Computer Science • 24(2) 2023 https://doi.org/10.7494/csci.2023.24.2.4761

Anmol Bansal
Arjun Choudhry
Anubhav Sharma
Seba Susan

ADAPTATION OF DOMAIN-SPECIFIC TRANSFORMER MODELS WITH TEXT OVERSAMPLING FOR SENTIMENT ANALYSIS OF SOCIAL MEDIA POSTS ON COVID-19 VACCINE

Abstract *Covid-19 has spread across the world, and several vaccines have been developed to counter its surge. To identify the correct sentiments that are associated with the vaccines from social media posts, we fine-tune various state-of-the-art pre-trained transformer models on tweets that are associated with Covid-19 vaccines. Specifically, we use the recently introduced state-of-the-art RoBERTa, XLNet, and BERT pre-trained transformer models, and the domain-specific CT-BERT and BERTweet transformer models that have been pre-trained on Covid-19 tweets. We further explore the option of text augmentation by oversampling using the language model-based oversampling technique (LMOTE) to improve the accuracies of these models – specifically, for small sample data sets where there is an imbalanced class distribution among the positive, negative, and neutral sentiment classes. Our results summarize our findings on the suitability of text oversampling for imbalanced small-sample data sets that are used to fine-tune state-of-the-art pre-trained transformer models as well as the utility of domain-specific transformer models for the classification task.*

Keywords Covid-19, vaccine, transformer, Twitter, BERTweet, CT-BERT, BERT, XLNet, RoBERTa, text oversampling, LMOTE, class imbalance, small sample data set

Citation Computer Science 24(2) 2023: 167–186

1. Introduction

Covid-19 vaccines were developed in response to the widespread devastation that was caused by the Covid-19 pandemic, which started in December 2019 and has been on the rise ever since. These vaccines have elicited a mixed response among the general public; these reviews help us understand just how these vaccines have affected people emotionally. Social media is the primary platform for finding solutions to health-based queries that are related to the Covid-19 pandemic [1]. One of the best ways to determine public opinion is to survey Twitter, which has millions of users every day and receives lots of tweets on vaccines from all over the world [34]. Most of the researchers who have used machine learning to analyze Covid-19 vaccine-related tweets have adopted unsupervised techniques to determine the sentiments of tweets [20]. A majority of researchers use the rule-based Valence Aware Dictionary for sEntiment Reasoning (VADER) [14], which assigns a sentiment score to each tweet in an unsupervised manner [20, 22, 39]. Other unsupervised techniques that are used for Twitter sentiment analysis include AFINN [27] and TextBlob [26]. However, unsupervised techniques are dependent on pre-defined rules that are meant for a general sentiment analysis that may not work out for the text data that is related to the ongoing pandemic. Supervised learning methodologies overcome the drawback of unsupervised techniques by learning appropriate text patterns that distinguish among positive, negative, and neutral sentiments [31, 43, 45]. Supervised learning for text classification is achieved in the present times by using transformers [12] that are now popularly replacing long short-term memory (LSTM) [25] and convolutional neural networks (CNNs) [10] in various natural language processing (NLP) tasks [12, 32]. Transformer models can be used for sequence modeling to predict the next word in a sentence and are usually trained on large corpora such as Wikipedia or Brown Corpus [8, 41]. These pre-trained models are generalized and are usually fine-tuned for downstream tasks such as classification and text generation.

One of the challenges that are faced in supervised machine learning is the small size of a data set, which leads to less accurate models. Such data sets are called small sample data sets (SSD) [15]. The situation is complicated when the class distribution is imbalanced i.e. when the majority class samples outnumber the minority class samples [42]. SSD provides fewer examples from which a model can identify patterns; therefore, this results in less accurate models. An imbalanced data set can also be detrimental to a model, as it results in biased results toward the majority class and, therefore, results in wrong predictions. This paper aims to explore solutions to the class-imbalance problem that are associated with the sentiment categories of Covid-19-related tweets when the size of the data set is small. We explore the viability of text-based oversampling as a possible solution. The generated synthetic tweets are from classes that have a lower population in order to balance the population of all classes while increasing the number of training samples at the same time. We explore different pre-trained transformer models for supervised learning from a small imbalanced sample data set that contains tweets on Covid-19 vaccine-related discussions

and compare their performance with that of domain-specific transformer models for the classification task. Specifically, we investigate the performance of the state-of-the-art RoBERTa, BERT, and XLNet transformer models as compared to the domain-specific CT-BERT and BERTweet pre-trained transformer models for a sentiment analysis of Covid-19 vaccine-related tweets. CT-BERT and BERTweet are pre-trained transformer models that have been obtained after intensive training on English tweets that are related to the Covid-19 pandemic. On the other hand, RoBERTa, XLNet, and BERT provide input embeddings for sentences that are written in English and are used for generalized natural language processing. The CT-BERT and BERTweet models have the advantage of being familiar with the text patterns that emanate from Covid-19-related discussions such as "Covid positive", which may not be well-understood by the RoBERTa, BERT, and XLNet models.

