PromptAug: Data Augmentation for Fine Grained Conflict Identification

Anonymous ACL submission

Abstract

Following the garbage in garbage out maxim, the quality of training data supplied to machine learning models impacts their performance. Generating these high-quality annotated training sets from unlabelled data is both expensive and unreliable. Moreover, social media platforms are increasingly limiting academic access to data, eliminating a key resource for NLP research. Consequently, researchers are shifting focus towards text data augmentation strategies to overcome these restrictions. In this work, we present an innovative data augmentation method, PromptAug, using Large Language Models (LLMs). We demonstrate the effectiveness of PromptAug, with improvements over the baseline dataset of 2% accuracy and 5% F1-score. Furthermore, we evaluate PromptAug over a variety of dataset sizes, proving it's effectiveness even in extreme data scarcity scenarios. To ensure a thorough evaluation of data augmentation methods we further perform qualitative thematic analysis, identifying four problematic themes with augmented text data; Linguistic Fluidity, Humour Ambiguity, Augmented Content Ambiguity, and Augmented Content Misinterpretation.

1 Introduction

Social media has exploded in popularity throughout society, as social media usage increases, the volume of interactions on social platforms also increases. A significant number of these interactions are negative. These negative interactions can have substantial harmful consequences for users. In order to reduce these consequences the negative interactions first have to be detected. While existing work focuses on detecting extreme forms of these negative interactions (Fortuna and Nunes, 2018; Alkomah and Ma, 2022; Poletto et al., 2021), less extreme negative interactions have still been shown to cause harm to users (Boroon et al., 2021; Kowalski, 2000; Wang et al., 2022; Ledley et al., 2006).

In this paper, we focus on fine grained identification of negative interactions, and cast the problem as complex multi-class classification comprising of a range of negative behaviours. This task is studied by Breitsohl et al. (2018) using netnographies (Kozinets, 2015), proposing a unique dataset that demands a model capable of discerning between six distinct conflict behaviors, Table 1. The imbalanced dataset showcases typical overlapping human behavior classes with blurred boundaries due to shared traits (Lango and Stefanowski, 2022).

A key to successful classification models, esp. in the era of neural models, is access to robust large-scale training data (Minaee et al., 2021; Fenza et al., 2021). Datasets are commonly obtained by collecting and annotating data from platform APIs, frequently utilizing annotation services such as MTurk (Aguinis et al., 2021). However, this approach has a number of faults. Platforms such as Facebook and X(Twitter) have restricted academic access to research data, placing access beyond reach or behind a paywall, which many researchers cannot afford. Additionally, whilst these services provide opportunities to easily produce labelled data many researchers have questioned the quality of the produced data (Welinder and Perona, 2010; Paolacci et al., 2010). These issues with labelling quality are magnified when dealing with the highly nuanced behavioural data present within our problem. There also exists a layer of ethical concern whereby data annotators are repeatedly exposed to negative and harmful content (Roy et al., 2023).

Data augmentation (DA) presents a solution to this issue and is a growing NLP research area (Shorten et al., 2021; Soudani et al., 2023). Researchers can use DA to expand datasets and increase reliability and performance of models, while preventing over-fitting to limited training data. A large number of DA methods are centered around substitution augmentation; e.g., synonym swapping, sentence manipulation, and word insertion or

Class	Size	Description
Teasing	208	Humorous communication without hostile intent (light jokes, banter, friendly provocation, mild irony that can be misunderstood).
Sarcasm	577	Humorous communication in a cynical tone (biting, bitter, hurtful tone, including swearwords)
Criticism	698	Constructive communication without hostile intent (superiority, factual disagreements, without humorous elements)
Trolling	1089	Provocative communication without targeting anyone (edging conflicts on, inciting anger, seeking disapproval, obvious fake news and misinformation, seeking response)
Harassment	1098	Abusive communication with hostile intent (including swearwords, profanities, discriminatory language; and no humorous elements)
Threats	482	Abusive communication with declared intention to act in a negative manner

Table 1: The classes in this paper's conflict dataset, the number of datapoints in each class, and their definitions.

reordering (Fellbaum, 2010; Wei and Zou, 2019). Conversely, other data augmentation techniques aim to generate entirely new datapoints (Anaby-Tavor et al., 2020; Yang et al., 2020; Quteineh et al., 2020). These models often rely on expensive state-of-the-art LLMs and require pre-training.

We argue existing NLP techniques are limited in the variety and depth of generated datapoints for conflict classification task. It is shown that substitution based methods, while easy to implement, offer incremental improvements with little diversity between the original and generated datapoints (Feng et al., 2021). They often do not retain datapoint identity and can change the context, legibility, and label preservation of datapoints. Figure 1 exhibit this behaviour in two examples generated by a text transmutation DA method, EDA (Wei and Zou, 2019). In the first example, there is a lack of legibility, and the context of singling a user out for negativity is lost. In example two, the substitution of two words completely changes the tone and subsequent datapoint class.

