https://doi.org/ 10.48047/AFJBS.6.10.2024.5758-5768

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Use of a refined Distil-BERT model and Data Augmentation to Identify Patronizing and Condescending Language

Bolukonda Prashanth[1,a)] and Dr P Vijaya Pal Reddy[2]

[1]Research Scholar, Department of CSE, University College of Engineering, Osmania University, Hyderabad, Telangana, India.
[2]Professor, Department of CSE, Matrusri Engineering College, Hyderabad, Telangana, India.

[a)]Corresponding author : boluk.cse@gmail.com

Article History

Volume 6,Issue 10, 2024

Received:24 May 2024

Accepted : 02 Jun 2024

doi: 10.48047/AFJBS.6.10.2024.5758-5768

Abstract

Patronizing and condescending language can be represented positively or negatively some times. Patronizing can mean “giving support to”. Condescending can mean “superior attitude towards others”. The proposed work takes the news articles from various countries, and detect whether the news article contains any patronizing and condescending language or not. We apply deep learning (DL) techniques to address this issue, approaching it like a typical text classification task. We suggest using a pre-trained model. Distil-Bert and Data Augmentation. The proposed model DistilBERT achieved a precision of 76.18 and f1 score of 72.48. Data Augmentation with Synonym replacement achieved a precision of 78.12 and f1 score of 73.87.

1. Introduction

The rapid surge in social media usage in recent years has also made it possible for the use of patronizing and condescending language to increase (PCL). Language that is patronizing and condescending presents actions that appear helpful or kind, but they actually reveal a sense of superiority toward other people. Although until now PCL was an understudied area of research, harmful language behavior such as hate speech[1], offensive language[2], fake news[3], rumor propagation or misinformation[4], and many others have been extensively studied in NLP.
Even for humans, identifying PCL can be challenging because to its subtlety and subjectivity. For example, something that one person finds condescending may be seen by another as an objective depiction of the circumstances, or some people may not see anything wrong with

explaining how those in a privileged position give what remains to those in need. Additionally, when reading how others refer to them, A person who is a part of a so-called vulnerable community may expect to feel more patronized than someone who is not.

SemEval 2022-Task 4's objective is to create a system that can determine which text contains PCL and whether it does or not. The condescension is expressed by the PCL category. Two datasets were made available by the organizers: one with PCL classifications and the other with annotations based on PCL intensity. Various authors are proposed different models in the context and achieved an intensive result. Our work proposed a pre-trained model Distil-Bert to detect PCL in a given text. Data augmentation also proposed in order to address the issue of class disparity. We consider PCL detection task for this work.

The remainder of the article is arranged as follows. Section 2 provides an overview of the related work. The dataset is described in Section 3. The experimental configuration of our suggested model is outlined in Section 4. Pre-processing and implementation specifics for the suggested model are involved. The outcomes and discussion are covered in Section 5. Finally, we wrapped up the report and offered some suggestions for more research in the Section number 6.

2. Related Work

The investigation of inequitable, deceptive, or objectionable language has garnered the attention of numerous scholars among the community of NLP researchers. The primary work in this particular situation revolve around identifying explicit, aggressive, and striking occurrences, such as detecting fake news or conducting fact-checking [5,6,7,8]. Other tasks include detecting propaganda techniques [9], modelling offensive language [10,11], identifying hate speech [12], and studying rumor propagation [13]. However, there are other more nuanced yet equally detrimental forms of language that have not received as much scrutiny from the NLP community. These forms, due to their subtle characteristics, are likely to be more challenging to identify. One example of this is detecting Patronizing and Condescending Language (PCL), which was the main topic of Task 4 at SemEval-2022.

Patronizing or condescending communication occurs when an entity's language reflects a sense of superiority towards others. When beliefs become normalized, they establish prejudice as a routine and make it less apparent[14]. Moreover, the utilization of PCL is frequently unintentional and well-meaning, particularly when discussing marginalized communities. The presence of this positive sentiment, known as good will, can amplify the adverse impact of PCL. This is because the audience, without strong resistance, is more susceptible to the negative consequences of this discriminatory language, frequently without realizing it.

