Identifying Citizen-Related Issues From Social Media Using LLM-Based Data Augmentation*

Vitor Gaboardi Santos[1], Guto Leoni Santos[1], Theo Lynn[1], and Boualem Benatallah[1]

Dublin City University, Dublin, Ireland
{vitorgaboardidos.santos, guto.santos, theo.lynn, boualem.benatallah}@dcu.ie

Abstract. Social media platforms, such as Twitter, offer an accessible way for people to share information and perspectives on a wide range of topics. Such citizen discourse can be a valuable source of information and offer policymakers and researchers insights into public sentiment, needs, and suggestions, guiding more informed and responsive planning and policy decisions. In this paper, we propose a novel approach using Large Language Models (LLMs) for data augmentation and multi-class classification to extract domain-specific data from tweets and identify issues raised by citizens thus providing policymakers and social science researchers with valuable data to formulate effective plans and policies for improving services. This approach involves initially collecting data from Twitter using specific keywords and manually labelling a subset of the acquired data. Then, we introduce a new data augmentation strategy employing a LLM that leverages the initial human-labelled data to enhance text diversity and address imbalances in the dataset. Finally, we use the manual-labelled and augmented data to fine-tune different LLMs to classify texts across multiple topics. We test our approach considering the identification of issues related to the cycling domain as case study, detecting tweets across eleven categories associated with infrastructure, safety, and accidents. Through fine-tuning BERT-based models and experimenting with zero- and few-shot prompts with GPT for tweet classification, we accomplished an accuracy of up to 90.9%.

Keywords: Tweet classification · BERT · GPT · LLM · Cycling

1 Introduction

Social media platforms, such as Twitter (currently known as X), provide a convenient and accessible way for citizens to express their opinions and engage in discussions about different topics [4]. They have been extensively used as a source of information, where users contribute with data that can be leveraged to evaluate and monitor the quality of citizen services and infrastructure, including transport [24], traffic congestion [12], urban green spaces [23], and health

* Supported by Dublin City University (DCU).

[21]. The collection and analysis of citizen discourse can empower national and local governments, and others stakeholders to develop effective plans and policies for improving services. Nevertheless, obtaining data related to issues raised by citizens on social media presents significant challenges. Manual approaches are typically time-consuming and labor-intensive [24], emphasizing the need for automated solutions leveraging Natural Language Processing (NLP).

Traditional NLP solutions involve training or fine-tuning standard models for classifying public opinion presented in texts on different topics. In this case, it is necessary to collect and prepare a substantial amount of training and testing data, select a model architecture, train or fine-tune the model, evaluate its performance, and finally deploy it for usage after a satisfactory performance is achieved [20]. Moreover, the model performance relies heavily on the quality and quantity of training data samples, which can be expensive and challenging to obtain depending on the topic [1]. On the other hand, generative LLMs like GPT [2] have been recently employed for classification purposes using zero- and few-shot prompts without the need to collect extensive data [3,32]. In this approach, the developers leverage the model's conceptual understanding of language to write prompts detailing a specific behaviour that the LLM must follow to generate a desired output [28]. Although this strategy allows prototyping NLP applications without the need for extensive training data, it introduces challenges such as high dependence on prompt quality, high costs associated with commercial API inference, scalability issues in real-world scenarios due to slow processing time, and privacy concerns [30,28].

In this paper, we propose an innovative pipeline for identifying specific issues raised by citizens on Twitter using LLMs fine-tuned with training data obtained through a novel human- and GPT-based data augmentation approach. This approach explores the robust performance of traditional NLPs solutions while using the capabilities of generative LLMs to leverage the dataset for fine-tuning models. We initially identify domain-specific topics, gather tweets using related keywords, and manually label a subset of the gathered data. Then, we use GPT to create paraphrases and address the challenge of limited annotated data. This strategy focuses on increasing text diversity and balancing the class distribution, thereby improving generalisation when employing this training data for fine-tuning a multi-class classifier. Next, we combine the manually labelled and augmented data to fine-tune LLMs, specifically, BERT [6] and BERTweet [19], and test GPT [2] under zero- and few-shot prompts for classification.

We evaluate the effectiveness of our approach by considering the domain of cycling-related issues as a case study. Despite the benefits to individual health, economics, and the environment, urban cycling faces several obstacles, including lack of dedicated cycling infrastructure, secure parking and storage, safety concerns, and traffic problems [14,7,15]. Through the identification and monitoring of these challenges in tweets, authorities and other stakeholders may gain valuable insights, enabling them to formulate strategies for enhancing urban cycling policymaking. Furthermore, we assess the performance of the fine-tuned models by computing their accuracy on a test dataset that was manually curated by the

authors. Results show that fine-tuning BERT achieved the highest performance with an accuracy of 90.9% in classifying among eleven cycling-related issues, highlighting that specialized models, fine-tuned on human-annotated and LLM-augmented data, generally outperform zero- or few-shot GPT classification.