The organization of this paper is as follows. Section 2 describes some related work on the text oversampling of Covid-19-related text data sets and also reviews the LMOTE algorithm that is used for text oversampling in the current work. Section 3 outlines several state-of-the-art pre-trained transformer models, including domain-specific transformer models. Section 4 presents the methodology, Section 5 contains a detailed analysis of the results, and Section 6 concludes the paper.

2. Preliminaries of text oversampling

2.1. Text oversampling of Covid-19-related text

The small sample size (SSS) problem refers to the availability of a small number of training samples in high-dimensional data sets [24]; this leads to inadequate training, rendering supervised learning a challenging task. This paper uses a small subset of annotated Covid-19 tweets to observe the effects of using a smaller text data set for fine-tuning pre-trained transformer models that are state of the art for implementing various NLP tasks. The situation is complicated when the class distribution is uneven and the number of majority samples is much more than the number of minority samples [42]. In the literature, there are several examples of text oversampling being applied to Covid-19-related social media posts due to the apparent scarcity of text samples that belong to some of the minority classes. We discuss some of these works next. In [21], Liu et al. proved that oversampling term-frequency inverse document frequency (TF-IDF) features using the synthetic minority oversampling technique (SMOTE) [9] improved the results of Covid-19 vaccine-hesitancy prediction. A support vector machine (SVM) was used for the classification. SMOTE was also used in [3] to oversample word embeddings for the sentiment analysis of Arabic tweets that were related to Covid-19 conspiracy theories. A recent work [35] investigated ensemble models for the classification of Covid-19 infodemic tweets that were oversampled using SMOTE. Mohsen et al. [29] recommended text oversampling using SMOTE edited nearest neighbor (SMOTEENN) for the sentiment analysis of Arabic tweets

that were related to the Covid-19 quarantine. Random oversampling was performed in [4] for detecting Covid-19 misinformation on Twitter.

2.2. Review of LMOTE algorithm for text oversampling

The language model-based oversampling technique (LMOTE) that was proposed in 2020 by Leekha et al. [19] is a language modeling-based synthetic data point-generation approach for tackling the problem of class imbalance in natural language-processing tasks. Previous synthetic data point-generation approaches for tackling class imbalance (like SMOTE and its variants [42]) lack the ability to allow for a proper qualitative analysis of generated synthetic data points since the synthetic samples were generated in Euclidean space. This made it difficult to concretely judge the semantic and contextual validity of the generated synthetic data points. Unlike SMOTE and its variants, LMOTE works specifically on textual data, and the synthetic data points that are generated by LMOTE allow for a more concrete and intuitive balancing of the data set.

In our current work on Covid-19 vaccine-sentiment analysis, there are three classes of sentiments: positive, negative, and neutral. The neutral tweets are large in number, as a lot of people tweet about generic information regarding vaccines without expressing any sentiments. Hence, neutral sentiment is the majority class in our problem. The algorithm for the text oversampling of those tweets that belong to the minority classes (positive and negative tweets in our case) using LMOTE is given below.

Algorithm Text oversampling of the minority class using LMOTE

Input: minority class (positive/ negative tweets)
Output: Balanced dataset with augmented minority class
1: Compile the tweets in the minority class into one text corpus
2: Find the most frequently occurring 5-grams in this text corpus (select top-100)
3: Train a Bidirectional LSTM language model on the selected 5-grams
4: Use the trained language model to generate tweets one word at a time
5: REPEAT for all minority classes.

3. Pre-trained transformer models for domain-specific tasks

The transformers that were introduced in 2017 by Vaswani et al. [44] relied on the concept of self-attention that involves the computation of the intra-attention between positions in the input sequence. Transformers are rapidly replacing LSTM and CNN in encoder-decoder models that incorporate attention mechanisms between the encoder and decoder [6, 12]. BERT [11] and XLNet [46] are bi-directional transformer models that are used for learning representations for various NLP tasks, with XLNet outperforming BERT on several tasks that involve learning from long sequences [2, 46]. BERT is a transformer-based model based on masked language modeling. BERT

