PromptAug: Informed Prompt Engineering for Data Augmentation

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, focusing on the design of distinct prompt engineering techniques for Large Language Models (LLMs). We concentrate on Instruction, Context, Example, and Definition prompt attributes, empowering LLMs to generate high-quality, class-specific data instances without requiring pre-training. We demonstrate the effectiveness of PromptAug, with improvements over the baseline dataset of 2% accuracy, 5% F1-score, 5% recall, and 2% precision. 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 Research Contributions

We make the following contributions in this paper:

- Developed a prompt-based data augmentation method for enhancing social media data for multi-class conflict classification.
- Evaluate the quality of generated data using quantitative and qualitative methods.
- Evaluate the effect of dataset size on classification performance and data generation.

2 Introduction

Many machine learning models have been successfully applied to classification tasks (Minaee et al., 2021). Robust training datasets are required to achieve this high performance level (Fenza et al., 2021). In NLP, datasets are commonly obtained by collecting and annotating datapoints from platform APIs, frequently utilizing annotation services, e.g. MTurk (Aguinis et al., 2021). This approach has however been jeopardised, platforms such as Facebook and X(Twitter) have restricted academic access to research data, placing access either beyond reach or behind a paywall, which many researchers cannot afford. Additionally, researchers have questioned the quality of data produced by online data labelling services (Welinder and Perona, 2010). Whilst these services provide opportunities to easily produce labelled data many question the varying levels of accuracy and precision (Paolacci et al., 2010). Data augmentation (DA) presents a solution to this issue and is a growing NLP research area (Shorten et al., 2021). By using DA techniques researchers can expand datasets, increasing the reliability and performance of models while preventing over-fitting to limited training data.

Within image and vision, a variety of DA techniques exist such as rotations, color space augmenting, mixing images, etc. (Shorten and Khoshgoftaar, 2019). However, many of these techniques can't be applied to text DA which presents a more complicated challenge as class labels depend on nuanced relationships between characters, words, and sentences (Li et al., 2022). We argue existing NLP techniques are limited in variety and depth of generated datapoints or require extensive, expensive pre-training. A large number of DA methods center around rule based augmentation e.g. synonym swapping or sentence manipulation. These methods restrict the variety present in augmented datapoints. Feng et al. recognise that rule based DA

methods are easy to implement but offer only incremental improvements with small diversity in generated datapoints (Feng et al., 2021). Conversely, other augmentation techniques aim to train models using existing data and subsequently generate entirely new datapoints (Anaby-Tavor et al., 2020; Yang et al., 2020; Quteineh et al., 2020). However, these models are more expensive to implement and require a quality training set with a suitable number of datapoints, something which is rarely present in real world scenarios that require DA.

These method's problems are worsened when dealing with complex multi-class classification tasks surrounding human behaviours, e.g. the conflict task discussed in this paper. This task involves a compact dataset sourced from netnographies (Kozinets, 2015) compiled by Breitsohl et al. (Breitsohl et al., 2018), that demands a model capable of discerning between six distinct conflict behaviors shown in Table 2. These behaviors exhibit common traits, resulting in blurred class boundaries and identity. Lango and Stefanowski also identify class imbalance, inter-relation and overlapping as key contributors to the difficulties in small multi-class classification tasks (Lango and Stefanowski, 2022). Due to the small dataset size, augmentation methods requiring pre-training struggle to generate eligible datapoints. The problems with substitution based methods are also evident with nuanced behaviour data. Although these approaches tend to use techniques such as synonym selection via Wordnet (Fellbaum, 2010), they often do not retain datapoint identity. Performing text transmutation methods such as word swapping, insertion, or reordering can change the context, legibility, and label preservation of the datapoint. This is shown in the two EDA, a text transmutation DA method, datapoints in Fig. 1 (Wei and Zou, 2019). In augmented datapoint example one, 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. Instead of critiquing another user's viewpoint on a woman the datapoint is turned into an offensive trolling behaviour, this would however not be reflected in the datapoint label which would remain as criticism.

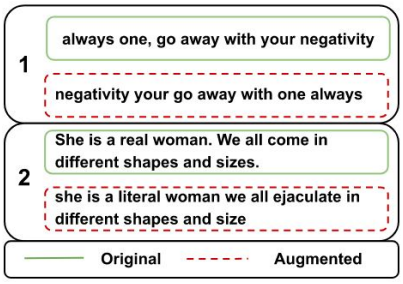

Figure 1: Example EDA Datapoints, showing a lack of legibility in "1" and change of context and label in "2".

3 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.

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), a DA method based on generating additional datapoints using an LLM then filtering the data using a classifier that is pre-trained 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.

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 2 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

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 1: Table showing the segments of the prompt.

perform qualitative thematic analysis over the augmented datapoints to verify their robustness.