In summary, we make the following contributions:

- Novel data augmentation pipeline using GPT-based prompts to address the challenge of limited annotated training data and leverage a balanced dataset to train or fine-tune domain-specific multi-class classifiers.
- Multi-class models to identify citizen-related issues by fine-tuning BERT and BERTweet using manual-labelled and augmented data, and prompting GPT under zero- and few-shot prompts employing only manual-labelled data.
- Collection of a dataset with cycling-related tweets labelled into multiple topics such as infrastructure, theft, parking, and accidents. The method to collect this dataset is applicable to other domains.
- Evaluation and comparison of the classifiers' performance considering both an augmented and manual-labelled data for BERT-based models, and different strategies for choosing samples in the GPT few-shot prompt.

2 Related Work

Text classification: Several works have been developed on text classification using social media data. Plunz et al. [23] focused on tweets from New York City parks, employing a logistic regression classifier with embedding features to determine if tweets related to green urban areas expressed a positive sentiment. Park et al. [21] analysed Google Maps reviews of hub airports in the US to identify the public perception of COVID-19 policy using text clustering to identify four topics. Taleqani et al. [25] conducted sentiment analysis on tweets using naive Bayes, logistic regression, and support vector machine to evaluate public opinion on dockless bike-sharing systems. Das et al. [5] proposed a framework for understanding the emotions of cyclists towards cycling through content published in tweets. They employed standard NLP techniques, such as text mining and sentiment lexicon.

The adoption of LLMs in text classification has become increasingly popular given their remarkable performance [26]. BERT [6] is a pre-trained LLM that excels in comprehending the context of words within a sentence, leading to noteworthy language understanding [10]. BERT has been fine-tuned in many downstream applications using social media data, such as aiding public safety personnel during disasters [31], analysing political campaign messages [13], and detecting COVID-19-related fake news [22]. BERTweet [19] is another pre-trained LLM with the BERT architecture, but fine-tuned on a corpus of 850M English tweets using the RoBERTa [16] pre-training procedure. BERTweet has demonstrated effectiveness in hate, offensive and profane detection [11] and health misinformation detection [29]. Although these solutions demonstrate satisfactory performance, they require substantial training data, which may be difficult to acquire depending on data availability.

On the other hand, GPT is a generative LLM pre-trained on vast amounts of text data, enabling it to understand and generate human-like text with remarkable coherence [10]. GPT can be used as a classification tool through zero- or few-shot prompting without much training data, achieving noteworthy results in hate speech [3] and agriculture topic classification [32]. Nevertheless, employing GPT for classification under prompting has drawbacks, such as limited input token length, high cost of conducting online inference, and scalability issues [30].

Data augmentation: GPT has also been employed as a data augmentation tool to increase training diversity without explicit data collection, and as a solution to improve model generalization and robustness for applications where collecting data is challenging [9]. Yoo et al. [30] proposed a method for synthesizing samples by including a specific number of real samples from each class in a GPT prompt and generating mixed sentences influenced by the included samples. Fang et al. [8] employed GPT-4 to augment student responses for automatic scoring using a prompt requesting for similar written-answers. They used the augmented samples and the original sentences to fine-tune a BERT variant model, leading to improved performance. Møller et al. [18] investigated the application of GPT-4 in generating synthetic data for low-resource domain applications in computational social science. Their findings highlight the benefits of data augmentation, particularly in improving performance on rare classes in multi-class tasks. Van Nooten et al. [27] leverage GPT 3.5 to create samples of anti-vaccination tweets in Dutch to augment a multi-label vaccine hesitancy dataset. They employed prompts that generate instances based on human-labelled data and assign labels to the augmented samples. They found that including the augmented data improves the overall classifier performance.

Our novel approach is designed to address applications with limited-available training data and unbalanced datasets by including a data augmentation pipeline that maximizes the use of input samples and enhances text diversity through paraphrase requests using GPT model. This allows leveraging data to fine-tune state of the art LLMs to identify citizen-related issues. Furthermore, we test our methodology in the case study of cycling-related topics, a domain with limited data availability [5]. We include detecting a broader range of topics than previous works [25,5], such as infrastructure, theft, parking, accidents, and traffic.

3 Methodology

In this section, we present our approach for identifying citizen issues from social media using LLMs. As illustrated in Figure 1, we use the Twitter API to collect tweets with topic-related keywords, and manually label a subset considering desired topics to monitor. Then, we employ GPT to augment the labelled tweets and use both human-annotated and augmented data to fine-tune BERT and BERTweet for multi-class classification, and employ some labelled data for zero- and few-shot classification using GPT. Finally, we evaluate all classifiers using a test dataset. We consider the case study of identifying cycling-related issues to evaluate our methodology.

