# Balancing the Scales: Text Augmentation for Addressing Class Imbalance in the PCL Dataset

Bolukonda Prashanth[1], P Vijayapal Reddy[2]

[1]Research Scholar, Department of CSE, Osmania University, Hyderabad, Telangana, India

2Professor , Department of CSE, Keshav Memorial College of Engineering, Hyderabad, Telangana, India..

**Abstract:** Identifying Patronizing and Condescending Language (PCL) is an essential critical task in natural language processing (NLP), relevant to content moderation, bias identification, and online discourse evaluation. The existence of substantial class imbalance in PCL dataset is a considerable difficulty for machine learning models, frequently resulting in biased predictions and inadequate generalization. This study investigates text augmentation methods to address class imbalance and enhance the efficacy of PCL classifiers. The suggested method analyses news stories from different nations and determines whether or not they use patronizing or condescending language and categorizes the detected PCL into various groups like 1) Unbalanced power relations(UPR), 2) Shallow solution(SS), 3) Presupposition(PS), 4) Authority voice(AV), 5) Metaphor(MP), 6) Compassion(CP) and 7) The poorer, the merrier(TPTM). This work use deep learning (DL) techniques to address this issue, approaching it like a typical multi label text classification problem. This work utilizes back-translation, synonym substitution, contextual embedding's to improve the representation of minority classes. The findings of this work indicate that purposeful augmentation boosts model resilience and improves memory for underrepresented categories. Additionally, assess the influence of augmentation on multiple categorization architectures, encompassing transformer-based model like DistilBERT. The findings underscore the efficacy of augmentation in mitigating label discrepancies, thus enhancing the equity and accuracy of PCL detection systems.

**Keywords**: Patronizing and Condescending Language, Text Augmentation, Class Imbalance, Data Balancing, NLP, Transformers

## 1. INTRODUCTION

The current proliferation of social media usage has facilitated an increase in patronizing and condescending language (PCL). Patronizing and condescending language manifests acts that seem benevolent or considerate, although they ultimately disclose a sense of superiority over others. Despite PCL being an inadequately explored study domain thus far, detrimental linguistic behaviours such as hate speech [1], offensive language [2], disinformation [3], rumor spreading [4], and others have been thoroughly examined in NLP.

Identifying PCL can be tough for humans due of its intricacy and subjectivity. For instance, what one individual perceives as condescending may be regarded by another as an impartial representation of the situation, while some individuals may not perceive any issue in elucidating how those in privileged positions allocate their

surplus to those in need. Furthermore, those belonging to a so-called vulnerable community may anticipate experiencing greater patronization compared to those who are not part of such a community.

The purpose of SemEval 2022-Task 4 is to develop a system capable of identifying texts that include PCL and determining their presence or absence. The condescension is conveyed through the PCL category. The organizers provided two datasets: one containing PCL classifications and the other featuring annotations based on PCL intensity. Various authors have developed alternative models within this setting and obtained significant results. Our research introduced a pre-trained model, Distil-Bert, for the detection of PCL inside a specified text. Data augmentation has also been recommended to mitigate the problem of class imbalance. This work focuses on the task of PCL detection and categorization.

| **Example** | **Category of PCL** |
|---|---|
| "We can be extremely proud of the current women winemakers" | Unbalanced power relations |
| "The inclusion of a refugee team" | Shallow Solution |
| "An immigrant to a developed country lives in two worlds" | Presupposition |
| "women must wake up" | Authority voice |
| "trapped in the prison of poverty" | Metaphor |
| "more than 400 suspected asylum seekers are awaiting their fate" | Compassion |
| "how talented disabled people can be" | The poorer, the merrier |

Table 1: Examples PCL for each category.

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, this work wrapped up the report and offered some suggestions for more research in the Section number 6.

## 2. RELATED WORK

The examination of inequitable, misleading, or unpleasant language has attracted the interest of many experts within the NLP research community. The primary task in this context involves identifying explicit, aggressive, and notable occurrences, such as detecting misinformation or performing fact-checking [5,6,7,8]. Additional tasks encompass the detection of propaganda strategies [9], modeling of objectionable language [10,11], identification of hate speech [12], and examination of rumor spread [13]. Nonetheless, there exist other more subtle yet equally harmful kinds of language that have not been subjected to as much examination by the NLP community. These types, owing to their nuanced traits, are likely to be more difficult to recognize. An instance of this is the identification of Patronizing and Condescending Language (PCL), which constituted the primary focus of Task 4 at SemEval-2022.

Patronizing or condescending communication transpires when an entity's discourse conveys a sense of superiority over others. When beliefs are normalized, they institutionalize prejudice, rendering it less conspicuous [14]. Furthermore, the use of PCL is sometimes inadvertent and well-intentioned, especially in

conversations on underprivileged communities. The existence of this favourable attitude, termed good will, can exacerbate the detrimental effects of PCL. The audience, lacking robust opposition, is more vulnerable to the adverse effects of this discriminating rhetoric, often without awareness.

Sociolinguistic study defines PCL as a subtle form of language that is often inadvertent yet yields negative and discriminatory consequences[15]. The process engenders and sustains biases [16], resulting in heightened marginalization [17], the spread of rumors, and the circulation of misinformation [18]. PCL tends to promote power-knowledge dynamics, promoting charitable behaviors above cooperation and depicting people who offer assistance as rescues of individuals in disadvantaged positions [19,20]. Moreover, PCL often obscures the individuals or groups responsible for entrenched societal difficulties, occasionally by directly or indirectly attributing blame to marginalized communities or individuals for their situations. Furthermore, it often depends on transient and superficial solutions [21]. Privileged groups are known to employ PCL, linked to the notion of "pornography of poverty". This communication technique depicts vulnerable circumstances using a language of pity to provoke charitable actions and empathetic responses from the target audience.

Despite the comprehensive analysis of the detrimental impacts of PCL on social interactions and economic and political discourse within social sciences, it remains a largely unexamined issue in the domain of NLP. Nonetheless, It is believed that the identification of PCL poses numerous substantial challenges for NLP research. Consequently, additional research in this domain is essential, especially given the prospective societal benefits that may ensue. Given its subjective and nuanced nature, one should expect that identifying PCL will be more difficult than tasks focused on more overt phenomena. Moreover, the identification of PCL often demands an implicit understanding of human ethics and values, necessitating a form of logical reasoning that NLP models are likely to struggle with.