Sociolinguistic research characterizes PCL as a nuanced kind of language that is frequently unintentional but has negative and discriminatory effects[15]. The process generates and perpetuates prejudices[16], leading to increased marginalization, dissemination of rumors, and dissemination of inaccurate information[17]. PCL also has a tendency to reinforce power-knowledge dynamics[18], advocating for acts of charity rather than cooperation and portraying those who can provide assistance as rescuers of individuals in a less advantaged position[19,20]. In addition, PCL has a tendency to obscure the persons or groups accountable for deeply ingrained societal issues, sometimes by indirectly or directly assigning blame to disadvantaged communities or individuals for their circumstances. Moreover, it frequently relies on temporary and simplistic remedies[21]. Privileged groups have been known to utilize

PCL, which is associated with the concept of "pornography of poverty". This communication style portrays vulnerable situations with a language of pity in order to elicit charity action and sympathetic attitudes from the intended audience.

Although the adverse effects of PCL, in terms of social interactions and in business and political discourse, have been thoroughly examined in the field of social sciences, it is still a relatively unexplored phenomenon in the field of NLP. However, we are of the opinion that the identification of PCL presents several significant hurdles for NLP study. Therefore, further research in this field is necessary, particularly considering the potential societal advantages that would arise. Considering how subjective and subtle it is, we can anticipate that detecting PCL will be more challenging compared to tasks that concentrate on more obvious occurrences. Furthermore, the identification of PCL frequently necessitates an implicit comprehension of human ethics and values, This requires logical reasoning of a kind that NLP models are probable not going to find easy.

As previously stated in the introduction, PCL has been thoroughly examined in the field of sociolinguistics by Margi´c[22], Giles et al.[23], Huckin[24], and Chouliaraki[25]. In the field of NLP, the study of patronizing discourse has not been given much focus. Wang and Potts[26] deviated from the norm by creating a collection of Reddit comments, specifically chosen for their use of condescending language, and annotated accordingly. It is important to mention that, unlike SemEval mission, their research did not expressly concentrate on marginalized communities. In the earlier study of Carle Perez[27] presented Don't Patronize Me!, which is, as far as we know, the initial annotated collection of PCL (Patronizing Language) targeted at communities that are at risk. This database served as the training data for the SemEval job. Related research has examined discourse types closely associated with condescension. For instance, Sap et al.[28] investigated how specific language usage reflects power dynamics. Mendelsohn[29] explored the depersonalization of marginalized groups through language, while Zhou and Jurgens[30] examined the interaction between sympathies and empathy expressed in online groups with authoritative voices.

3. Dataset

The primary source content for this undertaking is a dataset called "Don't Patronize Me!" (DPM). It is an annotated collection of language that is both patronizing and condescending towards marginalized people. This dataset was initially presented in earlier research conducted by Perez-Almendros et al. in 2020[27]. 10,469 paragraphs make up the dataset, which served as the SemEval assignment's training set. In order to generate the test set for this work, The authors meticulously annotated an additional 3,898 paragraphs, adhering to the identical procedure. Twenty English-speaking countries media outlets provided the news articles from which the paragraphs were taken. The original source of these pieces is the News onWeb (NoW) corpus, as given by Davies in 2013. We are using this dataset with the permission of authors [27] for research purpose.

4. Implementation

4.1 Pre processing

This section outlines the fundamental data pre-processing procedures that were implemented for our experiments:

1. Initially, we convert the Label column into a binary representation. If the value is either zero or one, we convert it to zero. If the number is two or three, we transform it to one. After this

modification, the Label becomes a Binary column containing two values: Zero, indicating the absence of PCL in the text, and One, indicating the presence of PCL. For sub-task one, we transformed the column to binary in order to ascertain the presence of PCL, without considering the extent of PCL.