As highlighted there are a multitude of shortcomings with the existing text DA methods for conflict classification task. In this paper, we propose a novel text DA method focusing on distinct LLM prompting techniques. We leverage open-source LLMs to generate new text datapoints which adhere to class identities and retain class boundaries. Specifically we make use of two open sources LLMs; Llama by Meta (Touvron et al., 2023), and Mistral by Mistral AI (Jiang et al., 2023). We generate high quality, creative text datapoints, expanding the training dataset whilst adhering to class definitions and boundaries. Our approach features a designed prompting scheme, consisting of four distinct components: instruction, context, examples, and definition (Table 2). We show the effectiveness of our DA approach by evaluating the generated data intrinsically and extrinsically.

To summarize, we make the following contributions in this paper:

- We study the critical task of fine-grained multi-class conflict classification and propose a prompt-based data augmentation method to address challenging properties of the task such as class imbalance and blurred class boundaries inherent to this task.
- We show that our DA method outperforms state-of-the-are substitution and LLM-based data augmentation methods and is highly robust under extreme data scarcity conditions.
- We perform an extensive analysis of the synthetically generated data, quantitatively by measuring lexical diversity, and qualitatively using human annotations, identifying four traits in mis-annotated datapoints.

These findings are of considerable importance in an academic landscape, where access to social media research data is becoming more restricted and the quality of available data is under scrutiny.

2 Related Work

EDA (Wei and Zou, 2019) is a widely used and referenced DA method, employing four operations; synonym replacement, random insertion, random swap, and random deletion. EDA demonstrated increased performance across a variety of classification tasks and restricted dataset sizes.

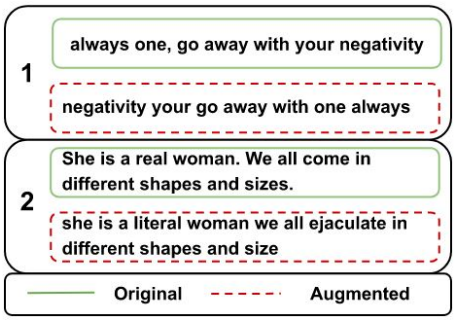

Figure 1: Example EDA datapoints, showing a lack of legibility in "1" and change of context and label in "2".

CBERT (Wu et al., 2019), is based on a BERT model where an additional label-conditional constraint is applied to the model task. The BERT model then creates augmented data whilst retaining contextual label information. CBERT showed increased performance in multiple classification tasks compared to baselines and other NLP DA methods.

Lambada (Anaby-Tavor et al., 2020), is based on generating additional datapoints using an LLM then filtering the data using a classifier that is pretrained on the original data to ensure quality data. The filtration works via the classifiers confidence score for each class, with the algorithm retaining the top N samples where the models classification matches the true label of the datapoint. However, filtering via classification model could introduce bias into the training dataset.

PromptMix (Sahu et al., 2023), is based on generating new datapoints near class boundaries using GPT3.5-turbo. The method generates mixed class datapoints, then uses the same LLM to relabel them to ensure correctness of labels. Whilst this method achieved SOTA performance in the paper, it presents a number of challenges in the conflict classification task, which features blurred boundaries and similar behaviour classes, introducing more examples along these blurred boundaries is not only difficult for LLMs to achieve due to the nuanced behaviours but also serves to reinforce the ambiguity within the dataset instead of providing more clarity. Additionally, the baseline classification performance in this task is low, due to the reasons mentioned in section 1. Therefore, the relabelling step within PromptMix is likely to increase rather than decrease the number of incorrect labels within the augmented dataset.

Outside of NLP classification tasks, Whitehouse et al. Whitehouse et al. (2023) explore the use of prompt formatting DA to improve performance in multilingual commonsense reasoning datasets. They make use of more powerful closed LLMs such as GPT-4, and identify that exploring open-source low resource LLMs, as we do in this paper, is a compelling direction for future work.

As a result of the problems identified in Section 1 and gap in related work identified here, we present a straightforward, easily implemented DA method. This approach is based on detailed prompt engineering for a low-resource LLM, harnessing the power of the LLM whilst removing the need for pre-training and specifically targeting augmentation with regards to class definition and identify. We evaluate the effectiveness of the DA method with respect to accuracy, f1-score, recall and precision over a variety of dataset sizes. We further perform qualitative thematic analysis over the augmented datapoints to verify their robustness.

3 Methodology

3.1 Defining the Method

For our DA approach we utilise an LLM to generate new datapoints which can be used to increase the size, diversity, and quality of training datasets. Firstly, we consider the set of classes C within the dataset. For each class $c \in C$, we divide the set of datapoints D_c into groups of size k. We set k=3. For each class c we also create a definition and additional adjectives and descriptors. Iterating through the set of classes C, for each group of examples within the class c, we prompt the LLM to generate 5 new examples belonging to class c in a numbered list. A breakdown of this prompt is shown in Table 2. Each section of the prompt was carefully designed and selected for a specific purpose.

