Amrita_CEN at SemEval-2022 Task 4: Oversampling-based Machine Learning Approach for Detecting Patronizing and Condescending Language

Bichu George, S Adarsh, Nishitkumar Prajapati, Premjith B, and Soman K.P
Centre for Computation Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
b_premjith@cb.amrita.edu

Abstract

This paper narrates the work of the team Amrita_CEN for the shared task on Patronizing and Condescending Language Detection at SemEval 2022. We implemented machine learning algorithms such as Support Vector Machine (SVM), Logistic regression, Naive Bayes, XG Boost and Random Forest for modelling the tasks. At the same time, we also applied a feature engineering method to solve the class imbalance problem with respect to training data. Among all the models, the logistic regression model outperformed all other models and we have submitted results based upon the same.

1 Introduction

Discriminatory language on the social media is lately creating hostile environment towards the vulnerable communities especially women and minorities. These are reflected in day to day conversations happening on popular social media sites. It is a high time now to build a technological solution to counter the discrimination against vulnerable communities. Here in this task, we consider one such issue known as "Patronizing and Condescending Language (PCL) Detection". When someone's language conveys a pompous attitude toward others or portrays them or their circumstances in a compassionate manner, eliciting feelings of sympathy and compassion, they are patronising or condescending. This is why it is important to develop a computational model to predict whether there is patronizing content in social media or not (Pérez-Almendros et al., 2020). This challenge can be solved by the applying Natural Language Processing (NLP) concepts. The Social media platforms reaches a huge audience, which might contribute to increased exclusion and inequity among vulnerable groups. Despite the fact that harmful language behaviour (such as hate speech, abusive language, fake news, rumour propagation, or disinformation) (Sreelakshmi et al., 2020), (Sreelakshmi et al., 2021) has been extensively investigated in NLP, PCL has remained a neglected field of research.

We implemented seven machine learning models which include three classical machine learning algorithms and four ensemble models: Support Vector Machine (SVM), Logistic regression, Naive Bayes, XG Boost and Random Forest for modelling the tasks (Soman et al., 2009), (Premjith et al., 2019), (Premjith and Kp, 2020). The class imbalance problem was dealt by a minority oversampling technique called SMOTE and comparative analysis of our algorithm was done by various evaluation metrics such as precision, recall and F1 score.

The remaining parts of the paper are described as follows: Section 2 contains dataset description along with works related to that. Section 3 describes the system overview. Section 4 explains the experimental setup. Section 5 discusses result and the paper is concluded in Section 6.

2 Related works

This section provides a brief review of the literature published for the detection of various offensive and abusive contents pertained to violence, cyberbullying etc. shared on the social media.

Adithya et.al (Bohra et al., 2018) analysed the hate speech data in code-mixed form and proposed classification models for the detection. They created a dataset consisting of Hindi-English code-mixed tweets. Machine learning algorithms like SVM, Random forest were used for the classification of tweets into different categories. Conroy et.al (Rubin et al., 2016) reported the problem of fake news detection in their paper and their study offered a classification of different types of truthfulness evaluation methods that fall into two categories: linguistic cue with machine learning and network analysis approaches. Zampieri et al (Zampieri et al., 2019) predicted the nature and victim of offensive content shared on social media. They used the Of-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 515 - 518

fensive Language Identification Dataset (OLID) for the analysis. They compared the performance of different machine learning models on this dataset. Wang and Potts (Wang and Potts, 2019) used a corpus called TALKDOWN for detecting the condescension in a text by incorporating the context. The dataset consist of annotated social media messages. They explored the issue of modelling condescension in direct communication from an NLP perspective. They used BERT-based models for developing the baseline models.

3 Task and Data Description

3.1 Task1

The competition mainly consisted of 2 sub tasks (Pérez-Almendros et al., 2022). The objective of the subtask 1 is to develop a model, which could predict whether a given paragraph contain condescension or not, which is a binary classification problem. The dataset used for subtask 1 consists of 10469 paragraphs. Each of the paragraphs describes the people belonging to vulnerable social categories. It contains excerpts from news items from 20 English-speaking nations that feature at least one of the following terms relating to potentially weaker sections of the society: vulnerable or women, refugee, hopeless, migrant, immigrant, in need, homeless, poor families, disabled, with Patronizing and Condescending Language (PCL) comments.

