A Novel Approach to Address Data Imbalance in Linking Tweets and News Articles using BERT and ChatGPT

ABSTRACT

We aim to link tweets and news articles from The New York Times, which is crucial for various purposes. It benefits computational social science research by providing insights into public responses to events. Additionally, it helps The New York Times create news content that is more appealing to Twitter users, while enabling Twitter to enhance its recommendation system. However, the task is challenging due to significant data imbalance, where some articles have numerous related tweets and others have very few. This imbalance can cause issues during training and evaluation stages. To address this, we propose using ChatGPT for text data augmentation, which involves rephrasing sentences in training samples into multiple conceptually similar but semantically different versions. We then train a BERT-based model to identify the most relevant tweets for a given news article. We conducted several experiments to explore the dataset and evaluate our model's effectiveness. We employed MapReduce to calculate TF-IDF for the dataset, enabling identification of important words and phrases. Visual aids were used to present dataset analysis findings. Additionally, we performed an experiment to assess the efficacy of ChatGPT and BERT.

KEYWORDS

data mining, test semantic matching, ChatGPT ,data augmentation, BERT

ACM Reference Format:

. 2023. A Novel Approach to Address Data Imbalance in Linking Tweets and News Articles using BERT and ChatGPT. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In computational social science, it's important to identify and link events across different document types to analyze public responses [11]. Tweets and news articles are valuable sources, but individually they have limitations. News articles provide contextual information, while tweets offer insights into public reactions. Our research benefits The New York Times and Twitter, helping them understand user preferences and improve content.

Integrating tweets and news articles poses challenges, including extreme data imbalance. Some articles have many related tweets, while others have few. Imbalance affects machine learning models during training and evaluation, leading to biased classification and misleading metrics. Text data augmentation is commonly used to address this, but existing methods lack correct labeling and diversity.

We propose using ChatGPT [17] for data augmentation. These large language models are trained in a self-supervised manner and store vast knowledge. ChatGPT, trained with reinforcement learning, provides informative and unbiased responses. We employ *Sentence-Bert* [18] to find relevant tweets by generating effective sentence embeddings using siamese network structures and cosine similarity.

After augmenting the data with ChatGPT, we analyze the dataset and conduct experiments. We use *MapReduce* to calculate *TF-IDF* and identify important words and phrases. Visual representations help identify dataset characteristics and patterns. Additionally, we compare model performance with and without ChatGPT-augmented data through a controlled experiment. This evaluation determines the effectiveness of ChatGPT for data augmentation.

2 KEY ISSUES AND CHALLENGES

2.1 Data imbalance

As previously mentioned, TF-IDF was used to calculate the baseline. However, there was a significant data imbalance where some news articles had over one hundred related tweets while others only had one. This imbalance can cause issues during both training and evaluation. During the training phase, the model may be biased towards the majority class (in this case, articles that relate to over 100 tweets), resulting in poor performance on the minority class (in this case, articles that relate to only 1 tweet). As a result, the model may not accurately classify instances from the minority class, leading to poor generalization and low overall accuracy. During the evaluation phase, data imbalance can lead to misleading performance metrics. For example, if the majority class has a much larger number of instances than the minority class, the accuracy metric may be high even if the model performs poorly on the minority class. This can lead to incorrect conclusions about the effectiveness of the model. Data imbalance can be particularly problematic in applications where the minority class is of greater interest. For example, a diagnostic system for a rare illness that incorrectly classifies all cases as healthy would still have high accuracy despite being ineffective. In these cases, it is important to address the data imbalance problem to ensure that the model is accurate and reliable. We propose using data augmentation to handle this problem. Data augmentation involves creating synthetic data from existing data to increase the number of instances in the minority class, thus improving the model's ability to generalize and perform well on the minority class.

The dataset contains a tweet file and a news file. The tweet file is a list of tweets with metadata such as mapreducestamp and text, along with their corresponding news links. The news file contains metadata of news articles such as published time and title, but not the content. We used a web crawler to obtain the news content

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

according to the links so that we could construct a dataset with news articles and their relevant tweets. Although we initially encountered incomplete tweet and news data, we ultimately resolved the issue.