PCL has been extensively analysed in sociolinguistics by Margić[22], Giles et al.[23], Huckin[24], and Chouliaraki[25], as previously mentioned in the introduction. The examination of patronizing discourse has received limited attention within the domain of NLP. Wang and Potts[26] diverged from conventional methods by compiling a selection of Reddit comments, deliberately selected for their condescending tone, and annotated accordingly. It is noteworthy that, in contrast to the SemEval goal, their study did not explicitly focus on underrepresented communities. In the previous study by Carle Perez[27], titled Don't Patronize Me!, which appears to be the first annotated compilation of Patronizing Language (PCL) aimed at at-risk communities. This database functioned as the training dataset for the SemEval task. Related research has scrutinized discourse styles closely linked to condescension. Sap et al. [28] examined the correlation between specialized language usage and power dynamics. Mendelsohn[29] investigated the depersonalization of marginalized groups via language, whereas Zhou and Jurgens[30] analyzed the interplay between sympathies and empathy articulated in online networks with authoritative voices.

## 3. DATASET

The principal source material for this endeavour is a dataset titled "Don't Patronize Me!" (DPM). It is a curated compilation of terminology that is both patronizing and condescending towards oppressed individuals. This dataset was first introduced in prior research by Perez-Almendros et al. in 2020[27]. The dataset comprises 10,469 paragraphs, which constituted the training set for the SemEval assignment. To create the test set for this study, the authors methodically annotated an additional 3,898 paragraphs, following the same procedure. News items from which the paragraphs were derived were sourced from media outlets in twenty English-speaking countries. The primary source of these materials is the News onWeb (NoW) corpus, as indicated by Davies in 2013. This dataset is utilized with the authors' permission [27] for research purposes.

## 4. IMPLEMENTATION

### 4.1 Preprocessing

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

1. Initially, This work convert the Label column into a binary representation. If the value is either zero or one, then convert it to zero. If the number is two or three, then 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, this work transformed the column to binary in order to ascertain the presence of PCL, without considering the extent of PCL.

2.all null value attributes from the dataset have been dropped, 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, it is transformed that the text into encoding and added with 2 more columns input_ids and attentionmasks to extract the features of datasets. Subsequently, the encoded sequence is sent into the model to execute the classification process.

### *4.2 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. The proposed method also applied the method of data augmentation to improve the results as the dataset has the class imbalance issue. 3 different data augmentation approaches discussed below.

4.2.1 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 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.

4.2.3 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. This work has used English to German and German to English back translation.

**4.3 PCL Detection**

***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.

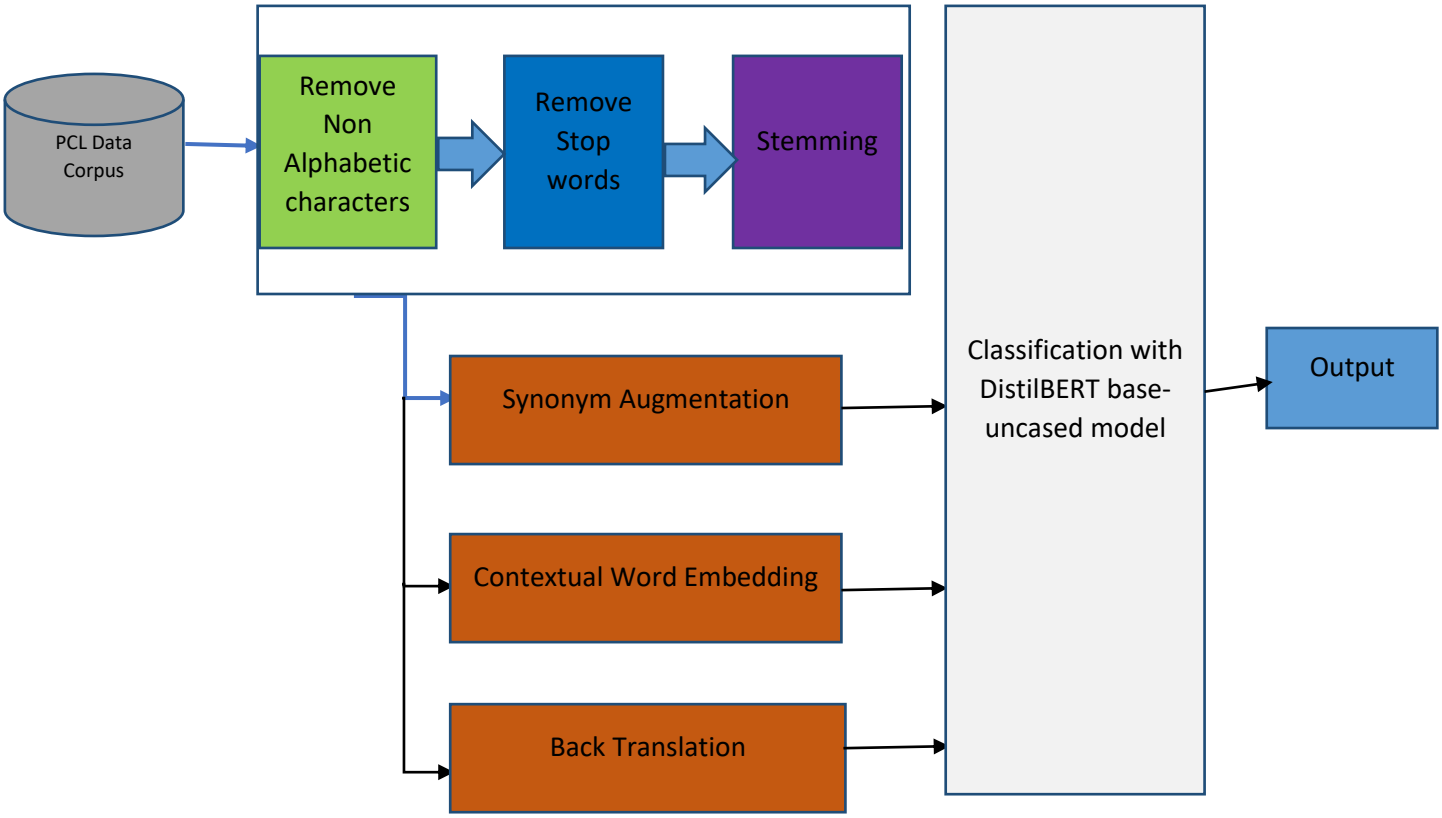


Figure 1: Proposed Model with Data Augmentation

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.