3.2 Prompt Components

When constructing our prompt structure we adhered to the CLEAR framework (Lo, 2023), which emphasises five components; concise, logical, explicit, adaptive, and reflective. The first two prompt features, *instruction and context,* directly relate to the framework, applying it's principles.

The *instruction* delivers a clear directive to the LLM. We experimented with different versions of the instruction and found it important to specify the output format ('In a numbered list...'). If not, the LLM sometimes generates erroneous datapoints, which could be related to the behaviour or completely random. Similarly, specifying '...

Instruction	In a numbered list, write 5 new social media comments containing {behaviour}...
Context	... directed at other social media users.
Examples	Here are some examples; {Examples one, two, three}.
Definition	{Behaviour} is defined as {type of} communication {list of additional adjectives and descriptors}

Table 2: PromptAug prompt segments.

write 5 new social media comments containing behaviour...' limited the randomness of the prompt output and provided the best quality responses. These components of the prompt enabled pattern matching in order to obtain the generated examples.

For the *context* portion of the prompt, we applied various role-playing scenarios. If the phrases 'As a social media user' or 'In response to a social media comment' were used, the LLM would often output advice on how to respond to the behaviour, not the behaviour itself. Simply using '... directed at other users' provided the best results, we theorise that this provides the LLM with enough context without making it the focus of the prompt.

The use of desired behaviour *examples* is key to our method, without which the LLM relies solely on the definition for creating datapoints. Including examples tethers the LLM to the existing dataset, retaining the current class boundaries whilst simultaneously having the freedom to create additional datapoints. This reasoning is supported by results from PromptMix (Sahu et al., 2023), where authors evaluated few-shot and zero-shot generation. They found that in all cases, few-shot generation outperformed zero-shot.

Finally, a vital part of our method is the inclusion of a clear, distinct desired behaviour *definition* with additional adjectives and descriptors. With numerous possible definitions for each behaviour, it is crucial the LLM understands the exact version of the behaviour it is generating. Strong behaviour definitions and additional descriptors allow the LLMs to generate creatively within the desired scope, contributing to the retention of class boundaries and good datapoint quality.

4 Experiments

We design three experiments to answer the following research questions.

- RQ1. Do data augmentation methods increase classification performance?
- RQ2. Do data augmentation methods retain performance within data scarce scenarios?
- RQ3. Do data augmentation methods generate good quality and diverse datapoints?

4.1 Experiment One: Data Augmentation Effects on Classification Performance

To answer RQ1, we evaluate the classification results of CNN, DistilBERT, and BERT models trained on the original datasets, and synthetically generated data by PromptAug, PromptMix (Sahu et al., 2023), EDA (Wei and Zou, 2019), and CBERT (Wu et al., 2019) DA methods. We apply the PromptAug method as described and PromptMix, EDA, and CBERT methods according to their papers. The EDA and CBERT methods produce a 1:1 ratio of datapoints. Both PromptMix and PromptAug produce higher ratios of augmented data. To conduct a fair comparison we randomly sample the generated datapoints from the PromptMix and PromptAug DA methods until this 1:1 ratio is achieved. Each DA method had the same original data, the training datasets then consisted of the original and newly generated DA datapoints. For the LLM based methods, PromptAug and PromptMix, we also evaluate generalisability by examining the effect on classification performance using different LLMs for data generation. We test using Llama2-7B and Mistral-8B.

In order to further evaluate the results we also include a breakdown of class performance in two heatmaps. This allows the analysis of the effect of augmentation on an individual class level, seeking to find trends related to class size or characteristic.

4.2 Experiment Two: Performance of Data Augmentation in Data Scarce Scenarios

DA techniques are frequently employed when there is a lack of available training data. Therefore, it is vital that the augmentation method retains its ability to create quality datapoints with limited data. As a result, we restrict the volume of training data available to the augmentation methods to 20%, 40%, 60%, and 80%. This experiment demonstrates not only the effect of training dataset size on classi-

fication models, but also the effectiveness of our augmentation method in data scarcity scenarios.