3 RELATED WORK

3.1 Approaches for semantic text matching

Semantic text matching is the process of comparing two pieces of text to determine their similarity in meaning. There are several methods without using deep neural network for semantic text matching, including [14] Cosine Similarity, Word Mover's Distance (WMD) [9], and Latent Semantic Analysis (LSA) [6]. Cosine Similarity measures the cosine of the angle between two vectors of text. It is a popular technique for measuring similarity but does not take into account the order of words in a sentence or their relationships. WMD measures how far apart two pieces of text are in terms of their word embeddings. It considers the meanings of individual words and their relationships, but it can be computationally expensive and does not work well with rare words. LSA uses a mathematical technique called Singular Value Decomposition (SVD) to find the underlying latent structure in a set of texts. It is a popular method but has limited effectiveness in capturing the nuances of meaning in text.

There are also several deep learning methods in NLP, such as Convolutional Neural Networks (CNNs) [15], Recurrent Neural Networks (RNNs) [12], and Long Short-Term Memory (LSTM) [19] networks. While CNNs are good at capturing local features, they may struggle with capturing long-range dependencies in a sentence. RNNs and LSTMs, on the other hand, are better at modeling sequential data and can handle long-range dependencies, but they may suffer from vanishing gradients and can be computationally expensive.

In summary, although other models have their strengths and can be effective in certain applications, we chose to use Sentence-Bert due to its ability to capture contextual relationships between words and sentences, giving it an edge over other methods in various NLP tasks.

3.2 Data Augmentation

To handle data imbalance, data augmentation is an effective and popular method which operates at different levels of granularity, such as character, word, sentence, and document levels. Character-level augmentation involves inserting, exchanging, replacing, or deleting characters, while optical character recognition (OCR) [4] augmentation simulates text recognition errors. Word-level augmentation includes methods like random swapping and deletion [3], synonym replacement using databases like PPDB [16] and WordNet [13], and counter-fitting embedding augmentation to adjust word embeddings. Contextual and back translation augmentations generate new text based on context or a different language. Document-level paraphrasing aims to maintain consistency.

Our project aims to use data augmentation to create new, diverse, and semantically consistent tweets based on the original tweets, thereby increasing the size of the dataset and addressing the problem of data imbalance. This approach can benefit both the training and evaluation of the model. By augmenting the data, the model can avoid being biased towards the majority class and perform better on the minority class during training. During evaluation, the risk of misleading performance metrics due to data imbalance can also be reduced.

4 PROBLEM STATEMENT

The task of semantic text matching plays a crucial role in various domains, such as information retrieval [20], social analysis [11] and question answering [10]. In the area of computational social science, associating events in different sorts of articles using semantic text matching is an essential task for analyzing public reaction and attitude to events. The main objective of semantic text matching is to identify the semantic similarities between two different pieces of text. The present study focuses on linking articles published by The New York Times with tweets shared on Twitter by using semantic text matching.

The main goal of this research is to identify the tweets that are most relevant to a particular news article.

This method can be used by The New York Times and Twitter as a powerful tool. The New York Times can use this approach to determine which types of news are more likely to appeal to Twitter users and spread widely, giving them a competitive advantage over other news sources. Similarly, Twitter can use our method to understand user preferences more effectively and recommend news articles more accurately. For instance, if a Twitter user has recently posted several tweets relating to the election of USA, the user will probably be interested in news about the election. Consequently, Twitter can recommend news such as the rate of support of the candidate in election or the prediction of the result of the election. These news should be very catchy and appealing to the user.

Overall, our research has broad implications for social media and news media. By linking news articles with tweets using semantic text matching, we can obtain valuable insights into user behavior and preferences, which can ultimately help organizations improve their content and gain a competitive advantage.

5 METHODS

Our research aims to determine the most appropriate tweet for a particular news article. However, we encountered a significant issue in the form of data imbalance, where some news articles had hundreds of related tweets while others had only one in the current dataset. To address this problem, we employed data augmentation techniques using ChatGPT, and then utilized *Sentence-Bert* to establish the connection between tweets and news articles using the augmented data. Finally, we conducted our experiment by utilizing MapReduce and visualization methods. In the following passage, we will discuss the three primary methods involved in our project, which include *ChatGPT*, *Sentence-Bert*, and *MapReduce*.