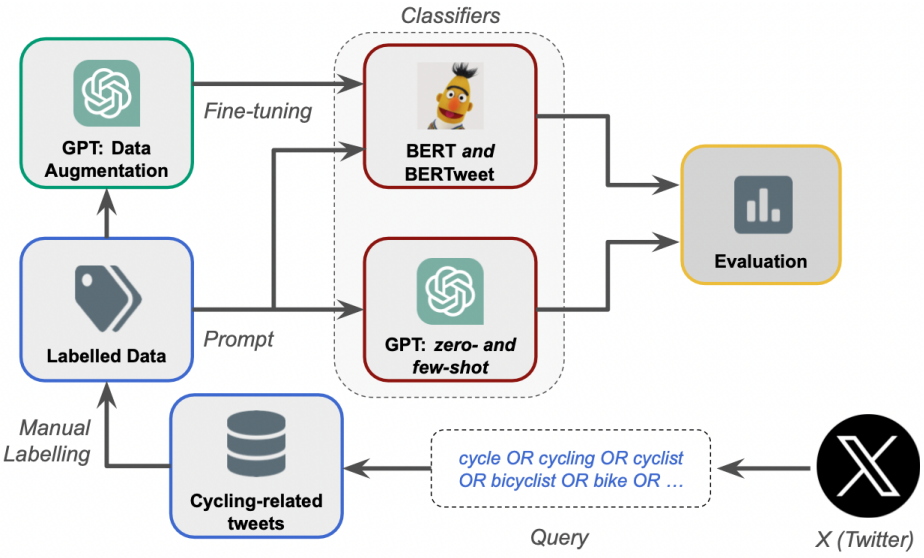


Fig. 1. Approach overview.

3.1 Data collection and human-based labelling

The Twitter data collected comprises 6,744,591 tweets posted between December 1st, 2022 and December 31st, 2022, acquired through the Twitter API. We collected tweets that mention one of the following terms: cycle, cycling, cyclist, bicyclist, bike, bicycle, cycle lane, cycle path, electric bike, city bike, ebike, e-bike.

Urban cycling faces many challenges, including inadequate infrastructure, insecure parking, safety concerns, poor way finding, and traffic issues [14,7,15]. In response to these issues, we selected the following categories for monitoring: *Accidents, Behaviour, Rental, Infrastructure, Journey statistics, Parking, Routes, Sales, Signage, Traffic, and Theft*. Moreover, we labelled tweets into these classes and "others", enclosing tweets unrelated to these topics.

We selected 1,650 tweets from the data collected and manually labelled them. However, we identified two issues after concluding this process. The first is the high number of False Positives (FP), accounting for 43.73% of the labelled data. This problem occurred due to the selection of keywords that occasionally identified tweets discussing other events related to "cycle", such as "cycle of life" or "cycle of abuse", instead of referring to the intended action of riding a bicycle. Moreover, there were FPs originating from product reviews, advertisements for bicycle accessories, and general tweets that did not align with any predefined categories. The second issue is the highly imbalanced dataset distribution.

To improve the dataset with high-quality samples, we manually labelled an additional 1,100 tweets. However, we employed a distinct approach when selecting the samples to label. We computed the most frequent sentence-level bigram

for each class considering the previously labelled data, and used this information to filter new samples from the entire dataset using only the top five most frequent bigrams per class. The intuition is to reduce the number of FP by selecting tweets with word combinations closely aligned with each class. As a result, we observed a decrease in the rate of FP to 36.53% in the second batch. Some sentence-level bigrams and their frequency are displayed in the Figure 2.

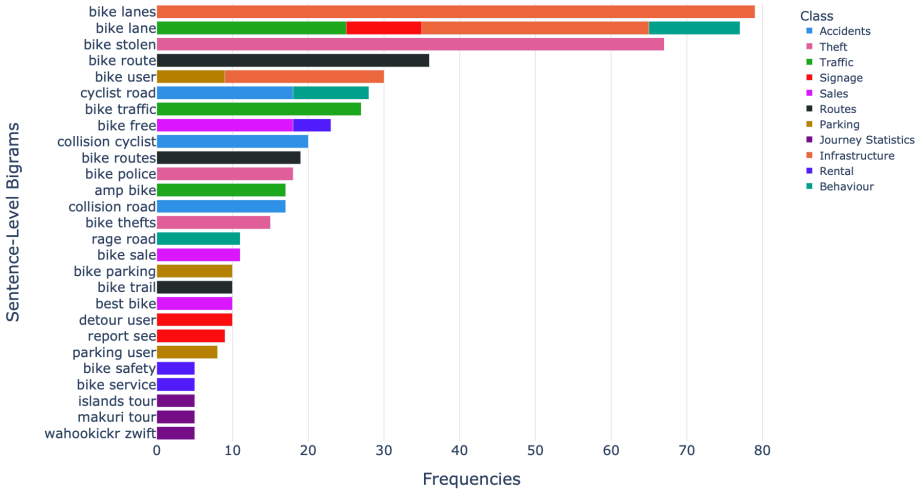


Fig. 2. Five most frequent sentence-level bigram per class.

Some sentence-level bigrams appear across multiple classes. For instance, "bike" and "lane" are found in tweets belonging to the classes *Traffic*, *Signage*, *Infrastructure*, and *Behaviour*. This happened because it is common for users to refer to these topics using both words within the same sentence. This idea of a sentence-level bigram belonging to multiple classes emphasizes how context-dependent the classification of tweets in this application is, where pairing two cycling-related words can lead to different classes.