and its advanced versions (such as RoBERTa [23] and ALBERT [16]) are trained on English Wikipedia and Brown Corpus. Pre-trained BERT models generate word embeddings that can be used for text understanding and classification. BERT can also be fine-tuned to adapt to specific tasks. XLNet is an autoregressive (AR) language model that uses permutation language modeling during the pre-training phase; even though it is similar in architecture to BERT, it differs in its pre-training objective (due to this, it surpasses BERT in various NLP tasks). XLNet is pre-trained by using only a subset of output tokens as a target. Like BERT, pre-trained XLNet models can also be used for many other downstream tasks while also increasing the limits for sequences. In a recent work, the XLNet transformer was used successfully for the sentiment analysis of unlabeled Covid-19 tweets by transfer learning [7]. The XLNet transformer was pre-trained on US Airlines tweet data set that was unconnected with Covid-19. Several language-specific models have been developed (such as CamemBERT for French [28] and GottBERT for the German language [40]); these specific language models have provided better results than the BERT multi-language model. The generalized transformer models can be further trained on downstream tasks to create separate models for different tasks and different languages. We discuss some of these domain-specific models here; each is trained on a specialized corpus that is relevant to the topic at hand and is more effective in this domain.

1. SciBERT (biomedical and computer science literature corpus): this is a BERT-based language model for performing scientific tasks; it was introduced in 2019 by Beltagy et al. [8].
2. FinBERT (financial services corpus): this is a pre-trained NLP model that was proposed in 2019 by Araci [5] for analyzing the sentiments of financial statements and is trained using a large financial corpus.
3. BioBERT (biomedical literature corpus): this NLP model that was pre-trained on biomedical corpora outperformed BERT and various state-of-the-art models in a variety of biomedical text-mining tasks; it was introduced in 2020 by Lee et al. [18].
4. ClinicalBERT (clinical notes corpus): this model focuses on clinical notes and their representations using bidirectional transformers and uncovers the relationships between medical concepts and humans, as discussed in 2019 by Huang et al. [13].
5. mBERT (corpora from multiple languages): mBERT is a single BERT model that was proposed in 2019 by Pires et al.; it was trained on 104 different languages [36]. Those languages with less data were oversampled, and those with a surplus of data were undersampled to balance the corpus.
6. patentBERT (patent corpus): the patentBERT model is a fine-tuned pre-trained BERT model for patent classification that was proposed in 2019 by Lee and Hsiang [17]. The fine-tuning was done by using more than 2 million patents and CNN with word embeddings.

7. RoBERTa (optimized pre-training approach for BERT): RoBERTa is a robustly optimized pre-trained model that was based on BERT [23]. It is implemented on PyTorch and modifies the key hyperparameters of BERT and was trained with much larger mini-batches and learning rates; this helps it achieve better downstream performance in the masked language-modeling approach of BERT.
8. COVID-Twitter-BERT or CT-BERT (Covid-19 tweets): this is a domain-specific transformer-based model that was pre-trained on 160 million Twitter messages that were specifically related to Covid-19 [30]; the aim was to understand the content of social media posts that were related to the Covid-19 pandemic. Muller et al. proposed this model in 2020 [30] and applied it for five different classification tasks. The model gave an improvement over BERT on Covid-19 data sets but needed more pre-training to achieve a similar performance on out-of-domain contents.
9. BERTweet (English tweets): this model is a large-scale pre-trained language model for English tweets and has the same architecture as the base BERT model [33]. Experiments have proven that this model outperforms strong RoBERTa-base and XLM-R-base baselines on various NLP tasks. BERTweet is the first public large-scale model that was pre-trained on English tweets; it was trained using the RoBERTa pre-training procedure using a corpus of 850 million English tweets that were comprised of 845 million tweets that were streamed from January 2012 through August 2019 and 5 million tweets that were related to the Covid-19 pandemic.

4. Implementation details

In our work, we test the suitability of text data augmentation for Covid-19 vaccine-related tweets applied for fine-tuning pre-trained transformer models. Data augmentation of the minority class is a popular remedy for class imbalance [38]. Due to the significant class-imbalance among positive, negative and neutral tweets, we oversample the positive and negative tweets only (minority class), concatenating the synthetic data points to the original data set to generate a more balanced data set. We adapt the LMOTE model for augmenting the data that is given as input to the pre-trained transformer models.

The data set used is a small subset of the larger set of Covid-19 tweets presented by Gabriel Preda [37] which is annotated for positive, negative, and neutral sentiments by FullMoonDataScience. The data set contains 6000 tweets with 3680 tweets belonging to the neutral class, 1900 tweets to the positive class, and 420 tweets to the negative class. The data set is thus both small in size and highly imbalanced. The data format is shown in Table 1.

The input text sequences from the Covid-19 tweets were tokenized. We use the hugging-face library in Python for the implementation of the transformer models. The text in the tweets was pre-processed by removing hashtags, links, emails, punctuation, and extra spaces using regex. The sequential pre-processing steps are shown for an

example tweet in Table 2. In addition to the steps that are shown, tabs and extra spaces were also removed.