2. We have dropped all null value attributes from the dataset, then divided the dataset in an 80:10:10 ratio into train, test and validation sets. Tokenization is applied on the split datasets then created as batches. Pre-trained models are not compatible with raw text. Therefore, we transformed the text into encoding and added with 2 more columns input_ids and attention masks to extract the features of datasets. Subsequently, the encoded sequence is sent into the model to execute the classification process.

4.2 Classification

4.2.1 DistilBERT:

Classification is implemented using a fine-tuned pre-trained distilBERT-base-uncased model. The distilBERT-base-uncased model is a widely used version of the BERT (Bidirectional Encoder Representations from Transformers) family. It is particularly created to be smaller, quicker, and more effective, while still maintaining a significant portion of BERT's performance. The intricacies of this paradigm and comprehend its attributes and use. The team at Hugging Face introduced DistilBERT as a means to develop more streamlined and effective iterations of the BERT paradigm. The objective is to achieve a harmonious equilibrium between the performance of the model and its computational efficiency. DistilBERT does this by decreasing the quantity of layers and parameters in comparison to BERT, while utilizing a process called "knowledge distillation". The uncased approach is commonly favored when case sensitivity is not crucial, which simplifies the tokenization process. DistilBERT has 6 transformer layers and will take 66 million parameters which are fewer then BERT 110 million parameters. Figure 1 represents the DistilBERT model architecture.

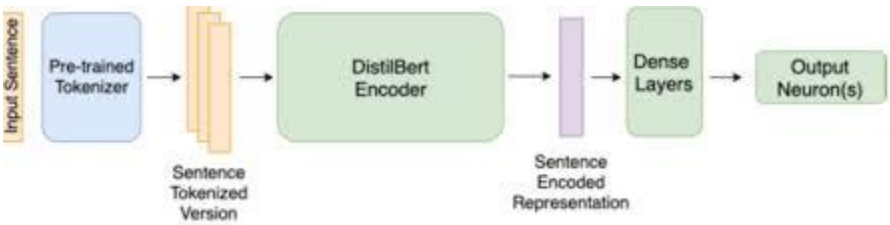

Figure 1: DistilBERT model

Although it is smaller in size, DistilBERT achieves competitive performance on a range of NLP benchmarks. It attains approximately 97% of BERT's performance on workloads like the GLUE benchmark, which makes it a compelling choice for environments with limited resources. The training process involves utilizing distillation of knowledge, a technique in which a smaller model known as DistilBERT is taught to imitate a bigger model referred to as BERT. This technique enables the student model to encapsulate a significant portion of the teacher's expertise in a more condensed format.

The hyper parameters used in training the model epoch=3, batch size=8 and learning rate at 2e-5. Which are given an extensive result.

4.2.2 Data Augmentation:
Data augmentation is a potent method for enriching training datasets and improving model performance across diverse data types. When it comes to Natural Language Processing, techniques like synonym substitution, reverse translation, and contextual expansion play a crucial role in generating varied and resilient training datasets. Applying data augmentation proficiently can greatly enhance the ability of models to generalize and perform well, particularly in situations where there is a scarcity of data.

We have also applied the method of data augmentation to improve the results as the dataset has the class imbalance issue. We have followed 3 different data augmentation approaches discussed below.

4.2.2.1 Synonym Replacement: Synonym substitution is a straightforward yet impactful data augmentation approach for textual data. It entails substituting terms in a sentence with their equivalents to create novel sentences that are both syntactically and semantically comparable. This contributes to the diversification of the training data and enhances the model's capacity to generalize.

4.2.2.2 Contextual Augmentation: Contextual augmentation is an advanced method of augmenting text data that utilizes the context within sentences to create more relevant and grammatically accurate variants. This approach commonly utilizes pre-trained language models such as BERT, GPT, or their variations. These models have the ability to comprehend and produce words based on the context provided by the surrounding text.