Figure 3 displays the distribution of manually labelled samples across different classes. It is possible to observe that some classes still have a limited number of tweets, notably *Rental* and *Signage*, each with less than 50 samples. Finally, we pre-processed all tweets from this training dataset to remove URLs and punctuation and replace username mentions with "@user".

3.2 Data augmentation

The percentage of FP decreased after filtering the tweets to be manually labelled based on sentence-level bigrams, which led to a higher number of samples for all categories. However, the training dataset remains imbalanced and some classes

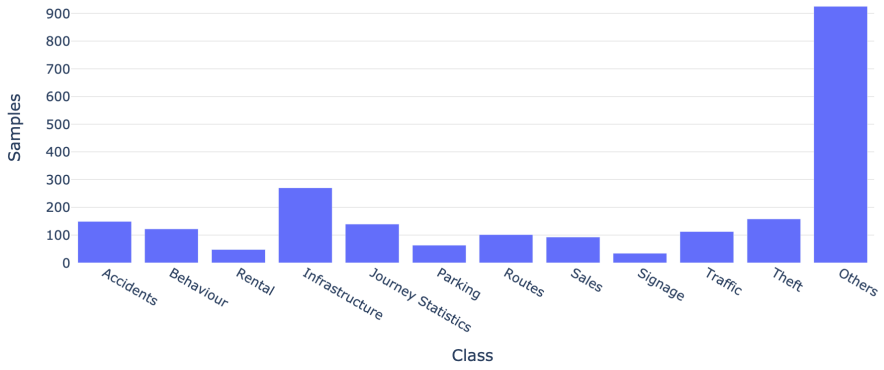


Fig. 3. Number of manually labelled samples per class.

still have few samples. To address these challenges, we prompted `GPT-3.5-turbo` model to perform data augmentation by creating paraphrases of tweets using the already labelled tweets as inputs.

The number of paraphrases requested for each tweet *(n)* varied based on the amount of labelled data available for a particular class. For instance, as shown in Figure 3, the *Signage* class have few samples, so we requested more paraphrases for each tweet of this class when compared to the *Accidents* one, which had a larger sample pool. Our approach aimed to ensure that each class had a minimum of 400 tweets (number of desired samples per class). We chose this threshold because it demonstrated good performance when fine-tuning BERT for multi-class problems, as documented in Liu et al. [17]. Furthermore, we prompted GPT to create paraphrases considering all the samples from the labelled data, enabling a better text diversity when creating the new instances. We did this for every class except for *Infrastructure*, since it already contains many instances, allowing to accumulate 400 samples without using all labelled data.

An overview of our data augmentation process is shown in Figure 4. We first extract all tweets of one particular class, and compute the number of paraphrases to be generated for every tweet within this class by considering the number of labelled data available and number of desired samples. Then, we iterate through each tweet within the selected class and request GPT to generate n paraphrases based on a specific prompt instruction, followed by few-shot examples specific to the class, and the tweet to be paraphrased. Next, we process the GPT's responses using the structure outlined in the prompt and integrate them into the dataset.

After testing multiple prompt instructions to generate diverse lexical and syntactical paraphrases, we used the following one: *You are an expert in paraphrases and creating tweets for cycling content. For each given tweet and subject, I want you to create other tweets that are lexically and syntactically different. You can change names, dates, and places, and you must make it as different as possible while maintaining the subject of the sentence.* This instruction was

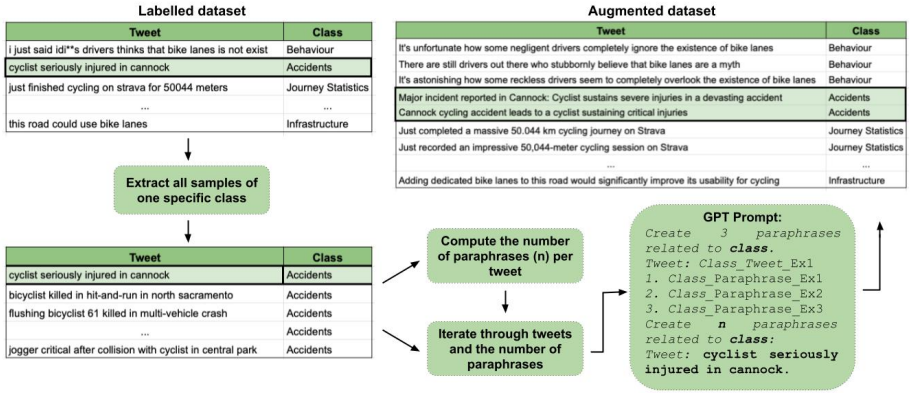


Fig. 4. Data augmentation process overview.

followed by a 3-shot prompt, including an example of a tweet and three corresponding paraphrases. We manually crafted examples for each class with the goal of enhancing the model's contextual understanding.