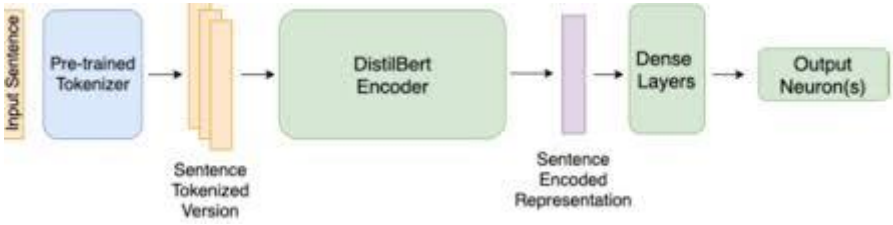


Figure 2: DistilBERT model

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.

**4.4 PCL Categorization**

***4.2.1 DistilBERT:*** The process of categorization is executed by utilizing a meticulously adjusted pre-existing distilBERT model with base uncased. The DistilBERT model with base uncased is a popular variant of the BERT (Bidirectional Encoder Representations from Transformers) family. It is specifically designed to be smaller, faster, and more efficient, while yet keeping a substantial percentage of BERT's performance. Understand the complexities of this paradigm and grasp its characteristics and use. Hugging Face's team introduced DistilBERT as a method to create more efficient and effective versions of the BERT model. The goal is to attain a balanced equilibrium between the model's performance and its computing efficiency.

DistilBERT achieves this by reducing the number of layers and parameters compared to BERT, while employing a technique known as "knowledge distillation". The uncased technique is typically preferred when case sensitivity is not essential, as it simplifies the tokenization process. DistilBERT consists of 6 transformer layers and utilizes 66 million parameters, which is a smaller number compared to BERT's 110 million parameters. Figure 2 depicts the architectural design of the DistilBERT model, which incorporates Data Augmentation.

Despite its modest size, DistilBERT demonstrates competitive performance across a variety of NLP benchmarks. It achieves a performance level of around 97% compared to BERT on tasks such as the GLUE benchmark, making it an appealing option for resource-constrained applications. The training procedure entails the utilization of knowledge distillation, a technique in which a smaller model called DistilBERT is trained to mimic a larger model known as BERT. This method allows the student model to capture a substantial part of the teacher's knowledge in a more concise form. The proposed model is shown in the Figure 1. The model was trained using the following hyper parameters, they are, learning rate of 2e-5, batch size=8 and epoch=3. Which are provided with a comprehensive outcome.

## 5 RESULTS AND DISCUSSION

Proposed model DistilBERT-base-uncased for detecting PCL has performed well. This work used the metrics Precision, Recall and F1-Score to judge our model performance. Proposedmodel achieved the precision score 76.18, Recall score 69.13and F1 score 72.48 at testing phase which shows the increase in the metrics with comparing models. This work compare 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].

This work proposes 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.

This work proposes 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. proposed work 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.

It is here by 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. it 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 3, 4, and 5 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.