4.3 Experiment Three: Quality Analysis of Augmented Datapoints

To answer RQ3, which focuses on generated datapoint quality, we analyse diversity within the datapoints generated by the DA methods. To this end, we follow the diversity evaluation outlined by Joko et al. (2024), employing two diversity metrics used in their work; Distinct-n (Dist-n) (Li et al., 2015) and Self-BLEU (Zhu et al., 2018). Dist-n evaluates the ratio of distinct unigrams and bigrams to the total numbers of unigrams and bigrams, respectively. Self-BLEU examines the diversity present within a corpus by calculating the BLEU score between each datapoint in the corpus and the rest of the corpus datapoints. To obtain the Self-BLEU score an average of all BLEU scores is taken. To obtain each BLEU score we use NLTK's BLEU methods, and set the weights for 1,2,3, and 4 n-grams to 0.25 each. Finally, prior to computing both of the diversity metrics we follow Joko et al. (2024)'s advice to employ normalisation. We randomly sample from each set of augmented datapoints until a set number of words is reached. Joko et al. (2024) note that without normalisation diversity metrics such as Dist-n are bias towards datasets with fewer words.

To qualitatively analyse data quality, we sampled 150 datapoints from the augmented EDA and PromptAug data and conducted a blind annotation by two researchers, one from outside the paper. We conduct percentage annotator agreement and calculate Cohen's Kappa statistic according to McHugh (2012). To evaluate trends and patterns in the mis-annotated generated datapoints we employ Thematic Analysis (TA), a widely used research method in the social science domain formally established by Braun and Clarke (2006). Additional work by Braun and Clarke (2021) outlines the six step process for TA that we follow; familiarisation of data, generate initial codes, identify themes, review themes, define themes, and report findings. One researcher coded the mis-annotated datapoints and identified themes, a second then reviewed the identified codes and themes. The researchers then discuss the codes, patterns, and themes before finalising findings, which are reported with the identified themes, definitions, descriptions, and examples included for robustness and reproducibility.

4.4 Implementation and Evaluation Setup

For classification model description and hyperparameters see Table 7 in the appendix. All models were standard implementations and were trained using the same setup over four epochs, learning rate of 2e-5, AdamW (Loshchilov and Hutter, 2017) for optimization, and Cross Entropy Loss. For each dataset size interval the same training (80%), validation (10%), and test (10%) sets were used, the only difference being the new generated data. Importantly, no augmentation occurred in the validation or test sets and the training set's augmented datapoints were based only on the original training set. This is vital to ensure no cross contamination between the train, validation, and test splits.

4.5 Dataset

The dataset used in this research was created by a sixteen-month netnography of four online Facebook brand communities (Breitsohl et al., 2018), where authors identified consumer conflicts and their different forms. Double coding was conducted by two social science researchers to ensure annotation integrity. The dataset, shown in 1, contains six conflict classes: Teasing, Criticism, Sarcasm, Trolling, Harassment, and Threats. After conducting the netnography some classes were severely lacking datapoints, we therefore supplemented the dataset with datapoints from other open source datasets. These additional data points from Khodak et al. (Sarcasm) (Khodak et al., 2017), Wulczyn et al. (Threats) (Wulczyn et al., 2017), and Aggarwal et al. (Trolling) (Aggarwal et al., 2020) were chosen because the annotation guides and descriptions within the papers for each of the classes aligned heavily with the characteristics and definitions of the classes within this paper. This practise has been supported by other research (Fortuna et al., 2018), (Salminen et al., 2020), which suggests that not only is the practise acceptable in terms of dataset robustness but can also lead to increased model performance.

5 Results and Discussion

5.1 Experiment One: Data Augmentation Effects on Classification Performance

Table 3 shows the effect of changing the LLM used for datapoint generation using the two implemented LLM-based DA methods, our method PromptAug and PromptMix. PromptAug improved over the baseline dataset when using both Mistral-8B and

		Acc	F1	R	P
Original Dataset		0.69	0.61	0.61	0.65
Llama2	PAug	**0.71**	**0.66**	**0.66**	**0.67**
	PMix	0.71	0.61	0.63	0.61
Mistral	PAug	**0.67**	**0.64**	**0.63**	**0.65**
	PMix	0.65	0.61	0.61	0.63

Table 3: BERT Classification performances for LLM-based DA methods using Llama2-7B and Mistral-8B.

Llama2-7B. PromptMix however, only achieved increased performances using Llama2-7B. We see that for both methods, Llama2-7B results in a stronger classification performance. We therefore use this LLM for further experiments.

Next we analyse classification performance of three models trained using the augmented datasets generated by the DA techniques. The results are displayed in Table 4, and show that PromptAug achieves best performance. Using BERT as classifier, both PromptAug and PromptMix achieve the same increase in accuracy over the original dataset. However, PromptMix achieves no increase in F1-score whilst PromptAug shows an increase of 5%. Additionally, PromptAug outperforms both EDA and CBERT in accuracy (3%) and F1-score (2%). Similar out-performance is present for CNN, PromptAug besting the original dataset by 5% accuracy and 6% F1-score, whilst scoring higher than EDA by 5% accuracy and 4% F1-score and higher than CBERT by 4% accuracy and 5% F1-score. The effects of DA are less evident but still present with DistilBERT, with PromptAug achieving the best performance in accuracy and joint highest performance in F1-score.