4 Methodology

We sought to exploit LLM's power and their ability to generate coherent text. Specifically we make use of 7B-LLaMA (Touvron et al., 2023), a low resource open source LLM by Meta. We leverage prompt engineering to generate high quality, creative text datapoints, expanding the training dataset whilst adhering to class definitions and boundaries. We designed a prompting scheme, spliting the prompt into four distinct parts; instruction, context, examples, and definition, (Table 1).

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 generated erroneous datapoints, which could be related to the behaviour or completely random. Similarly, specifying '... write 5 new social media comments containing behaviour...' limited the randomness of the prompt output and provided the best quality responses.

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.

Using examples of the desired behaviour is key to our method. Without examples present the LLM relies solely on the definition for creating datapoints, by providing examples the LLM is tethered to the existing dataset, therefore retaining the current class boundaries whilst simultaneously having the freedom to create additional datapoints.

Finally, a vital part of our method is the inclusion in the prompt of a clear, distinct desired behaviour definition. With numerous possible definitions for each behaviour, it is crucial the LLM understands exactly what version of the behaviour it is generating. Strong behaviour definitions also contribute to the retention of class boundaries as the classes expand. We experimented with using adversarial definitions and/or class behaviours alongside the desired class definition. E.g. '...avoid the following behaviours; X, defined as ... and Y, defined as ..., etc'. Ultimately these didn't work, often confusing the LLM, leading it to produce new behaviour definitions or refusing to produce an output.

5 Experiments

5.1 Implementation, Hyperparameter Details and Metrics

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

5.2 Research Questions

- R.Q.1 Could employing data augmentation using prompts enhance the classification performance?
- R.Q.2 How does dataset size affect DA method performance?
- R.Q.3 Do the generated data points exhibit good quality?

5.3 Experiment One:

To answer R.Q.1, we evaluate the classification results of CNN, DistilBERT, and BERT models trained using the original, PromptAug, EDA, and CBERT datasets. We apply EDA and CBERT DA

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

Table 2: Table showing the dataset classes and their definitions.

methods as described in their papers with each original datapoint generating one additional datapoint. We apply our PromptAug method as described in our methodology with three original datapoints generating five additional datapoints. Due to the LLM producing unexpected outputs and occasionally refusing to produce negative content this results in roughly the same 1:1 ratio. Each augmentation method had the same original datapoints, the classifier training datasets then consisted of the original and newly generated DA datapoints.

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.

5.4 Experiment Two,

Answering R.Q.2, 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 the size of the training dataset has on classification models but also the effectiveness of our augmentation method in data scarcity scenarios.

5.5 Experiment Three,

Investigating R.Q.3, focuses on examining generated datapoint quality. Firstly, we produce a visualisation of augmented behaviour classes of PromptAug and EDA vs the original classes. To do this we apply t-SNE (Van der Maaten and Hinton, 2008) to the additional datapoints generated for each class by the DA methods, allowing us to plot a 2-d visualisation of the datapoints. This allows us to analyse how closely the newly generate datapoints resemble their original class counterparts.

We randomly selected 150 datapoints from the augmented EDA and PromptAug datasets and then conducted a blind annotation by two researchers, one from outside the paper. We conduct % annotator agreement and calculate Cohen's Kappa statistic according to McHugh (McHugh, 2012). To evaluate trends and patterns in the mis-annotated generated datapoints we employ thematic analysis. Formally established by Braun and Clarke (Braun and Clarke, 2006), thematic analysis is a widely used research method in the social science domain for identifying themes and patterns within a set of data, e.g. the DA method's generated datapoints. Additional work by Braun and Clark (Braun and Clarke, 2021) outlines the six step process for thematic analysis we follow in this work; familiarisation of data, generating initial codes, identifying codes, evaluating codes, reviewing themes, evaluating significance of themes, and reporting findings. One researcher coded the mis-annotated datapoints, a second researcher then reviewed the identified codes and themes. The researchers then discuss the codes, patterns, and themes before finalising the findings. These findings are then reported with the identified themes, definitions, descriptions, and examples included for robustness and reproducibility.

		Acc	F1	Rec	Pre
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
	PromptAug	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
	PromptAug	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
	PromptAug	0.71	0.66	0.66	0.67

Table 3: Table showing DA method's classification performances.