5.1 ChatGPT

In our model, we utilized *ChatGPT* [17] with the version "gpt-3.5-turbo" for data augmentation. *ChatGPT* is a variant of the *GPT* (Generative Pre-trained Transformer) language model that has been specifically designed for conversational applications. It is a neural network model of considerable size trained on a comprehensive corpus of text data, allowing it to generate human-like responses to a broad range of natural language inputs. *ChatGPT* surpasses many

traditional models in performance as it can store an enormous amount of knowledge due to its large parameter space, and its extensive pre-training enables it to encode rich factual knowledge for generating language in specific domains.

5.2 Sentence-BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that is specifically designed for sentence encoding, with *Sentence-Bert* (SBERT) [18] being one of its variations. SBERT (Sentence-BERT) is a pre-trained language model that encodes sentence meaning into fixed-length vectors. It uses a siamese or triplet neural network architecture and surpasses other models in several areas. One advantage is its contextual understanding, unlike word2vec and GloVe, which lack context awareness. SBERT utilizes self-attention to process word sequences, capturing word relationships and contextual meaning. Another benefit is its bidirectional processing, comprehending text in both directions simultaneously. This grants SBERT a deeper understanding of word context, resulting in more accurate predictions. Additionally, SBERT's pre-training enables efficient fine-tuning for various downstream tasks, even with limited training data.

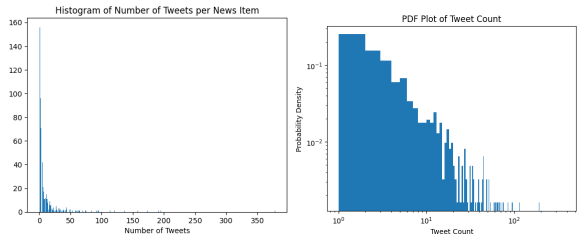

Figure 1: The left figure show Number of tweets per news article. The right figure shows the probability density distribution of number of tweets per news article.

6 EXPERIMENTS

6.1 Dataset

We used the dataset provided by [7]. In the dataset, there are two files: one containing tweets and their corresponding news links, along with meta data such as timestamp and text, while the other includes metadata about news, such as title and publish time, but not the actual content. To create a dataset comprising news articles and their related tweets, we employ a web crawler to retrieve the news content from the links. By inspecting the dataset, we found two problems, namely duplicate and invalid news links which were charged by the *New York Times*. We remove the duplicate and invalid links and finally construct a dataset contains a list of news and relevant tweets pairs. Each pair contains a news article and its relevant tweets. Specifically, the dataset contains 678 news articles and 6,601 tweets.

To gain a detailed understanding of the data imbalance, we counted the number of tweets for each news article using a histogram and a probability density distribution, as shown in Figure 1. We also calculated 25_{th} percentile, median, and 75_{th} percentile of the tweet count which are 1, 3, 10 respectively. The two distributions and the percentile revealed that the dataset suffers from severe imbalance in terms of the number of tweets per news article. Researches [1, 21] have shown that data imbalance can significantly influence the performance of semantic matching models because if there are many more instances of one class than another, the model may not learn to distinguish the less frequent class effectively so there is need to mitigate the imbalance problem.

6.2 Data Augmentation with ChatGPT

To handle data imbalance, data augmentation is an effective and popular method. The ability of pre-trained language models to augment a dataset by generating new samples with similar semantic meaning has been demonstrated by previous research [2], making it a valuable tool for real-world applications. Our study aims to explore the use of *ChatGPT*, a widely-used language model that is built upon the GPT-3 architecture [6]. GPT-3.5 was trained on a vast and diverse corpus of web data. *ChatGPT* was specifically trained using Reinforcement Learning from Human Feedback (RLHF), which involves integrating human feedback into the model's generation and selection process to enhance its performance. In addition, *ChatGPT* have demonstrated an impressive ability to perform In-Context Learning (ICL) [5]. By utilizing only a few input-label pairs as demonstration, these models can predict the label for an unseen input without the need for additional parameter updates. Given these features, we contend that *ChatGPT* is the optimal choice for generating high-quality tweets with human-level features.