Results show PromptAug is an effective DA technique that can easily be used to improve classification performance. We highlight PromptAug's robustness by comparing performance against one SOTA and two common DA methods, and it's generalisability through increased performance over the original dataset using two different generative LLMs. Additionally, the lack of pre-training and ease of access means that PromptAug maintains a simple approach, enabling it's application to other tasks, only requiring an open source LLM, task instruction and context, existing class examples, and class definitions; elements that researchers will already have when constructing datasets.

Investigating class-wise performance, two heatmaps of BERT's classification performance across the original and PromptAug datasets are presented in Figure 2. We observe large performance increases of 0.15 within Teasing and Criticism classes, marginal performance increase in Trolling, and no performance increase in the Threat class. Despite an increase in overall performance, there were class performance decreases of 0.11 in Sarcasm and 0.05 in Harassment. Within the original dataset Teasing and Criticism were most frequently misclassified as Harassment. This trend was reduced across almost all classes after augmentation . We propose that PromptAug increased these classes' profiles, reinforcing their identities as separate behaviours to Harassment. This highlights the ability of PromptAug to be effective in scenarios with strong overlap between class boundaries and complex class behaviour. Class size could also be a contributing factor to performance. The smallest and worst performing class is Teasing, with the next smallest class being more than twice it's size. It therefore could have had the most to gain from an increase in datapoints. PromptAug more than doubled the Teasing class performance, demonstrating the effectiveness of PromptAug within a small, imbalanced multiclass dataset.

		Acc	F1	R	P
CNN	Original	0.45	0.40	0.40	0.43
	EDA	0.45	0.42	0.42	0.44
	CBERT	0.46	0.41	0.42	0.42
	PMix	0.49	0.42	0.42	0.42
	PAug	**0.50**	**0.46**	**0.46**	**0.48**
Distil	Original	0.65	0.55	0.57	0.54
	EDA	0.65	0.56	0.56	0.54
	CBERT	0.65	**0.57**	0.57	**0.56**
	PMix	0.64	0.56	0.57	0.55
	PAug	**0.66**	**0.57**	**0.59**	0.55
BERT	Original	0.69	0.61	0.61	0.65
	EDA	0.68	0.64	0.63	0.64
	CBERT	0.68	0.64	0.64	0.65
	PMix	**0.71**	0.61	0.63	0.61
	PAug	**0.71**	**0.66**	**0.66**	**0.67**

Table 4: DA methods classification performances. For LLM based methods Llama2-7B is used.

5.2 Experiment Two: Performance of Data Augmentation in Data Scarce Scenarios.

Experiment two evaluates the effect of DA methods under data scarcity conditions. Figure 3 shows that for the original dataset, classification performance worsens as dataset size decreases. The same is true for DA methods but at a lower rate, with DA

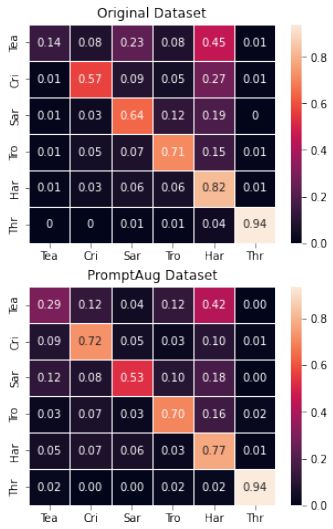

Figure 2: Class breakdown of BERT performance.

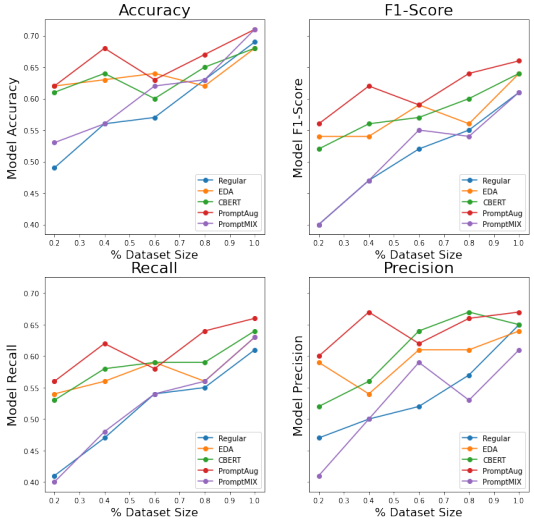

Figure 3: Line graphs of performance vs dataset size.

techniques reducing the impact of shrinking dataset size on performance. Of the methods, PromptAug improves the most over the original dataset. With accuracy increase of 13%, 12%, 6%, 4% and 2% over dataset sizes of 20%, 40%, 60%, 80% and 100%, respectively. This suggests that, for accuracy, DA is effective at all dataset sizes but has greater effect at lower dataset sizes. For F1-score, over the same size intervals, PromptAug improves over the baseline by 16%, 15%, 7%, 9%, and 5%. PromptAug therefore has greater impact on F1-score compared to accuracy at higher size intervals. PromptMix follows the same trend as the other DA methods with increased in accuracy over the original dataset, but does not follow the trend of increased F1-scores.