Despite instructing GPT to generate diverse paraphrases with entity variations, the model often produced text using the same entities as the original tweet. Also, the writing style in these paraphrases tends to be more formal when compared to typical language used on Twitter. However, the content of the paraphrases remained accurate and effectively conveyed the intended meaning of the respective class in most instances. Finally, Figure 5 displays the amount of training samples acquired by merging the manually labelled data and augmented datasets, resulting in a more balanced dataset and increased sample size. In total, the training dataset was composed of 5,599 tweets, of which 2,213 are human created and 3,386 are GPT created.

3.3 Test dataset

To evaluate the classifiers' performance, we built a test dataset consisting of 220 unique tweets (20 per class) posted between September 5th, 2023, and September 7th, 2023. These tweets were manually gathered by the authors from the X website, using the name of each class as a keyword, along with the following terms: cycling, cyclist, bike, and bicycle. We selected tweets from a distinct timeframe than those in the training dataset to minimize potential biases of testing the models with similar content used during fine-tuning and better assess the robustness of the models.

3.4 Tweet classification

To perform multi-text classification of cycling-related tweets, we will evaluate the performance of different LLMs and identify the most effective one. In par-

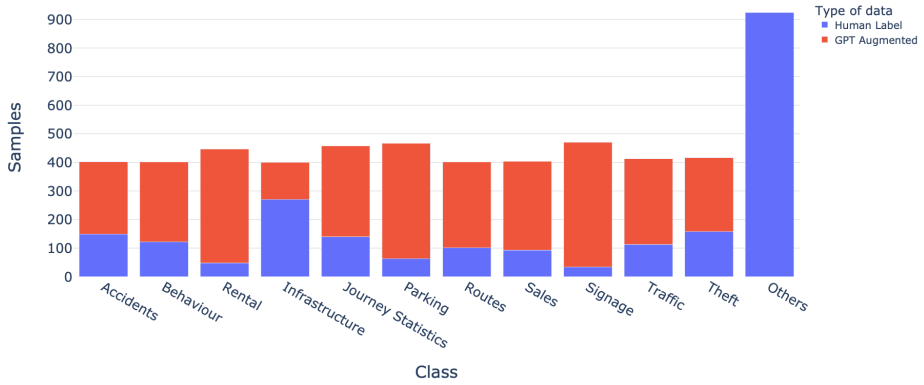


Fig. 5. Training data distribution.

ticular, we employed GPT as classifier under zero- and few-shot prompt, and fine-tuned BERT and BERTweet using the training dataset discussed in section 3.2. Although tweets may cover multiple cycling-related topics, we treat the classes as mutually exclusive. Therefore, each tweet is assigned to a single class that best represents its content.

GPT-based classifier While GPT is primarily known for its use in text generation, it can also be employed for classification tasks, despite its cost and resources [3,32]. In this configuration, we send GPT a prompt with instructions including all possible output classes, and request the model to classify a new input text into one of the mentioned classes. We assess GPT's performance as a classifier considering two configurations: zero-shot prompt, where only an instruction is provided to the model; and few-shot prompt, which include some few examples to enhance the model's contextual understanding of the task. In both cases, we used the following base prompt: *Classify the tweet related to cycling and delimited by triple backticks into one of the following 11 different categories: Accidents; Behaviour; Rental; Infrastructure; Journey Statistics; Parking; Routes; Sales; Signage; Traffic; Theft, and Others (if the tweet is not related to cycling). Return only the name of the class that represents the given tweet.*

We prompted `GPT-3.5-turbo` model for this task. Some parameters can be adjusted to improve GPT's performance as a classifier. The first parameter is the *"temperature"*, which controls the level of randomness in the model's output. Higher values result in a more random output, while lower values lead to a more deterministic outcome. We set the temperature value to 0 to ensure that the model consistently selects the class with the highest probability. Also, GPT can output any word from its extensive vocabulary, although our request is to generate only the class name. Consequently, there may be cases where the first token in the output does not correspond to any of the desired classes. To ensure that the output is consistent, we restricted the output tokens to be

within a predefined set by increasing the *"logit bias"* of all tokens associated with the classes names and the *"end of text"* special token to 100. Lastly, it is better to request the model to output the class names instead of a numerical representation, since it provides a more semantically meaningful approach [2].

BERT-based classifier We fine-tuned `BERT-base-cased` and `BERTweet` using the complete human-labelled and LLM-augmented training data described in sections 3.1 and 3.2. The fine-tuning process was performed considering the following hyperparameters: 5 epochs, 5^{-6} learning rate, AdamW optimizer, batch size of 8 samples, and early stopping if the accuracy does not improve after one epoch. The training and experiments were performed using a computer with Intel(R) Core(TM) i7-12700 CPU at 2.10GHz, 32 GB RAM, and Nvidia GeForce GTX 16660 SUPER.

4 Results

Table 1 summarizes the performance for each model. GPT zero-shot presents the worst results considering all metrics, including an accuracy of 77.27%. However, this outcome is particularly noteworthy as it was acquired without employing any training data for classification, and it serves as our accuracy baseline.

