be solutions for it

2 Review of Offensive Language Detection Methods (Task 2)

The paper [5] proposes a deep learning approach for detecting offensive language in tweets. The authors trained a Convolutional Neural Network (CNN) model on a dataset of 25,000 annotated tweets, achieving high accuracy in identifying offensive language. They also compared their model's performance with other machine learning methods and found that the CNN approach outperformed them. The study suggests that deep learning techniques can effectively detect offensive language in social media, which has potential applications in content moderation and online safety.

The Paper [7] introduces a new large-scale semi-supervised dataset for identifying offensive language in text. The dataset contains over 25,000 annotated examples and includes a wide range of offensive language types. The authors propose a new method for training models on this dataset that combines both supervised and unsupervised learning. The experimental results show that their method achieves state-of-the-art performance on multiple benchmark datasets.

The paper [10] discusses the ZYJ123 system, which took part in the DravidianLangTech-EACL2021 shared task on detecting offensive language. The system leverages the XLM-RoBERTa model and DPCNN architecture to identify offensive language in tweets across four Dravidian languages. The results showed that the system had a competitive performance on the task.

The paper [4] The article discusses DEEP, a machine learning framework that participated in the HASOC2019 shared task. DEEP uses various models, including LSTM, SVM, and BERT, to identify hate speech and offensive language in tweets written in English, Hindi, and German. The framework's outcomes demonstrate effective performance on the shared task.

The paper [1] suggests a method for identifying hate speech and offensive language on Twitter using machine learning, which uses n-grams and TF-IDF methods to extract features from tweets and train a model to classify them as offensive or not. The proposed approach is tested on a dataset of tweets, demonstrating favorable results in detecting hate speech and offensive language.

2.1 Critical Discussion (Task 2)

The paper [5] presents a deep learning-based approach for detecting offensive language in tweets, which has the potential to outperform traditional machine learning techniques. The proposed approach includes pre-processing steps to deal with noisy data and feature extraction using word embeddings. on the contrary, the paper has limitations such as only being evaluated on a small dataset of 4,000 tweets, which may not be representative of all

tweets. The performance of the approach is not compared to other state-of-the-art methods, limiting its effectiveness assessment in comparison to other approaches. Moreover, the paper does not discuss the ethical implications of using the proposed approach for detecting offensive language in tweets.

The paper [7] introduces a large-scale semi-supervised dataset that can be utilized to train machine learning models to detect offensive language in various languages. The dataset is diverse, covering multiple languages and platforms, which makes it useful for building models that can generalize across platforms and languages. Furthermore, the paper provides valuable insights into the characteristics of offensive language across different languages and platforms, which can help in the development of more efficient detection models. But the disadvantages of this paper is rather than proposing a new method for detecting offensive language, the paper concentrates on presenting a dataset for this task. The dataset is semi-supervised, which means that it still necessitates a considerable amount of labeled data for training. Additionally, the dataset only includes a limited number of languages, and offensive language usage can differ substantially in other languages. Finally, the paper does not address the potential biases in the dataset or the ethical considerations associated with using the dataset to create offensive language detection models.

The paper [10] achieved competitive performance on the DravidianLangTech-EACL2021 shared task on offensive language identification. The paper also provides insights into the challenges of detecting offensive language in multilingual settings. But, focuses on offensive language detection in tweets written in four Dravidian languages, limiting the generalizability of the proposed approach to other languages and text types. The paper does not compare the performance of the proposed approach to other state-of-the-art methods, which makes it difficult to assess its effectiveness in comparison to other approaches. Finally, the paper does not discuss the ethical implications of using the proposed approach for detecting offensive language.

The paper [4] The framework uses various deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to extract features from text data and classify it as offensive or non-offensive. The proposed framework achieved promising results in the HASOC2019 shared task on hate speech and offensive language detection. But, The paper only evaluates the proposed framework on a single dataset and does not compare its performance to other state-of-the-art methods. The proposed framework may also require a large amount of labeled data for training, which can be challenging to obtain in some contexts. Finally, the paper does not discuss the ethical implications of using the proposed approach for detecting hate speech and offensive language.