Concluding experiment two, as shown in Fig. 3, decreasing dataset size has an adverse effect on performance, this effect can be reduced using DA. PromptAug is the most effective DA technique, increasing Accuracy and F1-score performance at all dataset sizes with the exception of 60% where it is matched in F1-Score by EDA at 0.59 and outperformed in Accuracy by EDA by 1%. By demonstrating PromptAug's ability to effectively operate in data scarce scenarios we show its suitability for DA, where tasks seeking to employ DA are frequently struggling with extreme data scarcity.

5.3 Experiment Three: Quality Analysis of Augmented Datapoints

The diversity metric analysis in Table 5 highlights two findings. Firstly, that substitution based DA methods exhibit more diversity within the generated datapoints than LLM based methods. This can be attributed to substitution based augmented datapoints closely mirroring those present within the original dataset, they therefore retain the diversity present within the original data. Secondly, that of the two LLM based DA techniques PromptAug exhibts more diversity than PromptMix over Dist-1, Dist-2, and Self-BLEU.

The thematic analysis performed on misannotated datapoints from the EDA and PromptAug datasets produced four themes: (i) "Linguistic Fluidity," (ii) "Humour Ambiguity," (iv) "Augmented Content Ambiguity," and (vi) "Augmented Content Misinterpretation." Both EDA and PromptAug methods experienced Linguistic Fluidity and Humour Ambiguity.Augmented Content Ambiguity was identified within EDA data, and Augmented Content Misinterpretation was identified within PromptAug data (Table. 6). For the PromptAug, data annotators had an agreement rate of 67% and Cohen's K of 0.36, described as fair agreement by Landis and Koch (1977). For EDA, data annotators had an annotation agreement of 46% and Cohen's K of 0.14, described as slight agreement. Conducting TA to identify these themes provides an evaluation of DA beyond quantitative metrics. These themes can be used to target weaknesses that may be found in all NLP DA methods such as linguistic fluidity and humour ambiguity, or used to target specific weaknesses within methods such as augmented content ambiguity for EDA or augmented content misinterpretation for PromptAug.

The *Linguistic Fluidity* theme encompasses fluid or blurred boundaries between classes. Although

	Dist-1	Dist-2	Self-BLEU ↓
EDA	**0.131**	**0.636**	**0.122**
CBERT	0.104	0.534	0.132
Prompt Aug	0.114	0.482	0.453
Prompt Mix	0.070	0.252	0.662

Table 5: Diversity metrics for the DA models. ↓ indicates a lower result is better.

datapoints have dominant behaviours, they can contain aspects of multiple behaviours. Jhaver et al. (2017); Kim et al. (2022) identify ambiguous class boundaries investigating; how Criticism develops into Harassment, the inter-relation between the two behaviors, and subjectivity of true class identity. This theme is also present in hate research. Fortuna et al. (2020) discuss how terminology differs across the hate domain, leading to fluidity between behaviour classes in different datasets and misinterpretation of the behavioural identities.

The second theme, *Humour Ambiguity*, relates to the difficulty of identifying nuanced humour. Humour has been recognised as a challenging NLP area. It is largely subjective and often relies on subtle cues. For example, the first humour ambiguity datapoint in Table 6 belongs to 'Trolling' but was mis-annotated as 'Teasing.' There are two difficulties in identifying this datapoint. Firstly, the border between teasing and trolling behaviours can be subjective, what one individual finds humourous may incite a negative response from others. Secondly, humour is often nuanced, and as mentioned relies on subtle clues, DA within humourous behaviours may result in further ambiguity and blurring of class boundaries as words and phrases are altered.

The third theme, *Augmented Content Ambiguity*, relates to the DA method's ability to produce coherent augmented datapoints interpretable by humans, whilst retaining class labels. When human interaction behaviours are involved, class labels can depend on subtle text features, DA can obscure and sometimes remove vital clues for human coders. In the two given examples, we can observe that text transmutation has compromised the sentence composition, resulting in difficult interpretation for human coders. In their survey of NLP DA, Chen et al. (2023) note a similar problem of text transmutation changing the meaning of sentences.

The final theme, *Augmented Content Misinterpretation*, occurs within the PromptAug data. Although the prompt is designed to produce quality examples of desired classes, it occasionally produces erroneous responses, e.g., other negative classes, advice on dealing with the behaviour, and random data. These responses are difficult to filter and hinder model performance as they do not accurately reflect the desired classes. These erroneous responses are often a result of safety nets employed by the LLM, which are used to ensure safe AI practices. Other researchers identify this issue when generating negative behaviour datapoints. Lermen et al. (2023) investigated harassment and hate classes within their work, which is relevant to this paper's data. They found that Llama can refuse to produce harassment and hate examples around 75% and 70% of the time.