3.2 Task2

The objective of the subtask 2 is to develop a model, which could predict whether a given paragraph comes under any of the top 7 PCL taxonomies namely, Unbalanced power relations, Shallow solution, Presupposition, Authority voice, Metaphor, Compassion, The poorer, the merrier, which is a multi-label classification problem. The dataset used for subtask 2 consists of 993 paragraphs. Each of the paragraphs describes the people belonging to vulnerable social categories. It contains excerpts from news items from 20 English-speaking nations that feature at least one of the following terms relating to potentially weaker sections of the society: vulnerable or women, refugee, hopeless, migrant, immigrant, in need, homeless, poor families, disabled, with Patronizing and Condescending Language (PCL) comments.

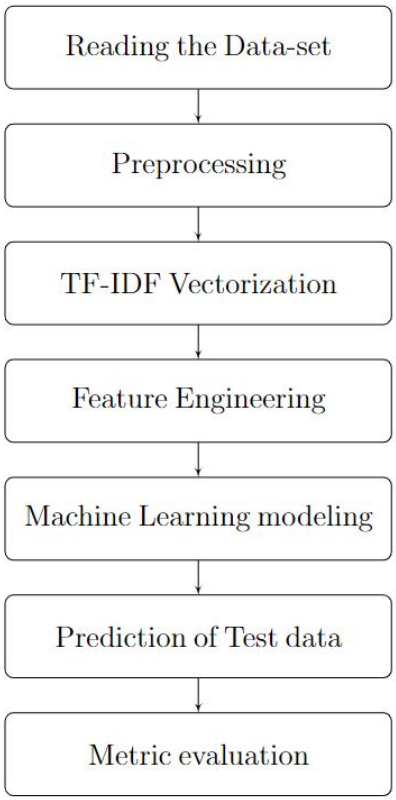

Figure 1: Flowgraph of the methodology

4 System Overview

This section discusses the procedure followed for developing models for each subtasks in completion. Figure 1 represents the block diagram of the workflow of the methodology.

This section explains the steps followed for developing models for the PCL shared tasks.

4.1 Preprocessing

Initially, we cleaned the data by removing stopwords, URLs and special characters. The cleaned texts were tokenized and lemmatized to obtain the root form of the word. It helped to reduce the vocabulary in the corpus, which further reduce the dimension of the sentence vector obtained using Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer algorithm.

4.2 Feature Engineering

We represented the textual data as vectors using TF-IDF for the further processing. In addition to that, we employed SMOTE (SMOTE: Synthetic Minority Over-sampling Technique) (Chawla et al., 2002), an oversampling algorithm to address the problem of class imbalance in the data. The

SMOTE algorithms synthetically generates random data for the minority classes to increase the size of the minority classes. It is done by selecting one or more of random k-nearest neighbour for each minority instances. We employed SMOTE after converting texts into vector using Term Frequency-Inverse Document Frequency vectorizer algorithm.

4.3 Machine Learning modelling

The dataset for subtask 1 consists of total 10469 instances and for subtask 2 it is 993 instances. We considered a train-test split ratio of 80:20. The parameter stratify was used for the purpose of making a split so that the share of values in the sample produced will be the equal to the proportion of values provide to parameter stratify. For prediction, we have a total of 2094 test instances in which 1895 belongs to class 0 and 199 belonging to class 1 in subtask 1 and 198 test instances. For logistic regression model, hyper parameter tuning was done using sklearn's GridSearchCV function [1]. The parameters that was given for tuning was penalty =$l1, l2$ and value of $C = array([0.01, 0.1, 1, 10, 100])$. After hyperparameter tuning using GridSearchCV, the best parameters were found to be $C(regularization\_term) = 10$ and $Penalty = l2$. For subtask 2, we set the class_weight hyperpaameter to be 'Balanced'. To predict the multi-label output, we used the 'MultiOutputClassifier' function from Scikit-learn [2]. For models other than logistic regression, we used default parameters available in Scikit-learn for classification.