Table 1
Data format

column	type
tweetID	integer
label	1, 2, 3
text	string

Table 2
Sequential pre-processing steps for example tweet

Pre-processing step	Output
Original tweet	"More #GoodNewsfrom @bopanc & @DovLieber! #PfizerBioNTech's #COVID #vaccine is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/ ZWwi00rIU via @WSJ https://t.co/7TMIPCkkBa"
Removing @ mentions	"More #GoodNews from & ! #PfizerBioNTech's #COVID #vaccine is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/QZWwi00rIU via https://t.co/7TMIPCkkBa"
Removing Hash tags	"More from & ! 's is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/QZWwi00rIU via https://t.co/7TMIPCkkBa"
Removing websites	"More from & ! 's is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data via "
Removing every Punctuation mark except – !?.	"More from amp ! 's is highly effective after just 1 dose amp can be stored in ordinary freezers for up to 2 weeks according to new data via "
Removing numbers	"More from amp ! 's is highly effective after just dose amp can be stored in ordinary freezers for up to weeks according to new data via "

A three-fold cross-validation was performed for all of the models in our experimentation (RoBERTa, XLNet, and BERT) and the domain-specific CT-BERT and BERTweet. The cross-entropy loss function and Adam Optimizer were used for training the models with a learning rate of 2e-5 and five epochs. The hyperparameter settings were chosen as per the guidelines in the original papers. Fig. 1 shows the process-flow pipeline that shows the training and testing phases.

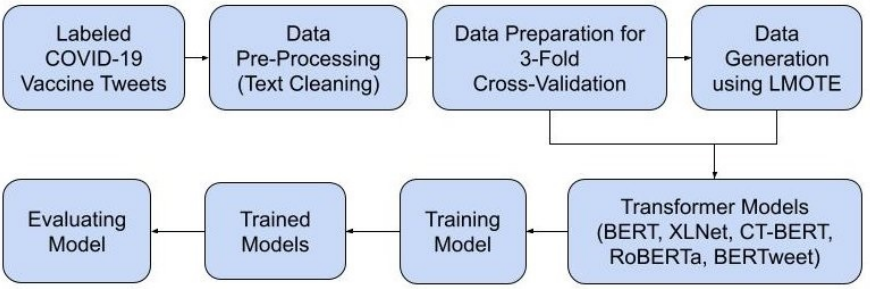


Figure 1. Process flow

The tokenized text was subject to text oversampling, where the positive and negative tweets (minority classes) were oversampled using LMOTE such that the populations of all three classes (positive, negative, and neutral) were balanced. Some of the generated tweets are shown in Table 3 for reference. Neutral tweets were not oversampled since they constituted the majority class. Negative tweets were highly augmented using synthetic samples to match the population of the neutral class.

Table 3
Some instances of tweets that were generated by text oversampling using LMOTE

text generated	label
reporting cases of new cases in toronto closed business church no family no hope canada get out of syria effective and tech	positive
the commission has secured million additional doses of vaccine bringing the total number of doses secured to billion europeans will have had the	positive
home no family no business no church lent no hope until jab unproven vaccine order russia get out of syria that we have no idea of the	negative
company trial participants have died had received amp s s now not not worry to show for any but i know that experienced after the first dose of my	negative

The augmented and balanced data set was applied in order to fine-tune the pre-trained models (RoBERTa[1], BERT[2], and XLNet[3]) and the domain-specific pre-trained models (BERTweet[4] and CT-BERT[5]).

[1]https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md
[2]https://github.com/google-research/bert
[3]https://github.com/zihangdai/xlnet
[4]https://github.com/VinAIResearch/BERTweet
[5]https://github.com/digitalepidemiologylab/covid-twitter-bert

5. Results

Our experiments were performed in Python (Version 3.8) on a 2.8 GHz Intel core PC. We have posted our code online[6] for research purposes. The Covid-19 tweets were pre-processed and tokenized as per the procedure that is outlined in Section 4. In order to explore the effects of text oversampling on our small sample data set, we performed the experiment twice; one with text oversampling, and one without. The text data was used to fine-tune the pre-trained models (RoBERTa, BERT, and XLNet) and the domain-specific pre-trained models (CT-BERT and BERTweet). The transformer models were trained using three-fold cross-validation on three different 80–20 splits of the data set; then, we obtained the mean of the performance metrics over the three runs.