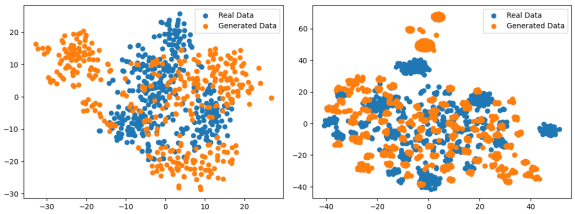

Figure 2: The left figure shows the t-SNE dimension reduction result of the tweets of the news article with maximum tweet count. The right figure shows the t-SNE dimension reduction result of the tweets of randomly selected 100 news article.

6.2.1 Pormpt Design. To utilize the ICL ability of *ChatGPT*, we designed prompts to generate tweets for each news article. Specifically, we created two prompts: one for generating 10 tweets per news article using a single-turn dialogue, and the other for generating an equal number of tweets for each news article using a multi-turn dialogue approach that maximizes the number of tweets generated.

Prompt for single-turn dialogue: We first use the system role to give the assistant (i.e.*ChatGPT*) a role using the prompt *You are twitter users.* Then we give the prompt *Write 10 tweets about the following news article:[news]+Here are an example tweet: [tweet]*

Prompt for multi-turn dialogue: We first use the system role to give the assistant (i.e. *ChatGPT*) a role using the prompt *You are twitter users.* Then we give the prompt *Write [number of tweets]*

tweets about the following news article:[news] Here are three example tweets: 1.[tweet1] 2.[tweet2] 3.[tweet3]. Since *ChatGPT* generate limited number of tweets per turn so we give the prompt *Go on.* to continue generating tweets. If the token of dialogue exceed the maximum number of tokens. We restart the dialogue.

6.2.2 Evaluation of Generated Data. To evaluate the fidelity (i.e., the proximity of the generated data samples to the original samples) and coherence (i.e., the extent to which samples of each category are sufficiently clustered for effective discrimination), we conduct two studies to inspect the two aspects from two perspectives as shown in Figure 2. By examining the t-SNE dimensionality reduction results of two images, we can easily see that the generated tweets' clustering overlaps with that of the original tweets. This implies that the generated tweets have a certain level of fidelity. At the same time, we can also observe that the clustering is not entirely identical, indicating that the generated tweets have diversity. Additionally, it is evident that the generated tweets themselves have clustering, meaning that the generated tweets are compact. This applies to both individual tweets and groups of tweets. So we can conclude that the generated tweets are reliable to some extent.

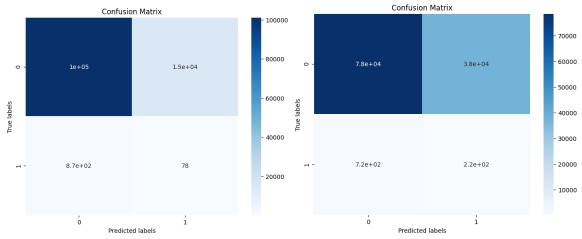

Figure 3: The left figure shows the linear svm classification result with the TF-IDF embedding on the vanilla training set. The right figure shows the linear svm classification result with the TF-IDF embedding on the augmented training set.

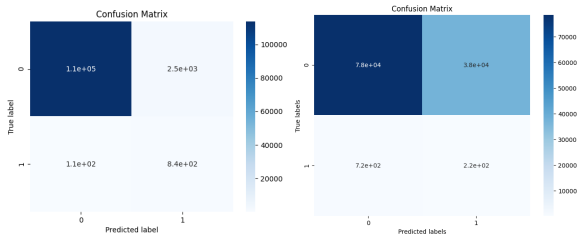

Figure 4: The left figure shows the classification result with *Sentence-Bert* on the vanilla training set. The right figure shows the classification result with *Sentence-Bert* on the augmented training set.

6.3 Text Matching Performance Comparison

6.3.1 Training Set and Development set Preparation. The authors divided the dataset into two parts: 80% and 20%, respectively. The former was used to construct the training set, and the latter was used to construct the development set. For both sets, we matched each news article with all the tweets in the same set, thus building the training and development sets. Therefore, there were 2,788,903 training pairs and 117,056 development pairs. For the augmented training set, we collected the same news as in the training set. Finally, the augmented training set had 5,189,898 training pairs. The augmented development set was the same as the test set.