4.4 Evaluation

The evaluation measures used for this work were macro average F1,precision and recall. Recall is ratio of correct positive predictions to the total number of positives and Precision is ratio of correct positive predictions to the total number of positive predictions. F1 score is the harmonic mean of precision and recall. Macro average is defined as the average of precision, recall, F1 score on different classes.

5 Results

In both the sub tasks we used three classical ML models and four ensemble techniques for classification. The three ML models were logistic regression ,SVM and DecisionTreeClassifier and the ensemble techniques were Bagging classifier, Random forest, GradientBoost and XGBoost. Validation dataset was used to get a comparative analysis of our algorithm. In this analysis we used evaluation metrics such as precision, recall and F1 score. The official evaluation metric was F1 score for positive class for subtask 1. For the validation dataset an F1 score of 0.41 was achieved for positive class and in case of test dataset an F1 score of 0.39 was obtained and our final rank for subtask 1 in the competition was 60. For subtask 2 we got a macro_average F1 score of 0.45 during the post evaluation phase.

Model	Recall	Precision	F1
Log Reg	0.73	0.64	0.66
SVM	0.53	0.75	0.53
Dec Tree	0.57	0.57	0.57
Bagging	0.54	0.61	0.55
Random For	0.51	0.64	0.49
GradBoost	0.53	0.75	0.54
XGBoost	0.56	0.68	0.58

Table 1: Comparitive analysis of our ML models for subtask 1 considering macro averages

Model	Recall	Precision	F1
Log Reg	0.48	0.45	0.45
SVM	0.30	0.58	0.32
Dec Tree	0.35	0.36	0.35
Bagging	0.29	0.41	0.33
Random For	0.27	0.56	0.31
GradBoost	0.28	0.45	0.32
XGBoost	0.33	0.47	0.36

Table 2: Comparitive analysis of our ML models for subtask 2 considering macro averages

From the Tables 1 and 2 we can clearly see that the macro F1 score of Logistic regression stood out among all the other models. Moreover the execution time for logistic regression was less compared to other models especially the ensemble techniques. Hence this model was used for the final prediction of the test dataset.

6 Conclusion

This paper narrates the work of Amrita_CEN with respect to SemEval 2022 Task 4 competition named " Patronizing and Condescending Language Detection ". A total of seven machine learning algorithms were used which include three classical ML models and four ensemble techniques. The problem

[1]GridSearchCV: https://rb.gy/lajkio
[2]MultiOutputClassifier: https://rb.gy/52vpax

of class imbalance was dealt with minority oversampling technique called SMOTE. Considering macro F1 score for both the sub tasks, logistic regression performed the best and the results were submitted using the same model. Coming to the future work, implementation using deep learning and BERT approaches can give better results compared to classical machine learning models.

7 Acknowledgements

With the invaluable guidance of our faculty Dr Sowmya V, this work and the research has been feasible. Her constant guidance, attitude of perfection and passion of detailing were an inspiration and helped us to achieve our objective.

References

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 Task 4: Patronizing and Condescending Language Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

B Premjith and Soman Kp. 2020. Amrita_cen_nlp@ wosp 3c citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 71–74.

B Premjith, KP Soman, M Anand Kumar, and D Jyothi Ratnam. 2019. Embedding linguistic features in word embedding for preposition sense disambiguation in english—malayalam machine translation context. In *Recent Advances in Computational Intelligence*, pages 341–370. Springer.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

KP Soman, R Loganathan, and V Ajay. 2009. *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd.

K Sreelakshmi, B Premjith, and Soman Kp. 2021. Amrita_cen_nlp@ dravidianlangtech-eacl2021: deep learning-based offensive language identification in malayalam, tamil and kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.