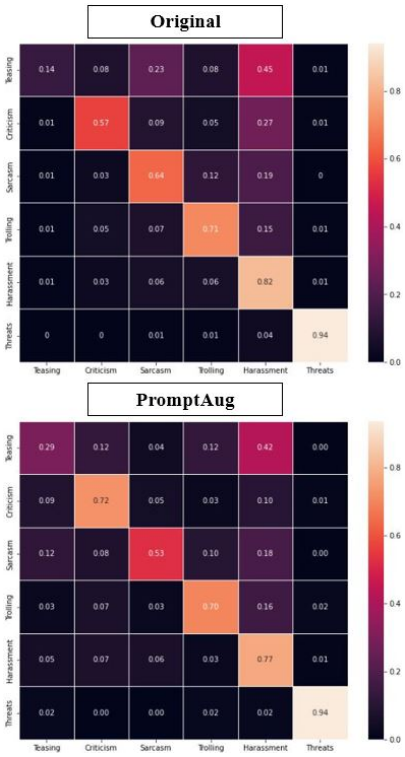

Figure 2: Heatmaps of class classification performance of BERT on the original and PromptAug datasets.

6 Results and Discussion

6.1 Experiment One

Experiment one shows that not only does PromptAug improve the classification performance of all the models trained on the original dataset but also outperforms other DA techniques. The BERT model trained on the PromptAug dataset outperforms the original dataset in accuracy (2%), F1-score (5%), recall (5%), and precision (2%). Additionally, it outperforms both EDA and CBERT DA methods in accuracy (3%) and F1-score (2%). Similar out-performance is present for CNN, PromptAug besting the original dataset by 5% accuracy, 6% F1-score, 6% recall, and 5% precision, 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 pronounced but still present with distilBERT.

These results show that PromptAug is an effective DA technique which can easily be utilised to improve classification model performance. By comparing performance against two SOTA DA methods we demonstrate PromptAug's robustness. Additionally, the lack of pre-training and ease of access means that Prompt Aug maintains a simple approach whilst improving performance. This enables the application of the technique to other tasks, only requiring an open source LLM, task instruction and context, existing class examples, and class definitions; all things that researchers will already have to hand when constructing datasets.

Investigating class-wise performance, the model trained with PromptAug is analysed and the results presented in two heatmaps (Fig. 2). We observe large performance increases of 0.15 within both Teasing and Criticism classes. We observe a marginal performance increase in Trolling whilst the Threat class performance remains the same. Interestingly, despite an increase in overall performance, Sarcasm and Harassment class performance decreased by 0.11 and 0.05 respectively.

Within the original dataset the most frequent misclassifications were Teasing and Criticism as Harassment. We propose that PromptAug increased these classes' profiles, reinforcing their identities as separate behaviours to Harassment. Class size could also be a contributing factor. Teasing is the smallest class within the imbalanced dataset, with the next smallest class being more than double it's size. It therefore could have had the most to gain from an increase in profile within the dataset.

To summarise, as shown in Fig. 2, the model originally struggled with Harassment misclassification. This was reduced across almost all classes after augmentation. This highlights the ability of PromptAug to be effective in scenarios with strong overlap between class boundaries and complex class behaviour. Furthermore, PromptAug more than doubled the Teasing class performance, demonstrating the effectiveness of PromptAug within a small, imbalanced multiclass dataset.

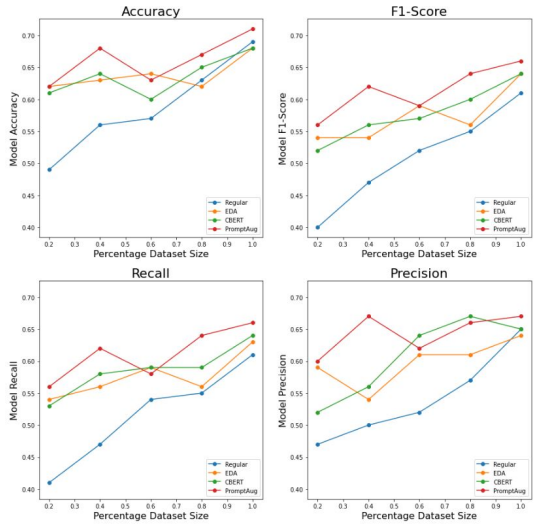

Figure 3: Line graphs of performance vs dataset size.

6.2 Experiment Two

Experiment two evaluates the DA method's ability to augment under increasing data scarcity. For the original dataset, classification performance worsens as dataset size decreases. The same is true for the DA methods but at a lower rate, with the DA techniques reducing the impact of shrinking dataset size on performance metrics. Of the DA methods tested, PromptAug continues to improve the most over the original dataset. With accuracy increases of 13%, 12%, 6%, 4% and 2% over the 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 data sizes. A similar trend exists over the same dataset size intervals for F1-Score with PromptAug improving over the baseline by 16%, 15%, 7%, 9%, and 5%. PromptAug therefore has greater impact on F1-score compared to accuracy at higher dataset size intervals. PromptAug and CBERT experience anomalies at 60% dataset size where performance metrics don't match the trend for other data sizes. The original and EDA datasets do not experience the same performance dip.