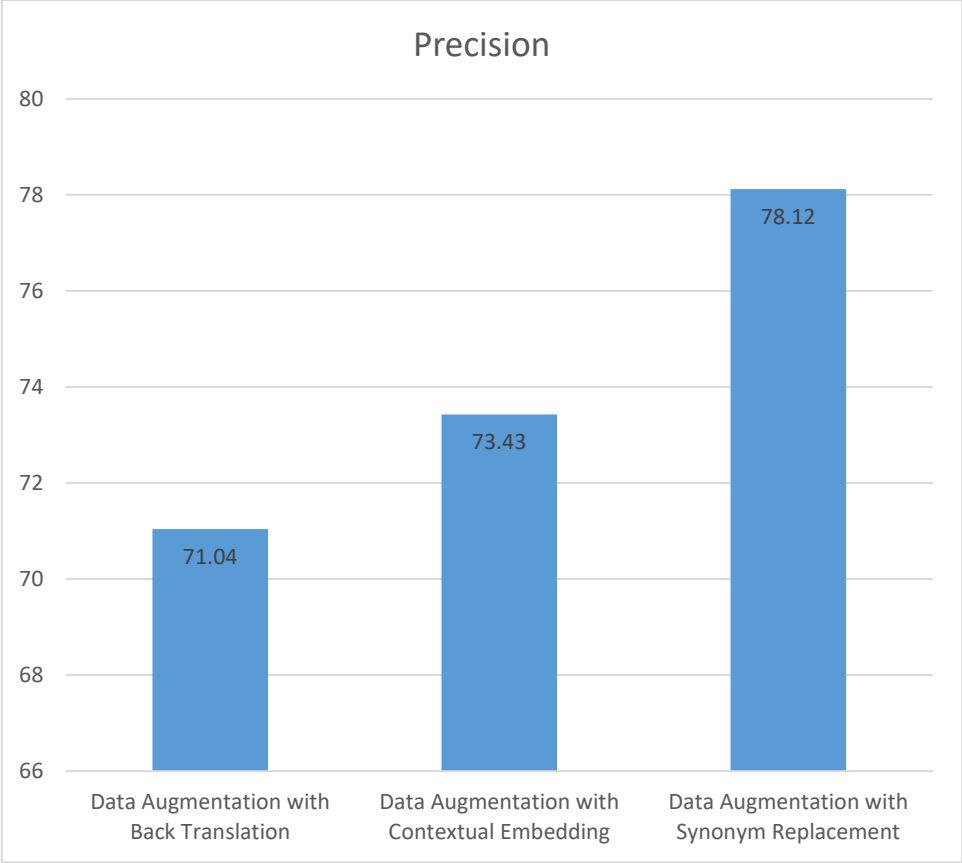


Figure 3: Precision score of proposed and comparing models

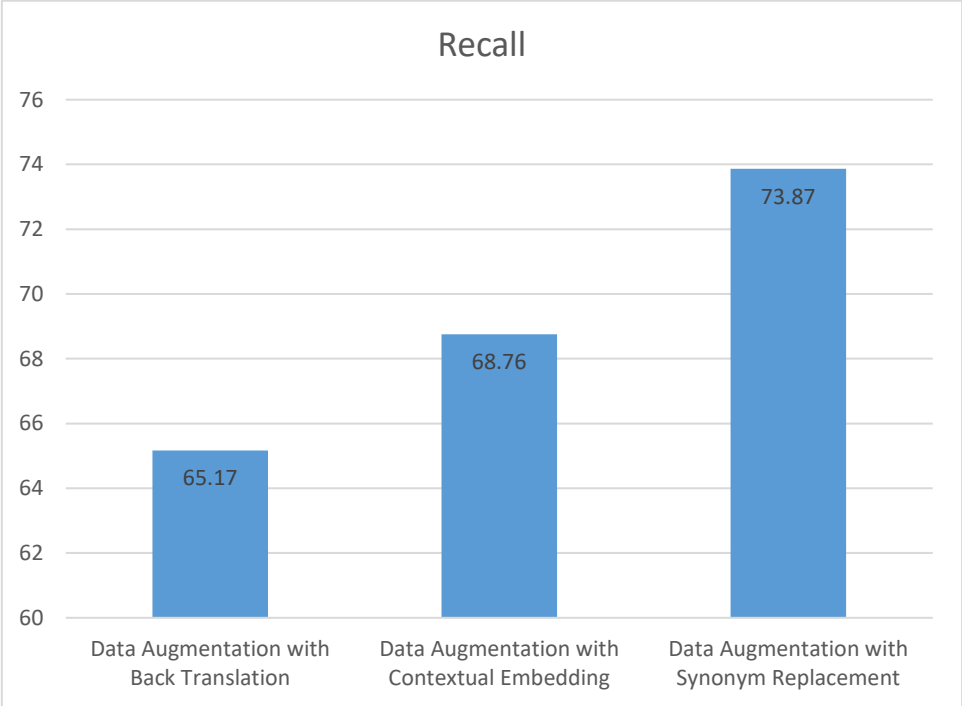


Figure 4: Recall score of proposed and comparing models

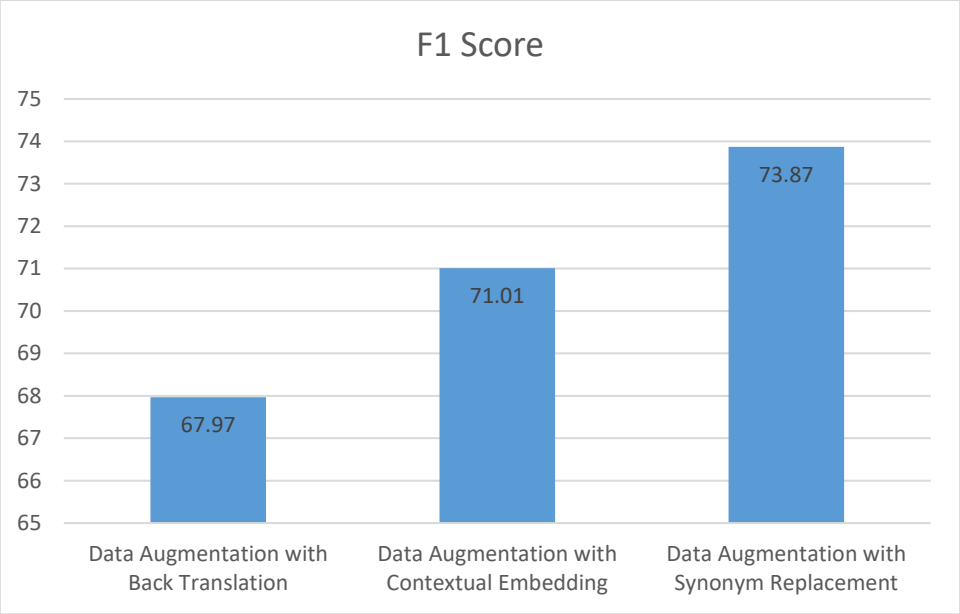


Figure 5: F1 score of proposed and comparing models

Proposed model DistilBERT-base-uncased with data augmentation for categorization PCL has performed well. The metric macro F1-score used to judge proposed model performance. Proposed model achieved the macro F1-score 49.26, precision score 52.95 and recall score of 46.10 at testing phase which given the increase in the metrics with comparing models. This work compares proposed model with prompt based models [37] has achieved Macro F1 Score 46.90 and Prompt training and label attention mechanism [38] has achieved Macro F1 score 43.90. Proposed model received a favourable score when compared to other criteria.Transformer models of larger size generally exhibit slower inference speeds compared to DistilBERT, which is particularly crucial for processing real-time data. During the training phase, larger models generally undergo fine-tuning at a slower pace compared to DistilBERT, resulting in greater benefits. DistilBERT effectively transfers acquired information throughout the distillation process across larger transformer models [39].

DistilBERT is typically fine-tuned on specific datasets for text classification tasks, enabling it to capture task-specific patterns and nuances. When comparing these fine tuning approaches to all purpose techniques. In prompt-based models, the level of accuracy is often higher. DistilBERT-base-uncased is significantly more resource-efficient than large prompt-based models. Modifying DistilBERT does not necessitate intricate rapid engineering or dynamic alterations. Instead, it employs a straightforward approach to alter the weights of the model by taking into account task-specific inputs. DistilBERT can outperform prompt-based model and prompt training and label attention mechanisms in terms of training and data requirements [39].