5.1. Results without text oversampling

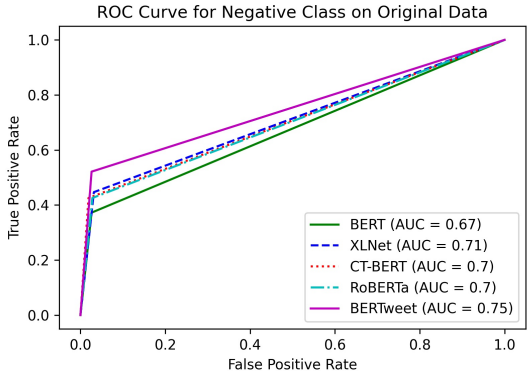
The performance metrics (test accuracy, F1-score, and Mathew's correlation coefficient [MCC]) are summarized in Table 4 for the five transformer models (RoBERTa, BERT, XLNet, CT-BERT, and BERTweet) in the absence of text oversampling. The corresponding receiver operating characteristic (ROC) curves (with area under curve [AUC] readings) are plotted in Figs. 2a, 2b, and 2c for the three sentiment classes.

It can be observed from Table 4 that the domain-specific CT-BERT and BERTweet models showed better performance as compared to the RoBERTa, BERT, and XLNet models in terms of test accuracy, F1-score, and MCC. RoBERTa was observed as being better than BERT and XLNet, while BERTweet proved to be almost as good as CT-BERT and better than RoBERTa. The reason for the better performance of the CT-BERT and BERTweet models is that they were specifically trained on Covid-19 tweets and, hence, were familiar with text patterns that were related to the pandemic-related discussions.

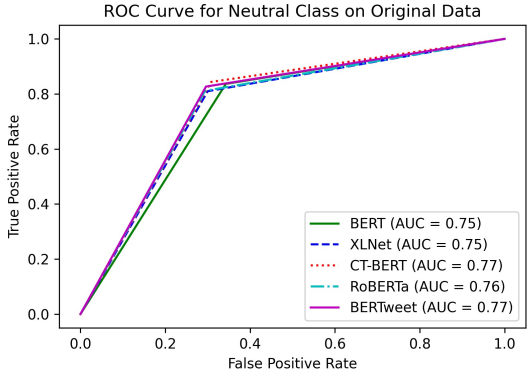
Table 4
Test accuracy, F1-Score, and Mathew's correlation coefficient (MCC) for five transformer models (w/o text oversampling)

Model	Accuracy [%]	F1-score	MCC
BERT	75.06	0.63	0.51
XLNet	75.64	0.67	0.53
RoBERTa	76.44	0.67	0.55
CT-BERT	77.70	0.70	0.57
BERTweet	77.25	0.70	0.56

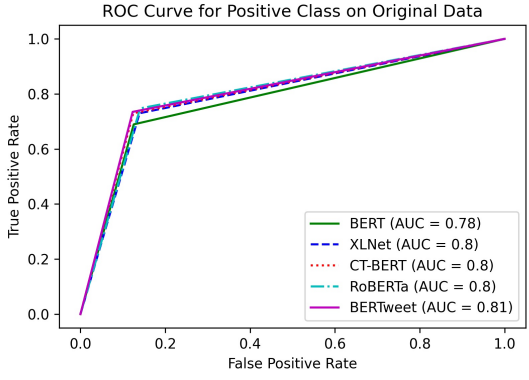
[6]https://github.com/Ace117MC/transformer-models-covid



(a) Negative class



(b) Neutral class



(c) Positive class

Figure 2. ROC curves and AUC values for different models (w/o text oversampling)

Table 5
Class-wise performance analysis of transformer models for positive, negative, and neutral sentiments (w/o text oversampling)

Model	Category	Precision	Recall	F1-score	MCC
BERT	neutral	0.79	0.83	0.81	0.49
BERT	positive	0.72	0.69	0.70	0.57
BERT	negative	0.44	0.34	0.38	0.34
XLNet	neutral	0.82	0.80	0.81	0.51
XLNet	positive	0.70	0.74	0.72	0.58
XLNet	negative	0.50	0.48	0.49	0.45
RoBERTa	neutral	0.82	0.81	0.81	0.53
RoBERTa	positive	0.71	0.75	0.73	0.6
RoBERTa	negative	0.51	0.45	0.48	0.44
CT-BERT	neutral	0.81	0.84	0.82	0.55
CT-BERT	positive	0.75	0.70	0.72	0.61
CT-BERT	negative	0.58	0.54	0.56	0.53
BERTweet	neutral	0.81	0.82	0.81	0.53
BERTweet	positive	0.73	0.73	0.73	0.6
BERTweet	negative	0.56	0.52	0.54	0.51

The detailed class-wise precision, recall, F1-score, and MCC readings are presented in Table 5 for the five transformer models. There are three sentiment classes: neutral (which was the majority class) as well as positive and negative (which were the minority classes, with the negative tweets being very small in number). As expected, the results were biased, with the neutral and positive classes performing better than the negative class (as can be observed in both Table 5 [F1-score and MCC] and Fig. 2 [AUC]). The negative tweets being very low in number were highly mis-classified (as evident from the poor performance of the negative class). CT-BERT was the best performer out of all five models, followed by BERTweet (in terms of test accuracy, F1-score, and MCC), RoBERTa, and XLNet. The accuracies of CT-BERT and BERTweet for the minority class (negative sentiment) were found to be significantly higher than the other models in Table 5, verifying that domain-specific transformer models mitigate the effect of a class imbalance to a certain extent – even in the absence of text augmentation.