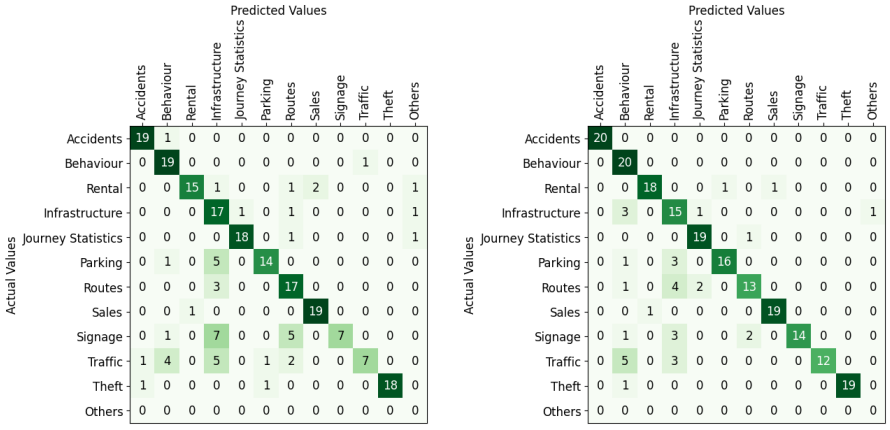
Table 1. Models comparison.

Model	Accuracy	Precision	Recall	F1-score
GPT zero-shot	0.7727	0.7710	0.7083	0.7074
GPT six-shot	0.8409	0.8062	0.7708	0.7755
BERTweet	0.8772	0.8156	0.8041	0.8062
BERT-base-cased	**0.9090**	**0.8418**	**0.8333**	**0.8345**

We provided six examples in the prompt to improve GPT's performance as a classifier, a configuration that we call GPT six-shot. The confusion matrix shown in Figure 6(a) shows that GPT zero-shot had more misclassifications for two particular classes: 21 tweets were incorrectly categorized as *Infrastructure*, and 10 tweets were misclassified as *Routes*. This result indicates that GPT lacks context to differentiate both classes. To address this issue, we provided additional context by adding three samples for each of the *Infrastructure* and *Routes* classes in the prompt, which were randomly chosen from our manual-labelled training dataset. As shown in Table 1, the results with the six-shot examples improved the classification performance by approximately 7% in all metrics except precision.

Figure 6(b) shows the confusion matrix considering the GPT six-shot prompt. The number of correct classification for both *Routes* and *Infrastructure* classes decreased. However, the number of FP also decreased, leading to a better performance overall. We also tested other combinations of examples, such as including

one tweet of all classes, but the results did not improve as much when compared to selecting samples from classes with the lowest performance.



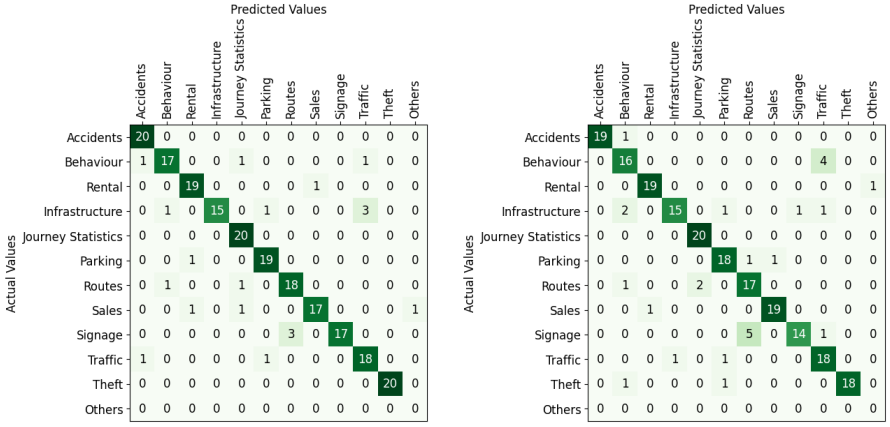
(a) Zero-shot (b) Six-shot

Fig. 6. Confusion matrix considering GPT as classifier.

We also considered fine-tuning BERT-based models. The BERTweet model outperformed both GPT prompts, obtaining more than 80% in all performance metrics. The vanilla BERT model obtained the best performance, being the only one that obtained an accuracy of over 90%. However, it is important to highlight that both BERT models were fine-tuned using the entire training dataset, while GPT used only a few samples.

Figure 7 displays the confusion matrix for BERT and BERTweet fine-tuned models. Both models demonstrated enhancements in the classification of *Infrastructure* class, where BERT exhibited no mistake of other classes with *Infrastructure*, while BERTweet had only one misclassification. For the *Routes* class, both BERT and BERTweet improved the true positive rates, although BERT and BERTweet confused *Signage* tweets with the *Routes* class on three and five occasions, respectively. Other notable results include BERT's confusion of *Infrastructure* with *Traffic* on three occasions, and BERTweet's misclassification of *Behaviour* with *Traffic* four times. Furthermore, the number of correct predictions (diagonals of the confusion matrices) also increased. For BERT, the class with the fewest accurate samples was *Infrastructure* (15 tweets), and for BERTweet, it was *Signage* (14 tweets).