The paper[1] approach is simple and easy to implement, making it useful for practical applications. The paper also provides insights into the effectiveness of different features for detecting hate speech and offensive language, which can aid in the development of more effective detection models. But, The disadvantage of the paper is it only evaluates the proposed approach on a small dataset of 10,000 tweets, which may not be representative of the entire population of tweets. The performance of the approach is also not compared to other state-of-the-art methods, which limits the ability to assess its effectiveness in comparison to other approaches. Additionally, the paper does not discuss the ethical implications of using the proposed approach for detecting hate speech and offensive language on social media platforms

3 OLID Dataset Characterization (Task 3)

Who made collected the data? Are they alright with you using it?

A dataset called OLID was created by a team of researchers for the OffensEval shared task in 2019. The individuals behind the paper "Predicting the Type and Target of Offensive Posts in Social Media" are Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. The dataset is available to the public under the CC BY-SA 4.0 license, permitting

sharing and adaptation of the material for any purpose as long as the license terms are met.

What is in the data? Is it what you need for your work?

There are 14,200 English tweets that have been annotated in the OLID dataset. The data is categorized into three different levels of annotation, which are Offensive Language Detection, Categorization of Offensive Language, and Offensive Language Target Identification. The dataset is appropriate for research concerning the automatic detection of offensive language on social media. However, researchers must assess whether the dataset is relevant to their specific research questions and objectives.

Where was this data produced? Where is it now?

The authors created the data for the OffensEval shared task held in 2019. The dataset is currently available to the public on the OLID dataset website and has been utilized in several research projects and studies.

Why was it produced? Can you trust it?

The primary aim of producing the OLID dataset was to enhance the advancements in automatic detection of offensive language on social media. To create the dataset, a group of workers annotated tweets by categorizing them as offensive or non-offensive and identifying the type of offense. The OLID dataset is extensively used in academic research and has led to the development of many successful models for detecting offensive language. Nevertheless, since it is a crowdsourced dataset, there may be limitations regarding the reliability and quality of annotations, and it is crucial to critically evaluate the data and methods used for collection and annotation.

When was it produced? What's happened to it since?

The OLID dataset was created in 2019 for the OffensEval shared task and has since been extensively utilized in academic research. Several universities have also incorporated it into their research projects.

4   Summary

The referenced papers helps us understand the different approaches and different technique used to solve the problems of text clas-

sification and text identification like offensive language identification. But, all the methods only helps us understand different methods used to solve the particular problem and helps us implement a better one by comparing and using the previous knowledge from the research papers.

References

[1] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*, 2018.

[2] R Janani and S Vijayarani. Automatic text classification using machine learning and optimization algorithms. *Soft Computing*, 25:1129–1145, 2021.

[3] Xiaoyu Luo. Efficient english text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3):3401–3409, 2021.

[4] Hamada A Nayel and HL Shashirekha. Deep at hasoc2019: A machine learning framework for hate speech and offensive language detection. In *FIRE (Working Notes)*, pages 336–343, 2019.

[5] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*, 2018.

[6] Maria Razno. Machine learning text classification model with nlp approach. *Computational Linguistics and Intelligent Systems*, 2:71–73, 2019.

[7] Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*, 2020.

[8] Sam Scott and Stan Matwin. Text classification using wordnet hypernyms. In *Usage of WordNet in natural language processing systems*, 1998.

[9] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, and Xin Li. An optimal svm-based text classification algorithm. In *2006 International Conference on Machine Learning and Cybernetics*, pages 1378–1381. IEEE, 2006.

[10] Yingjia Zhao and Xin Tao. Zyj123@ dravidianlangtech-eacl2021: Offensive language identification based on xlm-roberta with dpcnn. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 216–221, 2021.