Concluding experiment two, as shown in Fig. 3, decreasing dataset size has an adverse effect on performance metrics, this effect is reduced when DA techniques are employed. 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 scarcity scenarios we show its suitability for DA tasks, where tasks that seek to employ a DA technique are frequently struggling with small dataset sizes.

6.3 Experiment Three

Two findings can be observed from the t-SNE visualisation plots (Fig. 4). Firstly, EDA results in higher noise within the generated data than PromptAug, noise within data distorts content and affects classification performance (Agarwal et al., 2007). Secondly, PromptAug generates datapoints closer to the original class characteristics, seen by larger data plot overlap. This suggests that PromptAug expands the training set while retaining class labels.

The thematic analysis performed on misannotated datapoints from the EDA and PromptAug datasets produced four identified themes; both DA methods experinced "Linguistic Fluidity" and "Humour Ambiguity", "Augmented Content Ambiguity" identified within the EDA dataset, and "Augmented Content Misinterpretation" identified within the PromptAug dataset (Table. 4). For the PromptAug dataset annotators had an agreement rate of 67% and Cohen's K of 0.36, described as "fair agreement" by Landis and Koch (Landis and Koch, 1977). For the EDA dataset annotators had an annotation agreement of 46% and Cohen's K of 0.14, described by Landis and Koch as "slight agreement". By conducting the thematic analysis and identifying these themes we provide evaluation of NLP DA beyond classifier performance 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 techniques such as augmented content ambiguity for EDA or augmented content misinterpretation for PromptAug.

The linguistic fluidity theme encompasses fluid or blurred boundaries between class behaviours. Although datapoints tend to have dominant behaviours, they can contain aspects of multiple behaviours. Ambiguous class boundaries have been identified by both Jhaver et al. and Kim et al. (Jhaver et al., 2017; Kim et al., 2022) who identify how Criticism develops into Harassment, the interrelation between the two behaviors, and subjectivity of true class identity. This theme is also present in hate research. Fortuna et al. (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 within research.

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 4 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 examples given 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. (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 the desired behaviour, it occasionally produces erroneous responses, which can range from other negative behaviours, advice on dealing with the behaviour, to completely random. These responses are difficult to filter out and render the new datapoints useless 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 augmented negative behaviour datapoints. Lermen et al. (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.

6.4 Future Works

With the recent emphasis on responsible AI and growing focus on social bias within LLMs, examining how these bias present themselves within DA would be valuable research. A study adopting two methods suggested by Ferrara (Ferrara, 2023), 'Applying fairness metrics' and 'Human-in-the-loop approaches', would provide interesting insights of social bias present within generated data. Secondly, a work which quantifies the expense of DA methods would be of interest, highlighting trade-offs between expense and performance. Future work could also seek to employ PromptAug within other text datasets, evaluating it's generalisability.

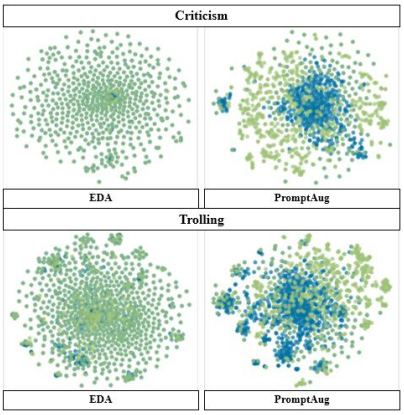

Figure 4: TSNE plots showing the distribution of EDA and PromptAug vs original Trolling & Criticism classes.

7 Conclusion

We present a novel few shot learning DA approach based on informed prompt engineering which targets 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 performance in standard classification metrics over the baseline dataset and other SOTA DA methods. We further demonstrate the effectiveness of the augmentation method in extreme data scarcity scenarios. 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. In addition to the quantitative evaluation of the augmentation methods through classification performance metrics, we also conduct a manual annotation and qualitative thematic analysis of the augmented datapoints to evaluate the quality of 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.

Table 4: Table showing the 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"

8 Limitations

Firstly, we only evaluate our model with regards to generalisability across multiple classification models and dataset sizes. Therefore we cannot make any assumptions about the generalisation of our method to other datasets with different classes and sizes. Additionally, we only use the open-sourced small 7B parameter LLama model for our LLM, so we cannot assume any generalisability with regards to the LLM used for prompting. We also do not investigate any social bias present within the datapoints generated by the LLM.

9 Ethical Concerns

In this paper we discuss harmful content such as harassment and threats, specifically how to generate them using LLMs. This presents an opportunity for individuals with malicious intentions 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

Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. 2007. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 3–12. IEEE.

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

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.

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, 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*.

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.

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.

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*.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

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.

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.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

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

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*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

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*.

A Appendix

Table 5: 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 6: 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