6.3.2 Training Schema. To show the effectiveness of the data augmentation method, we use two models (namely, *TF-IDF* and *Sentence-Bert* to conduct our experiments.

TF-IDF We calculated the TF-IDF value of a tweet and news pair on the corpus of all tweets and news pairs. Then we trained a linear support vector classifier to classify whether the pair was positive or negative.

Sentence-Bert We fine-tuned the pre-trained model "distilroberta-base" on both training sets with early stopping on single NVIDIA Tesla T4 GPU. *MultipleNegativesRankingLoss* [8] was used as the objective function. Finally, cosine similarity was used for classification with the threshold of 0.5.

6.3.3 Performance Comparison. In real scenario, news article or tweets are used to retrieve relevant tweets or news article so $preciison = \frac{TP}{TP+FP}$ is the metric that has practical significance. From the left figure in Figure 3 we can find that the most negative samples are correctly classified but the precision is quite low, which means most of the positive predictions are wrong. From the right figure in Figure 3, we can infer tha the augmented training set improve the performance of classification in terms of precision. From the Figure 4, we can infer that *Sentence-Bert* achieve a precision of 25.1% on vanilla training set and a precision of 64.0% on the augmented training set, which is a huge improvement.

7 CONCLUSION

In conclusion, our task is to link tweets and news articles from The New York Times. However, a significant issue of data imbalance exists, causing problems during training and evaluation. We were able to handle the problem of data imbalance effectively by utilizing ChatGPT to generate multiple samples that were conceptually similar but semantically different. Furthermore, our BERT-based model, which identifies the most relevant tweets for a given news article, demonstrated high accuracy.

We conducted a series of experiments that enabled us to gain deeper insights into the dataset and evaluate the performance of our model. By using MapReduce to calculate TF-IDF, we identified important words and phrases within the dataset. Visual representations of the dataset's features and statistics helped us to identify patterns and trends. We also conducted experiments to evaluate the effectiveness of ChatGPT for creating a dataset resembling tweets and for data augmentation. The results of these experiments provided valuable insights into the dataset's characteristics, the effectiveness of our model, and the potential of ChatGPT for data augmentation.

REFERENCES

[1] Haseeb Ali, Mohd Najib Mohd Salleh, Rohmat Saedudin, Kashif Hussain, and Muhammad Faheem Mushtaq. 2019. Imbalance class problems in data mining: A

review. *Indonesian Journal of Electrical Engineering and Computer Science* 14, 3 (2019), 1560–1571.

[2] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *Comput. Surveys* 55, 7 (2022), 1–39.

[3] Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* (2017).

[4] Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K Ghosh, Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. 2017. *Optical character recognition systems*. Springer.

[5] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta Optimizers. *arXiv preprint arXiv:2212.10559* (2022).

[6] Nicholas E Evangelopoulos. 2013. Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 6 (2013), 683–692.

[7] Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 239–249.

[8] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652* (2017).

[9] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. *Advances in neural information processing systems* 29 (2016).

[10] Hai Jin, Yi Luo, Chenjing Gao, Xunzhu Tang, and Pingpeng Yuan. 2019. ComQA: Question answering over knowledge base via semantic matching. *IEEE Access* 7 (2019), 75235–75246.

[11] Béatrice Mazoyer. 2020. *Social Media Stories. Event detection in heterogeneous streams of documents applied to the study of information spreading across social and news media*. Ph. D. Dissertation. Université Paris-Saclay.

[12] Larry Medsker and Lakhmi C Jain. 1999. *Recurrent neural networks: design and applications*. CRC press.

[13] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography* 3, 4 (1990), 235–244.

[14] Lailil Muflikhah and Baharum Baharudin. 2009. Document clustering using concept space and cosine similarity measurement. In *2009 International conference on computer technology and development*, Vol. 1. IEEE, 58–62.

[15] Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[16] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 425–430.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[18] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[19] Jürgen Schmidhuber, Sepp Hochreiter, et al. 1997. Long short-term memory. *Neural Comput* 9, 8 (1997), 1735–1780.

[20] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[21] Xiaobo Tang, Hao Mou, Jiangnan Liu, and Xin Du. 2021. Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports* 11, 1 (2021), 11849.