4.2.2.3 Back Translation Augmentation: Back translation is a data augmentation approach employed in Natural Language Processing (NLP) which entails translating text into a different language and subsequently translating it back into the original language. This approach is especially efficient in producing a wide range of top-notch training data while preserving the semantic significance of the original text. We have used English to German and German to English back translation.

5 Results and Discussion
Proposed model DistilBERT-base-uncased for detecting PCL has performed well. We used the metrics Precision, Recall and F1-Score to judge our model performance. Our model achieved the precision score 76.18, Recall score 69.13 and F1 score 72.48 at testing phase which is shows the increase in the metrics with comparing models. We compare our proposed model with ensemble of transformer models [31] has achieved Precision 64.6, Recall score 65.6 and F1 score 65.1 and Prompt based Learning [32] has achieved Precision 61.2, Recall 67.2 and F1 score 64.1. In comparison of metrics our proposed model achieved a good score. Larger transformer models typically provide slower inference than the DistilBERT, which is most important for real time data. Larger models typically fine-tune slower than the DistilBERT, which is most beneficial during the training phase. Learned knowledge is successfully transferred during the distillation process using DistilBERT over larger transformer models [33].

For text classification tasks, DistilBERT is usually adjusted on specific datasets, which let it to pick up on task-specific patterns and subtleties. Comparing these-fine tuning techniques to all purpose Prompt-based procedures, accuracy is frequently higher. When compared to huge prompt-based models, DistilBERT-base-uncased is much more efficient in terms of processing resources. Adjusting DistilBERT does not require complicated quick engineering or dynamic modifications. Instead, it uses a simple procedure to modify model weights based on task-specific inputs. When it comes to training and data requirements, DistilBERT can be more effective than prompt-based techniques [33].

We propose data augmentation with synonym replacement, It obtained an F1 score 73.87, a recall score of 70.03 and precision score of 78.12. In contrast to prompt-based learning and transformer-based models, data augmentation with synonym replacement may be the most effective method for text categorization when comparing them along multiple important parameters. This in-depth comparison illustrates the advantages of synonym substitution in text classification. The training dataset efficiently expands in size and diversity without the need for new labelled data by substituting words with their counterparts. This may improve the model's ability to generalize to new examples, which reduces the overfitting. Synonym substitution [34] is a simple method to incorporate into the data pre-processing pipeline and needs very little computer power. This is in stark contrast to the computation burden of employing big prompt-based models to generate replies, or training massive transformer models. Because synonym replacement augmentation of data does not require pre-trained models. It is a strategy that may be used to a wide range of languages and topics. Compared to huge transformer models or prompt-based systems, augmented dataset with synonym replacement are easier to scale and employ with smaller models that are easier to implement.

We propose data augmentation with contextual word embedding, It obtained an F1 score 71.01, a recall score of 68.76 and precision score of 73.43. We can investigate a few important characteristics of data augmentation with contextual word embedding [35], which suggests that it may be a better method for text classification than prompt-based learning and transformer-based models. Contextual word embedding produced by models like BERT or comparable ones are utilized to provide text data variations that are more realistic and semantically rich. As a result, simpler models may perform better and become competitive with more sophisticated strategies. When used for data augmentation, contextual word embedding offers rich, semantically relevant representations of words in their particular settings, improving generalization and performance. For specialized applications like Patronizing and Condescending Language detection, contextual embedding can be refined to capture pertinent nuances on domain-specific data, resulting in more efficient data augmentation. By adding contextual embedding to data, the model becomes more resilient to linguistic variances and adversarial inputs, resulting in stronger performance.

We propose another model data augmentation with back translation, It obtained an F1 score 67.97, a recall score of 65.17 and precision score of 71.04. We may look at a few important features that make Data Augmentation with Back Translation [36] the best method for text classification when compared to Transformer-based models and prompt-based learning. Back translation creates varied and semantically rich data variants by translating text into another language and then back into the original. This technique would improve model robustness and performance without adding to the computational load of more complicated models. Back translation, which involves translating a document into another language and back again, produces paraphrases that include diverse word choices and sentence structures, which improves the model's ability to generalize to various linguistic patterns. It Especially helpful for capturing a wider range of expressions, preserves the text's overall meaning while incorporating natural linguistic differences, in contrast to simple synonym replacement.