It has been proposed data augmentation with synonym replacement, and obtained an Macro F1 score 49.26, Macro recall score of 46.10 and Macro precision score of 52.95. When comparing prompt-based learning with transformer-based models, data augmentation with synonym replacement emerges as the most successful strategy for text categorization across many significant factors. This comprehensive comparison highlights the benefits of using synonym replacement in the process of text classification. The training dataset effectively increases in both size and diversity without requiring more labelled data by replacing terms with their

equivalents. Enhancing the model's capacity to generalize to novel instances may lead to a decrease in overfitting. Synonym replacement [40] is an uncomplicated technique, which may be easily included into the data pre-processing pipeline and requires minimal computational resources. This stands in sharp opposition to the computational load associated with using large prompt-based models for generating responses, or training enormous transformer models. Synonym substitution augmentation of data does not necessitate pre-existing models. This method can be applied to a diverse array of languages and subjects. When compared to large transformer models or systems that rely on prompts, using augmented datasets with synonym substitution is more convenient for scaling and can be used with smaller models that are easier to construct.

It has been proposed data augmentation with contextual word embedding, and obtained Macro F1 score 47.80, a Macro recall score of 44.74 and Macro precision score of 51.38. It is conducted an investigation on several significant attributes of data augmentation using contextual word embedding [41]. This research's findings indicate that this approach may offer superior performance for text classification compared to prompt-based learning and transformer-based models. Models like BERT or similar ones are used to generate contextual word embedding, which enhance the realism and semantic richness of text data changes. Consequently, less complex models may achieve superior performance and rival more advanced tactics. Contextual word embedding, when employed for data augmentation, provides comprehensive and meaningful representations of words within their specific contexts, hence enhancing generalization and performance. Contextual embedding can be enhanced to capture certain nuances in domain-specific data, leading to more effective data augmentation. This is particularly useful for specialist applications such as detecting patronizing and condescending language. By including contextual embedding into the data, the model gains increased resistance to variations in language and adversarial inputs, leading to improved performance.

It has been proposed another model data augmentation with back translation, and obtained Macro F1 score 47.34, a Macro recall score of 44.16 and Macro precision score of 51.06. Data Augmentation with Back Translation [42,43,44] is considered superior to Transformer-based models and prompt-based learning for text classification due to several crucial properties. Back translation generates diverse and linguistically nuanced data variations. This strategy would enhance the resilience and efficiency of the model without increasing the computational burden of more complex models. Back translation involves translating text into another language and then translating back into the original language and generates paraphrases that incorporate a wide range of vocabulary and sentence constructions. This enhances the model's capacity to apply its knowledge to other linguistic patterns. It is particularly advantageous for capturing a broader spectrum of phrases, as it maintains the basic sense of the text while retaining natural variations in language, as opposed to mere substitution of synonyms.

The suggested models showed improved performance compared to standard transformer-based models and prompt-based models, as well as Prompt training with label attention mechanism, when the DistilBERT-base-uncased model was combined with varied data augmentation strategies. The metrics considered in the comparing models are Macro F1 Score. This work also considers the same along with Macro Precision and Recall Score. Specifically, macro F1 score improved by 5.03%, indicating a reduction in false positives. Macro

Precision and Macro Recall saw a good score, showcasing better detection of true positives. Consequently, the Macro F1-score rose to 5.03% significantly, reflecting an overall balanced and enhanced performance. The improvements can be attributed to the model's efficient architecture and enhanced ability to generalize from enriched data, resulting in more dependable and precise predictions.

Table 2, 3, 4, 5 and 6 shows the score of proposed and comparing models. Figure 6, 7, 8, 9 and 10 compare the precision, recall, and F1-score of all the proposed models and comparing models that are being presented.

| | **Patronizing and Condescending Language Categories** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **UPR** | **SS** | **PS** | **AV** | **MP** | **CP** | **TPTM** | **Macro Avg** |
| **F1** | 65.75 | 53.92 | 37.98 | 41.86 | 34.97 | 52.98 | 47.12 | 47.80 |
| **Precision** | 69.99 | 55.84 | 40.33 | 47.25 | 39.39 | 56.47 | 50.36 | 51.38 |
| **Recall** | 61.99 | 52.13 | 35.89 | 37.58 | 31.44 | 49.90 | 44.28 | 44.74 |