Although this work involves tweet classification and BERTweet was fine-tuned considering English tweets, it had a lower performance when compared to the vanilla BERT. We suspect this occurred for two reasons. First, classifying tweets about cycling may not require specialized Twitter-related features that BERTweet is optimized for. BERT was pre-trained on a diverse corpus of text,



(a) BERT-based-case (b) BERTweet

Fig. 7. Confusion matrix considering BERT as classifier.

including a wide range of topics and writing styles. This general pre-training can make it more adaptable to various text domains. Thus, our tweets probably have a writing style closer to standard written language, which would favour the BERT model. Second, when performing data augmentation, the writing style of the tweets were actually more formal than the usual Twitter language. Therefore, our training data may be more similar to the information trained on BERT.

Furthermore, it is important to note the challenge of classifying tweets considering the classes adopted in this paper, as some texts can be related to multiple classes. Table 2 illustrates some cycling-related tweets with the classification of each model considered in this paper.

The first tweet shown in Table 2 discuss mainly about the traffic consequences when introducing a new cycling lane. Therefore, we consider that this tweet belongs to the *Traffic* class, although it is understandable to be also assigned to *Infrastructure* since it discusses new cycle routes, as classified by the GPT six-shot. Similarly, we consider that the second tweet is related to *Infrastructure* class, as it discuss the need to construct cycling pavements for people safety. However, both BERT and BERTweet model classified it as *Behaviour*, which is also comprehensible given that the tweet also mentions the behaviour of drivers in the city of Harare. Finally, the third tweet showcases an example discussing about *Infrastructure*, where only the GPT six-shot model misclassified.

Therefore, despite our best model achieved an accuracy of 90.9%, we consider this performance outstanding given the complexity of classifying the specific classes. Furthermore, some errors are acceptable since certain tweets may carry multiple interpretations and can belong to two or more classes simultaneously.

Table 2. Classification results of some tweets.

Tweet	GPT 0-shot	GPT 6-shot	BERTweet	BERT
just where do you think Sheffield can handle new cycle routes? Taking away a lane of traffic purely for cyclists will make the already abysmal traffic problem worse and drive people away from the city altogether	Routes	Infra.	Traffic	Traffic
In the new Zimbabwe, we need to advocate for rights of runners in the road! City of Harare should plan & construct (cycling & running pavements) for safety of people who want to exercise! Running in the road is never safe! Harare drivers do not know how to drive.	Others	Infra.	Behav.	Behav.
Kids in bike lanes are an indicator species for safe cycling infrastructure	Infra.	Behav.	Infra.	Infra.

5 Limitations and future work

A primary limitation in the data augmentation pipeline lies in using only OpenAI's `gpt-3.5-turbo` model. In future research, we plan to diversify our approach by integrating other generative LLMs like LLaMA and Falcon, aiming to reduce dependence on commercial models. Furthermore, a more comprehensive assessment of text diversity in the augmented dataset is necessary to ensure the creation of high-quality samples.

In the classification approach, we assumed that classes are mutually exclusive, indicating that tweets are labelled to a single class. However, some tweets may cover multiple topics simultaneously. For instance, in our cycling-related topics discussed in this paper, some tweets could be labelled as *Infrastructure* and *Behaviour*, or *Routes* and *Traffic*, among others. In future work, we plan to consider multi-label classification, assigning more than one class to a tweet.

Further research includes applying our method to other citizen-related issues, performing sentiment analysis to understand overall citizen opinion about each identified issue, considering other languages and input information such as video and images, and investigating alternative approaches involving multiple stages of manual labelling and fine-tuning to enhance the models' performance.

6 Conclusions

We proposed a novel technique for identifying citizen-related issues from social media using LLMs. Our approach involves gathering tweets of domain-specific

topics raised by citizens, followed by manual annotation of a subset of the collected data. Then, an innovative GPT-based data augmentation pipeline creates more training samples from the annotated data. This strategy aims to balance the dataset while also introducing diverse paraphrases. Finally, we fine-tuned BERT and BERTweet using both manually annotated and augmented data, and tested zero- and few-shot GPT prompts for classifying tweets in distinct topics. We assessed this approach by identifying eleven different issues within the domain of cycling, a case study with limited data availability.

The data augmentation strategy worked satisfactorily, producing new text with similar meaning to the original tweets, despite resulting in a more formal tone and not altering entities when requested, resulting in a training dataset more balanced. The fine-tuned BERT model achieved the best accuracy of 90.9% among all tested classifiers. Furthermore, both fine-tuned models performance surpassed zero- or few-shot GPT prompt classification.