5.2. Results with text oversampling

We next demonstrate the effects of text oversampling (of the positive and negative sentiment classes) using LMOTE to investigate the suitability of text augmentation prior to the training phase. The test accuracy, F1-score, and MCC values are summarized in Table 6 for the five transformer models (RoBERTa, BERT, XLNet, CT-BERT, and BERTweet) when text oversampling was performed using LMOTE.

A scrutiny of the results in Table 6 reveal a slight dip in the performance scores after text oversampling as compared to the readings in Table 4 (w/o text oversam-

Table 6

Test accuracy, F1-Score, and Mathew's correlation coefficient (MCC) for five transformer models on data set augmented using LMOTE

Model	Accuracy	F1-score	MCC
BERT	74.25%	0.62	0.49
XLNet	75.80%	0.65	0.53
RoBERTa	76.69%	0.67	0.55
CT-BERT	76.14%	0.67	0.52
BERTweet	77.78%	0.68	0.56

pling); this indicates that the text oversampling of minority classes for small sample data sets will not improve the classification accuracy. For the augmented data set, BERTweet performed the best, followed by RoBERTa and CT-BERT.

We also compared the performance of LMOTE with SMOTE; the performance scores for the five transformer models when performing text oversampling using SMOTE are compiled in Table 7. When comparing the scores of SMOTE (Table 7) with the results of LMOTE (Table 6), we note that the performance of LMOTE was found to be significantly stronger than that of SMOTE. This proves that text generation by language modeling is a better option for augmenting text corpora than the resampling strategies that are prevalent in data mining.

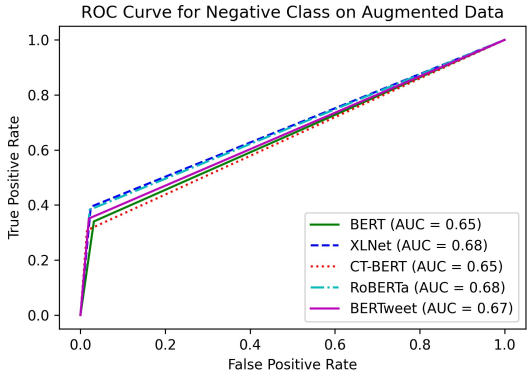
Table 7

Test accuracy, F1-Score, and Mathew's correlation coefficient (MCC) for five transformer models on data set augmented using SMOTE

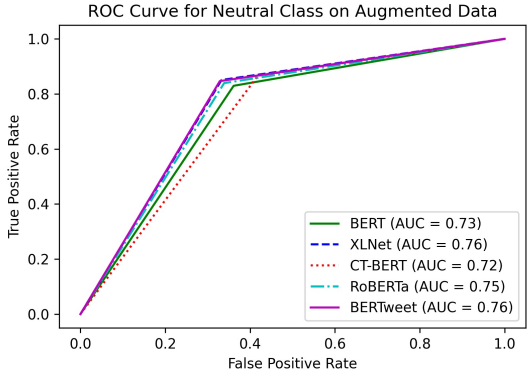
Model	Accuracy	F1-score	MCC
BERT	61.43%	0.54	0.37
XLNet	56%	0.51	0.36
RoBERTa	57.97%	0.54	0.38
CT-BERT	69.58%	0.63	0.47
BERTweet	69.02%	0.6	0.45

The detailed class-wise accuracies are presented in Table 8 for all five transformer models in the case of text augmentation using LMOTE. The corresponding ROC curves (with AUC readings) are plotted for the three sentiment classes in Figs. 3a, 3b, and 3c.

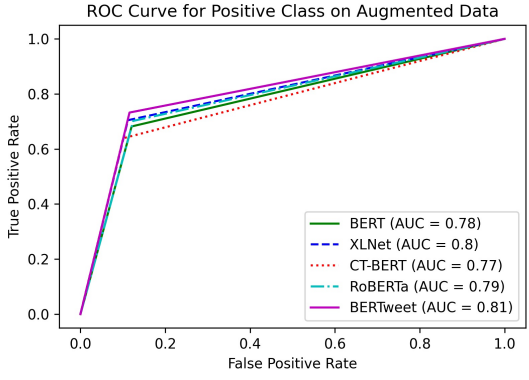
Both Table 8 and Fig. 3 indicate a decrease in the performance scores (after text oversampling) of the neutral class (majority class) as compared to the results of the original data set that was shown in Table 5 and Fig. 2. The positive class accuracies remained more or less the same, with a slight increase noted for some models. However, the negative class scores decreased distinctly for all of the models; this can be attributed to text oversampling.