Table 2: Scores of proposed Data Augmentation with Text Embedding

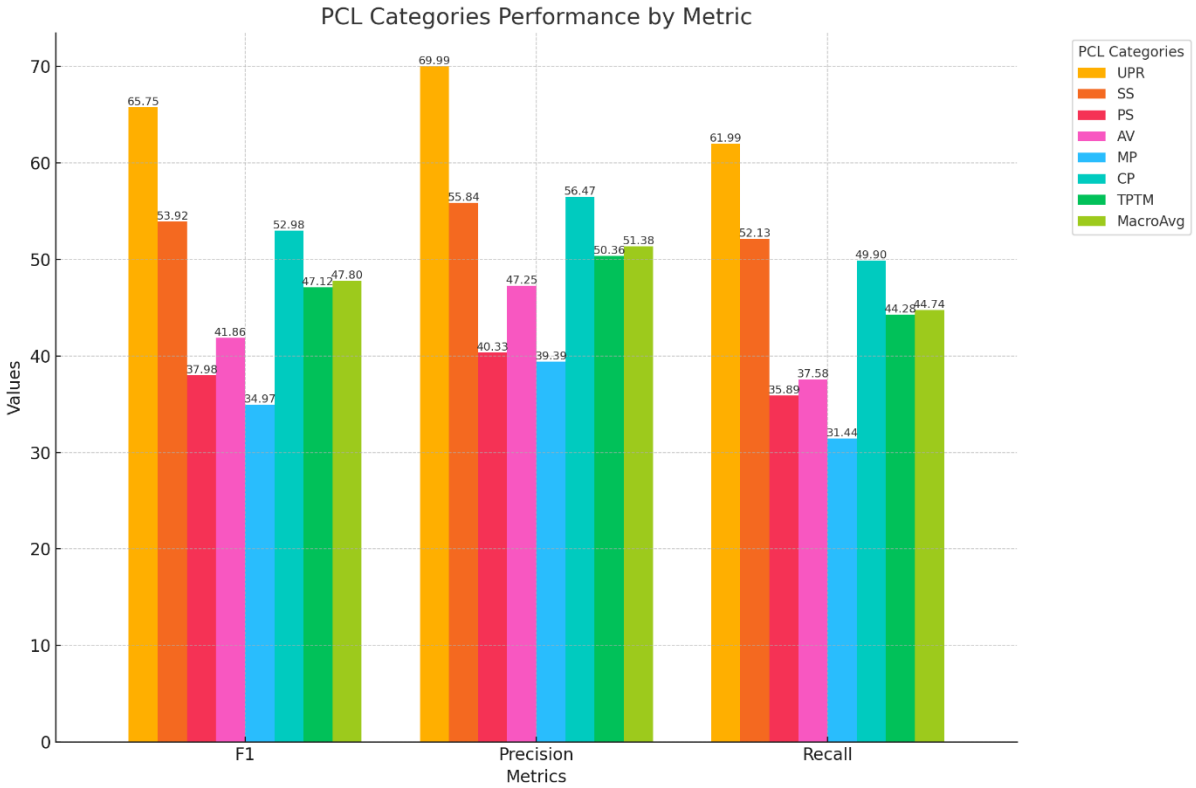


Figure 6: Metrics score of Fine-tuned DistilBERT with Data Augmentation Contextual Word Embedding

| | **Patronizing and Condescending Language Categories** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **UPR** | **SS** | **PS** | **AV** | **MP** | **CP** | **TPTM** | **Macro Avg** |
| **F1** | 65.98 | 54.21 | 38.45 | 42.58 | 42.87 | 53.48 | 47.24 | **49.26** |
| **Precision** | 70.49 | 56.14 | 41.40 | 48.60 | 46.81 | 56.56 | 50.62 | **52.95** |
| **Recall** | 62.01 | 52.41 | 35.89 | 37.88 | 39.54 | 50.71 | 44.28 | **46.10** |

Table 3: Scores of proposed Data Augmentation with Synonym Replacement

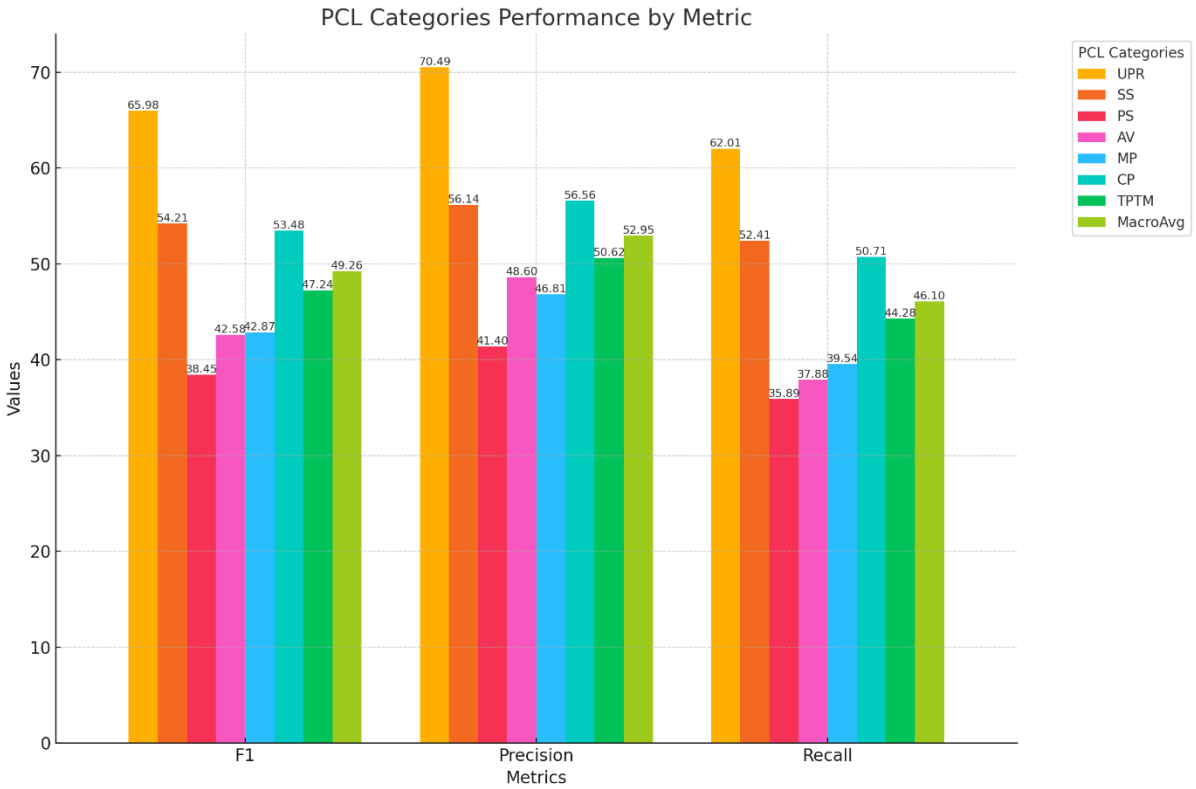


Figure 7: Metrics score of Fine-tuned DistilBERT with Data Augmentation Synonym Replacement

| Metric | Patronizing and Condescending Language Categories | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **UPR** | **SS** | **PS** | **AV** | **MP** | **CP** | **TPTM** | **Macro Avg** |
| **F1** | 64.74 | 55.48 | 37.15 | 41.47 | 34.22 | 53.18 | 45.15 | 47.34 |
| **Precision** | 68.75 | 59.14 | 40.40 | 46.60 | 38.11 | 56.27 | 48.18 | 51.06 |
| **Recall** | 61.17 | 52.25 | 34.39 | 37.36 | 31.05 | 50.41 | 42.48 | 44.16 |