Acknowledgement

This work was supported by funding from the European Consortium of Innovative Universities Bicizen project. This work was also conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

1. Adadi, A.: A survey on data-efficient algorithms in big data era. Journal of Big Data **8**(1), 24 (2021)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chiu, K.L., Collins, A., Alexander, R.: Detecting hate speech with gpt-3. arXiv preprint arXiv:2103.12407 (2021)
4. Daemi, A., Chugh, R., Kanagarajoo, M.V.: Social media in project management: A systematic narrative literature review. International Journal of Information Systems and Project Management **8**(4), 5–21 (2021)
5. Das, S., Dutta, A., Medina, G., Minjares-Kyle, L., Elgart, Z.: Extracting patterns from twitter to promote biking. IATSS research **43**(1), 51–59 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Dill, J.: Bicycling for transportation and health: the role of infrastructure. Journal of public health policy **30**, S95–S110 (2009)
8. Fang, L., Lee, G.G., Zhai, X.: Using gpt-4 to augment unbalanced data for automatic scoring. arXiv preprint arXiv:2310.18365 (2023)
9. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for nlp. arXiv preprint arXiv:2105.03075 (2021)

10. Ghojogh, B., Ghodsi, A.: Attention mechanism, transformers, bert, and gpt: tutorial and survey (2020)
11. Glazkova, A., Kadantsev, M., Glazkov, M.: Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. arXiv preprint arXiv:2110.12687 (2021)
12. Gu, Y., Qian, Z.S., Chen, F.: From twitter to detector: Real-time traffic incident detection using social media data. Transportation research part C: emerging technologies **67**, 321–342 (2016)
13. Gupta, S., Bolden, S., Kachhadia, J., Korsunska, A., Stromer-Galley, J.: Polibert: Classifying political social media messages with bert. In: Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC (2020)
14. Heinen, E., Maat, K., Van Wee, B.: The effect of work-related factors on the bicycle commute mode choice in the netherlands. Transportation **40**, 23–43 (2013)
15. Iwińska, K., Blicharska, M., Pierotti, L., Tainio, M., de Nazelle, A.: Cycling in warsaw, poland–perceived enablers and barriers according to cyclists and non-cyclists. Transportation research part A: policy and practice **113**, 291–301 (2018)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. Liu, Y., Dmitriev, P., Huang, Y., Brooks, A., Dong, L.: An evaluation of transfer learning for classifying sales engagement emails at large scale. In: 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). pp. 542–548. IEEE (2019)
18. Møller, A.G., Dalsgaard, J.A., Pera, A., Aiello, L.M.: Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. arXiv preprint arXiv:2304.13861 (2023)
19. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200 (2020)
20. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: a survey of case studies. ACM Computing Surveys **55**(6), 1–29 (2022)
21. Park, J.Y., Mistur, E., Kim, D., Mo, Y., Hoefer, R.: Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of covid-19 policy in transportation hubs. Sustainable Cities and Society **76**, 103524 (2022)
22. Pavlov, T., Mirceva, G.: Covid-19 fake news detection by using bert and roberta models. In: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). pp. 312–316. IEEE (2022)
23. Plunz, R.A., Zhou, Y., Vintimilla, M.I.C., Mckeown, K., Yu, T., Uguccioni, L., Sutto, M.P.: Twitter sentiment in new york city parks as measure of well-being. Landscape and urban planning **189**, 235–246 (2019)
24. Qi, B., Costin, A., Jia, M.: A framework with efficient extraction and analysis of twitter data for evaluating public opinions on transportation services. Travel behaviour and society **21**, 10–23 (2020)
25. Rahim Taleqani, A., Hough, J., Nygard, K.E.: Public opinion on dockless bike sharing: A machine learning approach. Transportation research record **2673**(4), 195–204 (2019)
26. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text classification via large language models. arXiv preprint arXiv:2305.08377 (2023)
27. Van Nooten, J., Daelemans, W.: Improving dutch vaccine hesitancy monitoring via multi-label data augmentation with gpt-3.5. In: Proceedings of the 13th Work-

shop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, July 2023; Toronto, Canada. vol. 1, pp. 251–270 (2023)
28. Viswanathan, V., Zhao, C., Bertsch, A., Wu, T., Neubig, G.: Prompt2model: Generating deployable models from natural language instructions. arXiv preprint arXiv:2308.12261 (2023)
29. Wahle, J.P., Ashok, N., Ruas, T., Meuschke, N., Ghosal, T., Gipp, B.: Testing the generalization of neural language models for covid-19 misinformation detection. In: International Conference on Information. pp. 381–392. Springer (2022)
30. Yoo, K.M., Park, D., Kang, J., Lee, S.W., Park, W.: Gpt3mix: Leveraging large-scale language models for text augmentation. arXiv preprint arXiv:2104.08826 (2021)
31. Zahera, H.M., Elgendy, I.A., Jalota, R., Sherif, M.A., Voorhees, E.: Fine-tuned bert model for multi-label tweets classification. In: TREC. pp. 1–7 (2019)
32. Zhao, B., Jin, W., Del Ser, J., Yang, G.: Chatagri: Exploring potentials of chatgpt on cross-linguistic agricultural text classification. arXiv preprint arXiv:2305.15024 (2023)