(a) Negative class



(b) Neutral class



(c) Positive class

Figure 3. ROC curves and AUC values for different models (with text oversampling)

Table 8
Class-wise performance analysis of transformer models for positive, negative, and neutral sentiments (for data set augmented using LMOTE)

Model	Category	Precision	Recall	F1-score	MCC
BERT	neutral	0.78	0.83	0.80	0.47
BERT	positive	0.72	0.67	0.69	0.56
BERT	negative	0.45	0.34	0.39	0.35
XLNet	neutral	0.80	0.82	0.81	0.51
XLNet	positive	0.72	0.73	0.72	0.59
XLNet	negative	0.48	0.38	0.42	0.39
RoBERTa	neutral	0.81	0.82	0.81	0.53
RoBERTa	positive	0.73	0.74	0.73	0.6
RoBERTa	negative	0.52	0.43	0.47	0.43
CT-BERT	neutral	0.81	0.83	0.82	0.49
CT-BERT	positive	0.73	0.71	0.72	0.56
CT-BERT	negative	0.55	0.47	0.51	0.46
BERTweet	neutral	0.81	0.84	0.82	0.54
BERTweet	positive	0.75	0.73	0.74	0.62
BERTweet	negative	0.54	0.41	0.47	0.43

5.3. Discussion

In our work, we investigated the utility of domain-specific pre-trained transformer models and text oversampling for the sentiment analysis of Covid-19 vaccine-related tweets from an imbalanced small sample data set. As observed from the performance scores in Tables 4–8, the domain-specific CT-BERT and BERTweet pre-trained transformer models significantly outperformed the pre-trained transformer models (RoBERTa, BERT, and XLNet). An instance of a Covid-19 tweet that was classified correctly by domain-specific CT-BERT and BERTweet but was classified incorrectly by all of the other transformer models is shown in Table 9, along with another instance of a tweet that was mis-classified by all of the models (including CT-BERT and BERTweet). The latter tweet was a mixture of positive and negative news even though the human annotation labeled it as being negative. Both CT-BERT and BERTweet labeled the tweet as positive due to the phrase "raised no safety concerns."

The following observations were made from the performance scores of the five transformer models before and after the text oversampling by LMOTE (augmentation of the minority classes only).

1. The domain-specific COVID-Twitter-BERT (CT-BERT) model performed significantly better than the pre-trained models (RoBERTa, XLNet, and BERT) for the original data set (Table 4) since it was pre-trained on a Twitter data set that consisted of only Covid-19 tweets. The CT-BERT results also outperformed the other domain-specific model (BERTweet) for the original non-augmented data set.

Table 9
Examples of tweet classification by pre-trained transformer models

Tweet	ground truth	BERT	XLNet	RoBERTa	CT-BERT	BERTweet
i will not be taking amp j or any other vaccine. it's clear now that governments have no idea of the safety profile short medium long term of these experimental vaccines	negative	neutral	neutral	neutral	negative	negative
there were six deaths during the late stage trials but the fda says this raised no safety concerns	negative	neutral	neutral	neutral	positive	positive

2. BERTweet (which has BERT as the base model and was trained on English Tweets) gave consistently better results and outperformed CT-BERT for the augmented data set. The results of CT-BERT and BERTweet were better overall than those of RoBERTa, XLNet, and BERT; this was expected since they were both trained on domain-specific information (tweets that were related to the Covid-19 pandemic).
3. Both CT-BERT and BERTweet performed well for the minority class that contained negative sentiments – even without text augmentation (Table 5); this indicates that pre-training with domain-specific information helps to mitigate the effects of an imbalanced class distribution.
4. The RoBERTa model was proved to be a better fit for the task at hand as compared to BERT and XLNet (as can be observed from the accuracy, F1-score, and MCC scores in Tables 4 and 6).
5. Training the models using synthetically generated textual data yielded worse results for the neutral class and a marginal increase for the positive class, while the scores of the negative class significantly decreased (as can be observed in Tables 5 and 8).
6. It was concluded that LMOTE worked poorly in multi-class settings, often degrading the performance of the majority classes. LMOTE was used to augment and balance the data set in our case so that training could be performed equally for all of the classes.

7. Text oversampling is, thus, not an advisable choice in the case of an imbalanced small sample multi-class data set since it can downgrade the precision and/or recall rate for the majority class (as can be observed from the drop in the performance scores of the neutral class in Table 8). The results of oversampling of Covid-19 tweets using LMOTE and SMOTE (Tables 6 and 7) were not encouraging, as the synthetically generated text for the low-population negative class was not of a high quality and degraded the performance of the neutral class (majority class).