Table 4: Scores of proposed Data Augmentation with Back Translation

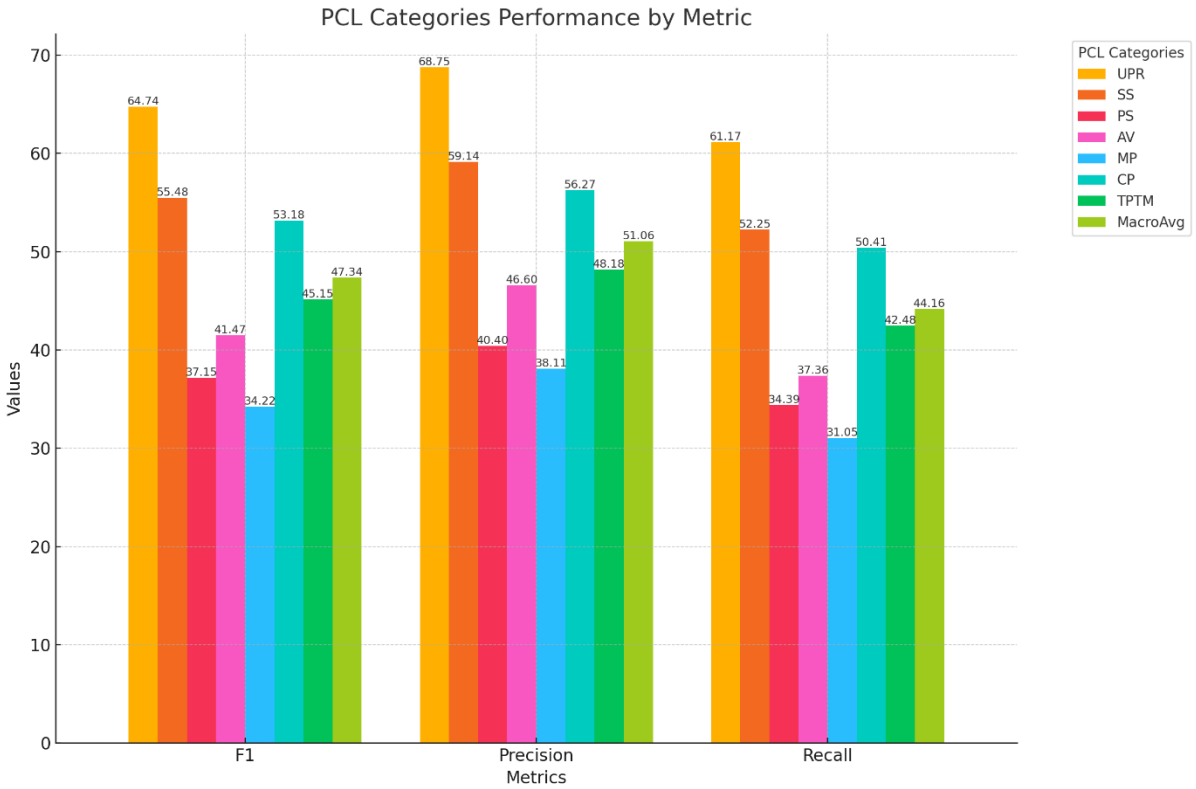


Figure 8: Metrics score of Fine-tuned DistilBERT with Data Augmentation back translation

| Models | **Model Name** | **UPR** | **SS** | **PS** | **AV** | **MP** | **CP** | **TPTM** | **Macro Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Proposed Models | Data Augmentation Synonym Replacement | 65.98 | 54.21 | 38.45 | 42.58 | 42.87 | 53.48 | 47.24 | **49.26** |
| | Data Augmentation Contextual Word Embedding | 65.75 | 53.92 | 37.98 | 41.86 | 34.97 | 52.98 | 47.12 | 47.80 |
| | Data Augmentation Back Translation | 64.74 | 55.48 | 37.15 | 41.47 | 34.22 | 53.18 | 45.15 | 47.34 |