In our proposed models, the DistilBERT-base-uncased model, when coupled with diverse data augmentation techniques, demonstrated superior performance over traditional transformer-based models and prompt-based-models. Specifically, precision improved by 20.93%, indicating a reduction in false positives. Recall saw a 4.25% increase, showcasing better detection of true positives. Consequently, the F1-score rose to 13.47% significantly, reflecting an overall balanced and enhanced performance. These improvements are attributable to the model's effective architecture and improved capacity for generalizing from enriched data, which together allow for more reliable and accurate predictions.

Figure 2, 3, and 4 compare the precision, recall, and F1-score of all the proposed models and comparing models that are being presented. The DistilBERT model's Normalized confusion matrix is displayed in Figure 5.

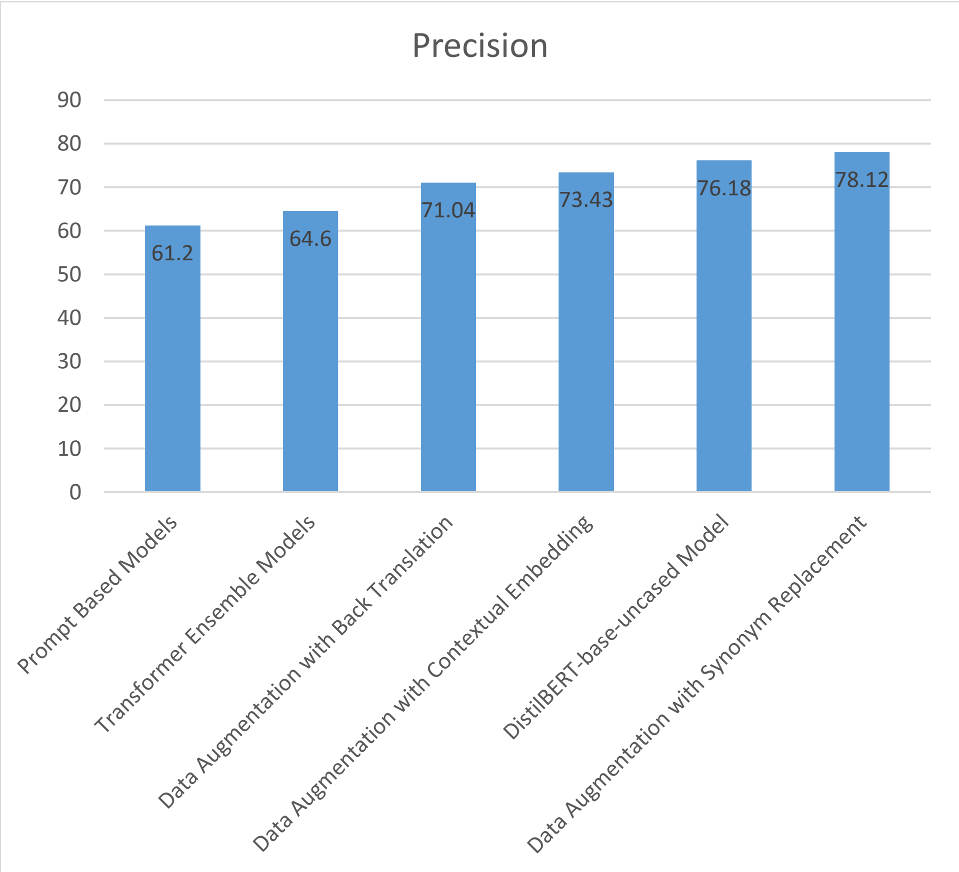

Figure 2: Precision score of proposed and comparing models

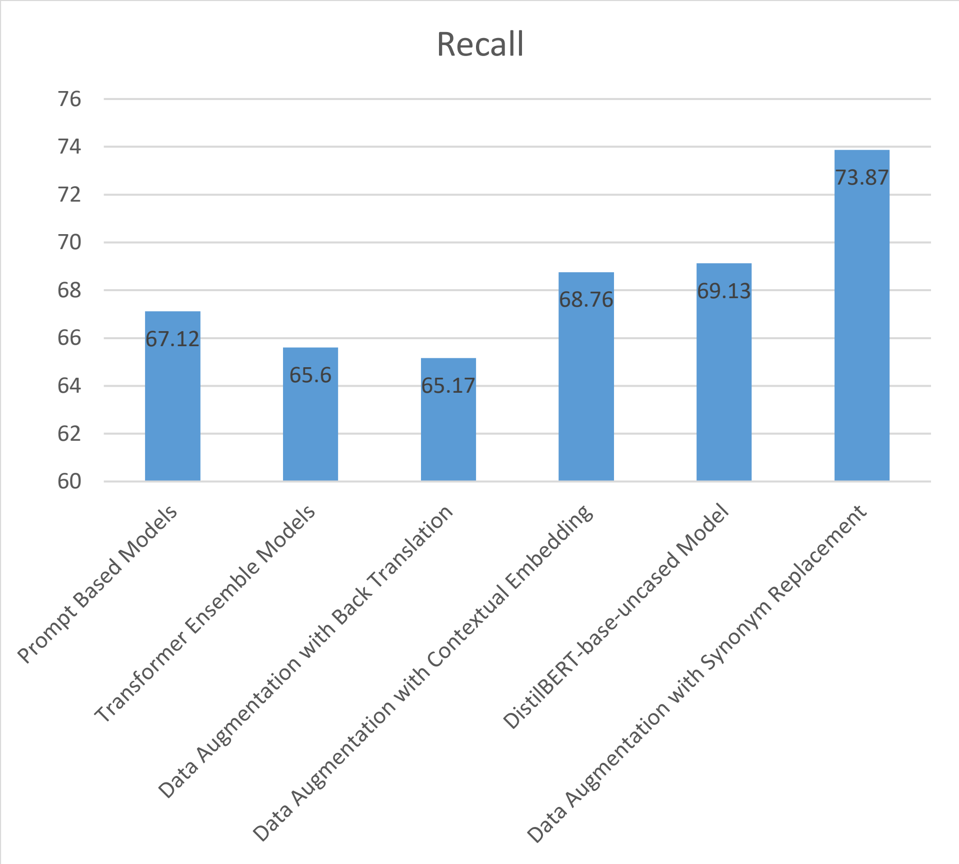

Figure 3: Recall score of proposed and comparing models

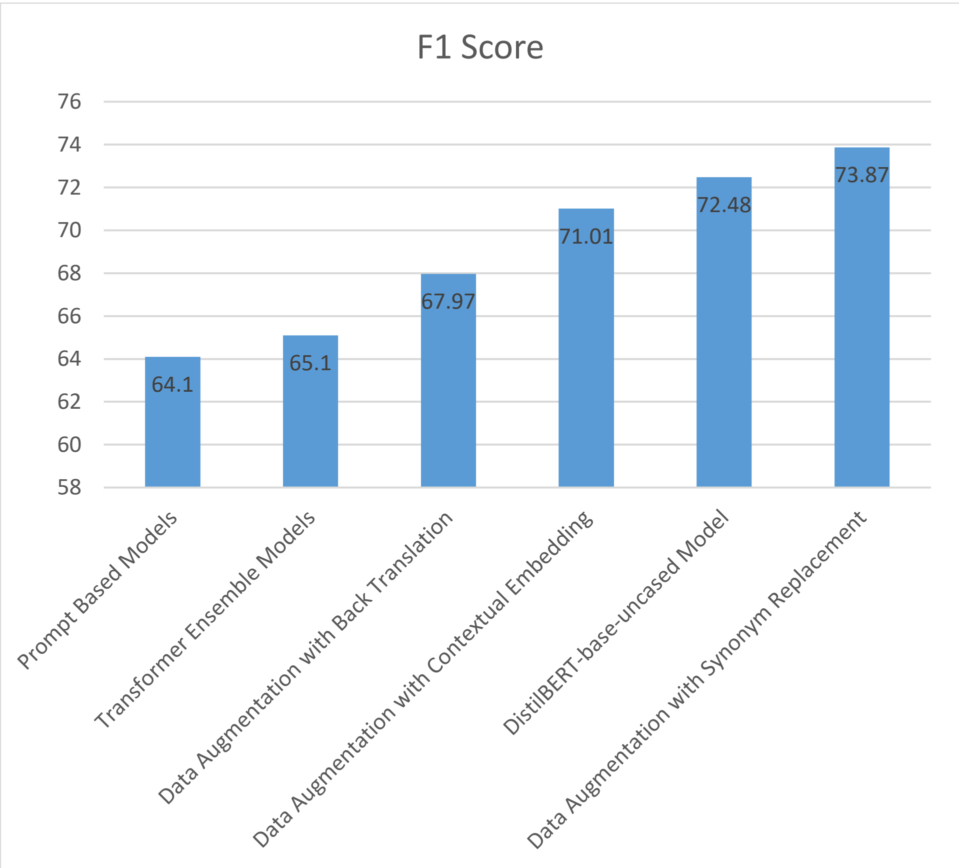

Figure 4: F1 score of proposed and comparing models

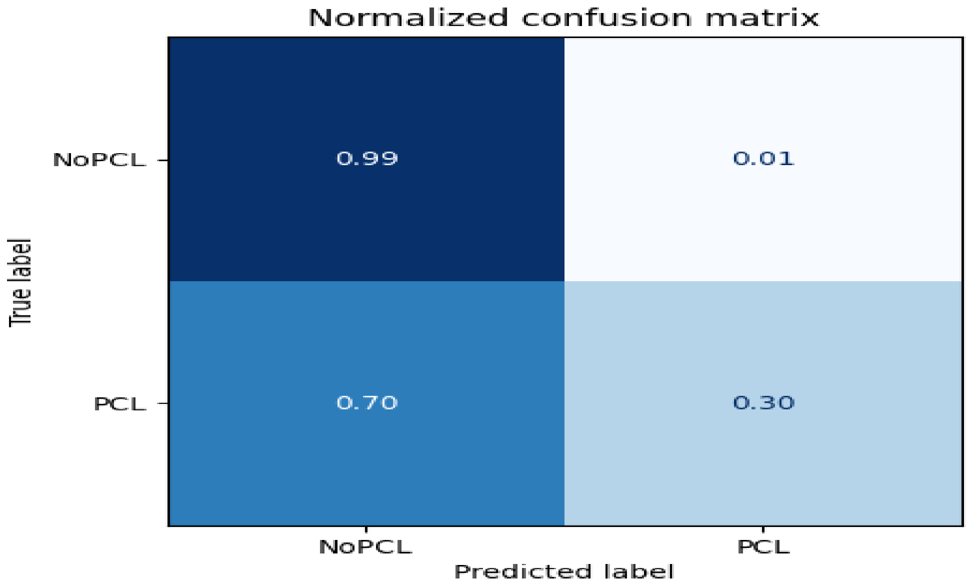

Figure 5: Confusion Matrix of DistilBERT model

6 Conclusion and Future Work

Patronizing and Condescending Language Detection is a Binary Text Classification problem. We proposed a fine-tuned DistilBERT pre trained model and Data Augmentation. Our models have performed well and achieved the best score than the previous models. Data augmentation with synonym replacement method also achieved a good precision score of 78.12. We have used back translation English to German and German to English. Data augmentation with Contextual Word Embedding also shown a better performance. In future large case pre-trained models can be applied on this dataset, also a different back translation language can be applied. In our proposed model due to high threshold it is able to detect true positives, there is a chance to miss some of the true positives due to high sensitivity. This also can be addressed in the further improvements.