6 Conclusion

We present a novel few shot learning DA approach based on LLM prompting, targeting class definition and identity within a small, imbalanced negative behaviour multi-class dataset. Our augmentation method harnesses the power of LLMs while being easily implemented, requiring no finetuning, and achieving superior classification performance over the baseline dataset and other SOTA DA methods. We further demonstrate the effectiveness of the augmentation method in extreme data scarce scenarios. We further analyse quality of the generated data by evaluating diversity within augmented datapoints. In addition to the quantitative evaluation, we conduct a manual annotation and qualitative thematic analysis of the augmented datapoints. We find that within augmented datapoints there are four main themes of mis-annotation; linguistic fluidity, humour ambiguity, augmented content ambiguity, and augmented content misinterpretation.

Future Directions. With recent emphasis on responsible AI and growing focus on social bias within LLMs, future study could examine how bias presents itself within DA. A study adopting two methods suggested by Ferrara (2023), 'Applying fairness metrics' and 'Human-in-the-loop approaches', would provide insights on social bias of generated data. Secondly, quantifying expenses of DA methods would be of interest, highlighting trade-offs between expense and performance. Future work could also employ PromptAug within other text datasets, evaluating generalisability.

7 Limitations

We evaluate our model's generalisability across classification models and dataset size. Therefore we cannot make any assumptions about the generalisation of our method to other datasets with different classes and sizes. Additionally, we only use Llama-7B and Mistral-8B as generative LLMs for our method, so we cannot assume any generalisability for more powerful LLMs such as GPT4 or GPT3.5 turbo. We also do not investigate any social bias present within the datapoints generated by the LLM.

8 Ethical Concerns

In this paper we discuss harmful content, e.g. harassment and threats, and how to generate it using LLMs. This presents an opportunity for individuals with malicious intent to use this research to cause harm. We argue that the purpose behind this work is to improve classification performance for harmful content along a negative behaviour spectrum. This increased capability to successfully identify harmful content on social media is ultimately a net positive for society. In addition we don't specify any additional techniques to completely bypass LLMs safety nets, instead we only note that our prompt structure does do so to some degree.

References

Karmanya Aggarwal, Pakhi Bamdev, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, et al. 2020. Trawling for trolling: A dataset. *arXiv preprint arXiv:2008.00525*.

Herman Aguinis, Isabel Villamor, and Ravi S Ramani. 2021. Mturk research: Review and recommendations. *Journal of Management*, 47(4):823–837.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Layla Boroon, Babak Abedin, and Eila Erfani. 2021. The dark side of using online social networks: a review of individuals' negative experiences. *Journal of Global Information Management (JGIM)*, 29(6):1–21.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Virginia Braun and Victoria Clarke. 2021. Can i use ta? should i use ta? should i not use ta? comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research*, 21(1):37–47.

Jan Breitsohl, Holger Roschk, and Christina Feyertag. 2018. Consumer brand bullying behaviour in online communities of service firms. *Service Business Development: Band 2. Methoden–Erlösmodelle–Marketinginstrumente*, pages 289–312.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Francesco Orciuoli, and Enrique Herrera-Viedma. 2021. Data set quality in machine learning: consistency measure based on group decision making. *Applied Soft Computing*, 106:107366.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Shagun Jhaver, Larry Chan, and Amy Bruckman. 2017. The view from the other side: The border between controversial speech and harassment on kotaku in action. *arXiv preprint arXiv:1712.05851*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P de Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. *arXiv preprint arXiv:2405.03480*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.

Haesoo Kim, HaeEun Kim, Juho Kim, and Jeong-woo Jang. 2022. When does it become harassment? an investigation of online criticism and calling out in twitter. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–32.

Robin M Kowalski. 2000. "i was only kidding!": Victims' and perpetrators' perceptions of teasing. *Personality and Social Psychology Bulletin*, 26(2):231–241.

Robert V Kozinets. 2015. *Netnography: redefined*. Sage.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Mateusz Lango and Jerzy Stefanowski. 2022. What makes multi-class imbalanced problems difficult? an experimental study. *Expert Systems with Applications*, 199:116962.

Deborah Roth Ledley, Eric A Storch, Meredith E Coles, Richard G Heimberg, Jason Moser, and Erica A Bravata. 2006. The relationship between childhood teasing and later interpersonal functioning. *Journal of Psychopathology and Behavioral Assessment*, 28:33–40.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Leo S Lo. 2023. The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4):102720.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Husam Quteineh, Spyridon Samothrakis, and Richard Sutcliffe. 2020. Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7400–7410.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128.

Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam H Laradji. 2023. Promptmix: A class boundary augmentation method for large language model distillation. *arXiv preprint arXiv:2310.14192*.

Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2023. Data augmentation for conversational ai. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5220–5223.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Qiong Wang, Ruilin Tu, Yihe Jiang, Wei Hu, and Xiao Luo. 2022. Teasing and internet harassment among adolescents: The mediating role of envy and the moderating role of the zhong-yong thinking style. *International Journal of Environmental Research and Public Health*, 19(9):5501.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Peter Welinder and Pietro Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32. IEEE.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Wikipedia talk labels: Personal attacks. *URL https://figshare.com/articles/dataset/Wikipedia_Talk_ Labels_Personal_Attacks/4054689/6*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Appendix

Table 6: Themes identified in the annotation of augmentation methods generated datapoints.

Theme % Misclassified Comments	Definition	Description	Examples
Linguistic Fluidity	A miscoding of an augmented datapoint that occurs due to the lack of defintional boundaries that are inherent to the interpretation of language.	A commonly known phenomenon in linguistics is that of multiple meanings to the same sentence, where interpretation depends on a multitude of unpredictable factors(e.g. one's mood, need for politeness etc;) Classes are not always clear cut, often having fluid boundaries. Datapoints can contain behaviour which could belong to more than one class, making it difficult for annotators to get it totally accurate.	Coded - "Harassment, Actual Class - "Sarcasm" "I'm not sure what's more impressive: your ability to take a selfie or your lack of self-awareness..." Coded - "Harassment, Actual Class - "Trolling" "I can't stand this YouTuber's voice. It's like fingernails on a chalkboard every time they speak."
Humour Ambiguity	A miscoding of an augmented datapoint that occurs when a message fails to convey that it was meant in humour and/or was good vs bad-natured.	Linguists have long recognised the lack of clarity inherent to humour as a quality on which humour often relied. Humour has been recognised as a particularly challenging area of NLP. Humour can often be taken two ways and is subjective meaning the dominant type of humour behaviour is often ambiguous within datapoints.	Coded - "Teasing", Actual Class - "Trolling" "Your favorite meme is so last year, get with the times" Coded - "Teasing", Actual Class - "Trolling" "he makes that phone look like a tablet"
Augmented Content Ambiguity	A miscoding of an augmented datapoint that occurs due to a lack of clarity within the datapoint produced by the augmentation technique, where the content makes no coherent sense.	Within NLP DA label preservation is a known challenge, where class boundaries can depend on specific and nuanced words, phrases, and subtleties. Text transmutation such as synonym swapping/ insertion, word deletion, and reordering can change the context and legibility of datapoints, severely impacting label and datapoint behaviour preservation.	Coded - "Harassment", Actual Class - "Criticism" "ua warrior same if probably steph had the of type won commercial for would've" Coded - "Trolling", Actual Class - "Harassment" "do you rattling have sex microsoft i dont believe you have sex what are you speak about"
Augmented Content Misinterpretation	A miscoding of an augmented datapoint that occurs due to the augmentation technique misinterpreting the augmentation task.	Although LLMs can be given specific prompt instructions they do not always generate datapoints within the specified boundaries. Occasionally, instead of generating examples of the requested behaviour the LLM would instead produce examples of responses to that behaviour. These erroneous examples tend not to adhere to the characteristics of the class behaviour and can vary drastically in their identity.	Coded - "Criticism", Actual Class - "Trolling" "Lol @ you thinking you're relevant, get a life troll" Coded - "Criticism", Actual Class -"Trolling" "I'm so tired of username constantly posting memes that are offensive and disrespectful. Can't they see how their humor is affecting others? #harassment #block"

Table 7: Tables showing classification model hyperparameters and Descriptions.

Model	HyperParameters and Descriptions
BERT	For the BERT model, we used the HuggingFace transformers BERT-Base uncased pre-trained model with 12 layers, 12 heads, 768 hidden size, and 110M parameters.
DistilBERT	For the DistilBERT model we used HuggingFace DistilBERT model with 6 layers, 12 heads, 768 hidden size and 66M parameters.
CNN	The CNN model was created using TensorFlow Keras sequential model, and had 3 convolution layers, 3 pooling layers, a flatten layer used as connection between the Convolution layer, and two dense layers.

Table 8: Tables showing package versions and URLs.

Package	**Version**	**URL**
Huggingface Hub	0.20.3	https://huggingface.co/
Accelerate	0.26.1	https://huggingface.co/docs/accelerate
Transformers	4.35.2	https://huggingface.co/docs/transformers/
Torch	2.2.0	https://pypi.org/project/torch/
Pandas	1.5.3	https://pandas.pydata.org/
Numpy	1.25.2	https://numpy.org/
Sklearn	1.4.1	https://scikit-learn.org/stable/
Meta Llama	Llama-2-7b	https://huggingface.co/meta-Llama