Table 5: Scores(macro F1) of Proposed Models for all the categories

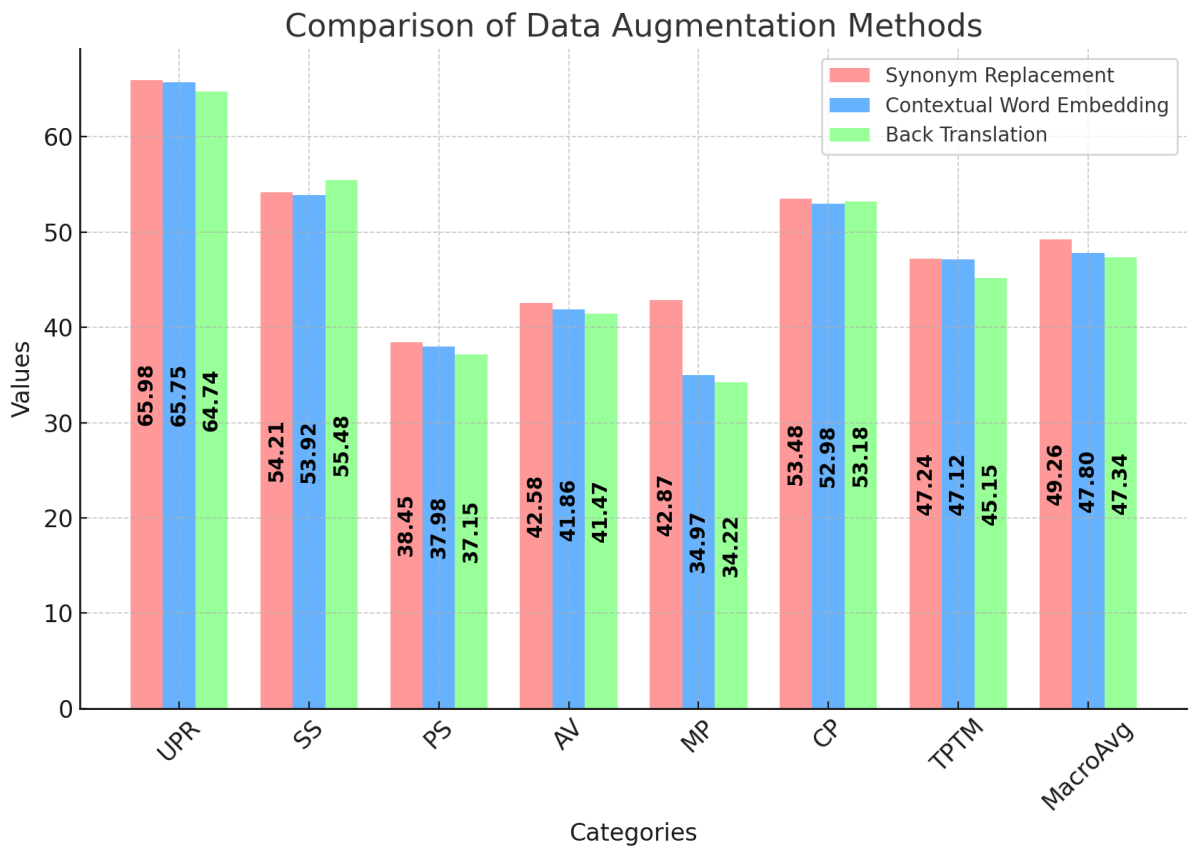


Figure 9: Macro F1 score of proposed and compared models

## 6 CONCLUSIONS AND FUTURE WORK

Patronizing and Condescending Language Detection is a Binary Text Classification problem. This work proposed a fine-tuned DistilBERT pre trained model and Data Augmentation. Proposed 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. This work has used backtranslation 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 backtranslation language can be applied. In 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. PreslavNakov, Alberto Barron-Cedeno, Tamer Elsayed, ReemSuwaileh, Lluis Marquez, WajdiZaghouani, PepaAtanasova, 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. PepaAtanasova, PreslavNakov, GeorgiKaradzhov, MitraMohtarami, 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, PreslavNakov, Giovanni Da San Martino, MaramHasanain, ReemSuwaileh, 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, RostislavPetrov, and PreslavNakov. 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, ShervinMalmasi, PreslavNakov, Sara Rosenthal, NouraFarra, 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, PreslavNakov, Sara Rosenthal, PepaAtanasova, Georgiaradzhov, Hamdy Mubarak, Leon Derczynski, ZesesPitenis, and Ca˘grıColtekin. 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, ElisabettaFersini, 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, KalinaBontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and ArkaitzZubiaga. 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, YuliaTsvetkov, 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 AkinaMikami. 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. LilieChouliaraki. 2010. Post-humanitarianism: Humanitarian communication beyond a politics of pity. International journal of cultural studies, 13(2):107– 126.
22. BrankaDrljaˇcaMargi´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. LilieChouliaraki. 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, YuliaTsvetkov, 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, MeizhiJin, Lianxin Jiang, Yang Mo and XiaofengShie. 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.
37. 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.
38. Ye Wang, YanmengWang, Baishun Ling, Zexiang Liao, Shaojun Wang and Jing Xiao. (2022). PINGAN Omini-Sinitic at SemEval-2022 Task 4: Multi-prompt Training for Patronizing and Condescending Language Detection. In Proceedings of 16th International Workshop on Semantic Evaluation (SemEval-2022). Pages 319-323.
39. 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.
40. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprint arXiv:1901.11196.
41. Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. arXiv preprint arXiv:1805.06201.
42. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. arXiv preprint arXiv:1606.04164.
43. Bolukonda Prashanth, Dr P Vijaya Pal Reddy. (2024). “Use of a refined Distil-BERT model and Data Augmentation to Identify Patronizing and Condescending Language”, African Journal of Biological Sciences, vol.6, no. 10, pp.575805768, June, 2024. *https://doi.org/10.48047/AFJBS.6.10.2024.5758-5768*

44. NandulaAnuradha, Dr P Vijaya Pal Reddy. (2024). "Deep Hybrid models with BERT for Cross Domain Suggestion Classification", African Journal of Biological Sciences, vol.6, no. 4, pp.1076-1087, June, 2024. https://doi.org/ 10.33472/AFJBS.6.4.2024.1076-1086.
45. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method" Computers and Electrical Engineering , 101 (2022) 107960
46. Ramdas Vankdothu,.Mohd Abdul Hameed "Adaptive features selection and EDNN based brain image recognition on the internet of medical things", Computers and Electrical Engineering , 103 (2022) 108338.
47. Ramdas Vankdothu,.Mohd Abdul Hameed,Ayesha Ameen,Raheem,Unnisa " Brain image identification and classification on Internet of Medical Things in healthcare system using support value based deep neural network" Computers and Electrical Engineering,102(2022) 108196.
48. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning" Measurement: Sensors Journal,Volume 24, 2022, 100440 .
49. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor MRI images identification and classification based on the recurrent convolutional neural network" Measurement: Sensors Journal,Volume 24, 2022, 100412 .
50. Bhukya Madhu, M.Venu Gopala Chari, Ramdas Vankdothu,.Arun Kumar Silivery,Veerender Aerranagula " Intrusion detection models for IOT networks via deep learning approaches " Measurement: Sensors Journal,Volume 25, 2022, 100641
51. Mohd Thousif Ahemad ,Mohd Abdul Hameed, Ramdas Vankdothu" COVID-19 detection and classification for machine learning methods using human genomic data" Measurement: Sensors Journal,Volume 24, 2022, 100537
52. S. Rakesh [a], NagaratnaP. Hegde [b], M. VenuGopalachari [c], D. Jayaram [c], Bhukya Madhu [d], MohdAbdul Hameed [a], Ramdas Vankdothu [e], L.K. Suresh Kumar "Moving object detection using modified GMM based background subtraction" Measurement: Sensors ,Journal,Volume 30, 2023, 100898
53. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Efficient Detection of Brain Tumor Using Unsupervised Modified Deep Belief Network in Big Data" Journal of Adv Research in Dynamical & Control Systems, Vol. 12, 2020.
54. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Internet of Medical Things of Brain Image Recognition Algorithm and High Performance Computing by Convolutional Neural Network" International Journal of Advanced Science and Technology, Vol. 29, No. 6, (2020), pp. 2875 – 2881
55. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Convolutional Neural Network-Based Brain Image Recognition Algorithm And High-Performance Computing", Journal Of Critical Reviews,Vol 7, Issue 08, 2020(Scopus Indexed)
56. Ramdas Vankdothu, Dr.Mohd Abdul Hameed "A Security Applicable with Deep Learning Algorithm for Big Data Analysis",Test Engineering & Management Journal,January-February 2020