References

1. Paula Fortuna and Sergio Nunes. 2018. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4):1–30.
2. Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In Canadian Conference on Artificial Intelligence, pages 16–27. Springer.
3. Ray Oshikawa, Jing Qian, and William Yang Wang.2018. A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770.
4. Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5):1–40.
5. Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4.
6. Preslav Nakov, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluis Marquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification

and verification of political claims. In International conference of the cross-language evaluation forum for European languages, pages 372–387. Springer.

7. Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Check-worthiness. In CEUR Workshop Proceedings, Lugano, Switzerland.
8. Alberto Barron-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! At clef 2020: Enabling the automatic identification and verification of claims in social media. In European Conference on Information Retrieval, pages 499–507. Springer.
9. Giovanni Da San Martino, Alberto Barron-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1377–1414.
10. 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). In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 75–86.
11. Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi aradzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Caˇgrı Coltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In Proceedings of the FourteenthWorkshop on Semantic Evaluation, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.
12. Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63.
13. Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 69–76.\
14. Sik Hung Ng. 2007. Language-based discrimination: Blatant and subtle forms. Journal of Language and Social Psychology, 26(2):106–122.
15. Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. Frontiers in Artificial Intelligence, 3:55.
16. Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. American psychologist, 48(6):621.
17. David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. Global Media and Communication, 9(1):53–70.
18. Michel Foucault. 1980. Power/knowledge: Selected interviews and other writings, 1972-1977. Vintage.
19. Katherine M Bell. 2013. Raising Africa?: Celebrity and the rhetoric of the white saviour. PORTAL Journal of Multidisciplinary International Studies, 10(1).
20. Rolf Straubhaar. 2015. The stark reality of the 'white saviour'complex and the need for critical consciousness: A document analysis of the early journals of a freirean educator. Compare: A Journal of Comparative and International Education, 45(3):381–400.

21. Lilie Chouliaraki. 2010. Post-humanitarianism: Humanitarian communication beyond a politics of pity. International journal of cultural studies, 13(2):107– 126.
22. Branka Drljaˇca Margi´c. 2017. Communication courtesy or condescension? linguistic accommodation of native to non-native speakers of english. Journal of English as a lingua franca, 6(1):29–55.
23. Howard Giles, Susan Fox, and Elisa Smith. 1993. Patronizing the elderly: Intergenerational evaluations. Research on Language and Social Interaction, 26(2):129–149.
24. Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. Discourse studies in composition, 155:176.
25. Lilie Chouliaraki. 2006. The spectatorship of suffering. Sage.
26. Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.
27. Carla Perez-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.
28. Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In Association for Computational Linguistics.
29. Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. Frontiers in Artificial Intelligence, 3:55.
30. Naitian Zhou and David Jurgens. 2020. Condolences and empathy in online communities. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 609– 626.
31. Dou Hu, Mangyuan Zhou, Xiyang Du, Mengfei Yuan, Meizhi Jin, Lianxin Jiang, Yang Mo and Xiaofeng Shie. 2022. PALI-NLP at SemEval-2022 Task 4: Discriminative Fine-tuning of Transformers for Patronizing and Condescending Language Detection. In Proceedings of 16th International Workshop on Semantic Evaluation (SemEval-2022). Pages 335-343.
32. Yong Dong, Chenxiao Dou, Liangyu Chen, Deqiang Miao, Xianghui Sun, Baochang Ma and Xiangang Li. 2022. BEIKE NLP at SemEval-2022 Task 4: Prompt-Based Paragraph Classification for Patronizing and Condescending Language Detection. In Proceedings of 16th International Workshop on Semantic Evaluation (SemEval-2022). Pages 319-323.
33. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
34. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprint arXiv:1901.11196.
35. Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. arXiv preprint arXiv:1805.06201.
36. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1606.04164.