6. Conclusion

In this paper, we explored the effectiveness of domain-specific pre-trained transformer models and text oversampling for learning from small sample data sets with an imbalanced class distribution. We considered the specific task of the sentiment analysis of Covid-19 vaccine-related tweets. The majority class was the neutral sentiment, while the positive and negative sentiments formed the minority classes. The performance scores of the negative sentiment class were the lowest; this occurred due to the small number of training samples in this class. In this scenario, the domain-specific pre-trained CT-BERT and BERTweet transformer models outperformed the RoBERTa, BERT, and XLNet transformer models, which are state-of-the-art pre-trained transformer models that are popularly used for text-classification tasks. Thus, we conclude that domain-specific transformer models are able to mitigate class imbalances to a certain extent. The text oversampling of the minority-class Covid-19 tweets was found to deteriorate the overall performance, with BERTweet performing better than the other models on the augmented data set. Hence, synthetic tweet generation via text oversampling for minority classes is not advisable for imbalanced small-sample text-data sets. We propose the adaptation of domain-specific transformer models for classifying Covid-19-related documents in digital repositories in our future works. Since both CT-BERT and BERTweet models are based on the BERT transformer model, we would like to explore the pre-training of more-recent transformer versions such as XLNet using Covid-19-related tweets in the future.

References

[1] Adeyemi I.O., Esan A.O.: Covid-19-Related Health Information Needs and Seeking Behavior among Lagos State Inhabitants of Nigeria, *International Journal of Information Science and Management*, vol. 20(1), pp. 171–185, 2022.

[2] Adoma A.F., Henry N.M., Chen W.: Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 117–121, IEEE, 2020.

[3] Al-Hashedi A., Al-Fuhaidi B., Mohsen A.M., Ali Y., Gamal Al-Kaf H.A., Al-Sorori W., Maqtary N.: Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories, *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022.

[4] Alenezi M.N., Alqenaei Z.M.: Machine learning in detecting COVID-19 misinformation on twitter, *Future Internet*, vol. 13(10), p. 244, 2021.

[5] Araci D.: Finbert: Financial sentiment analysis with pre-trained language models, 2019. ArXiv preprint arXiv:1908.10063., arXiv:1908.10063.

[6] Bahdanau D., Cho K., Bengio Y.: Neural machine translation by jointly learning to align and translate. In: *3rd International Conference on Learning Representations, ICLR*, 2015.

[7] Bansal A., Susan S., Choudhry A., Sharma A.: Covid-19 Vaccine Sentiment Analysis During Second Wave in India by Transfer Learning Using XLNet. In: *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 443–454, Springer, 2022.

[8] Beltagy I., Lo K., Cohan A.: SciBERT: A Pretrained Language Model for Scientific Text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, 2019.

[9] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.: SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[10] Dastgheib M., Koleini S., Rasti F.: The application of deep learning in persian documents sentiment analysis, *International Journal of Information Science and Management (IJISM)*, vol. 18(1), pp. 1–15, 2020.

[11] Devlin J., Chang M., Lee K., Toutanova K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.

[12] Goel R., Susan S., Vashisht S., Dhanda A.: Emotion-Aware Transformer Encoder for Empathetic Dialogue Generation. In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–6, IEEE, 2021.

[13] Huang K., Altosaar J., Ranganath R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019. ArXiv preprint arXiv:1904.05342., arXiv:1904.05342.

[14] Hutto C., Gilbert E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the international AAAI conference on web and social media*, vol. 8(1), pp. 216–225, 2014.

[15] Kou G., Yang P., Peng Y., Xiao F., Chen Y., Alsaadi F.: Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, *Applied Soft Computing*, vol. 86, p. 105836, 2020.

[16] Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R.: Albert: A lite bert for self-supervised learning of language representations, 2019. ArXiv preprint arXiv:1909.11942., arXiv:1909.11942.

[17] Lee J., Hsiang J.: Patentbert: Patent classification with fine-tuning a pre-trained bert model, 2019. ArXiv preprint arXiv:1906.02124., arXiv:1906.02124.

[18] Lee J., Yoon W., Kim S., Kim D., Kim S., So C., Kang J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, vol. 36(4), pp. 1234–1240, 2020.

[19] Leekha M., Goswami M., Jain M.: A multi-task approach to open domain suggestion mining using language model for text over-sampling. In: *European Conference on Information Retrieval*, pp. 223–229, Springer, Cham, 2020.

[20] Liew T., Lee C.: Examining the Utility of Social Media in COVID-19 Vaccination: Unsupervised Learning of 672,133 Twitter Posts, *JMIR public health and surveillance*, vol. 7(11), p. 29789, 2021.

[21] Liu J., Lu S., Lu C.: Exploring and Monitoring the Reasons for Hesitation with COVID-19 Vaccine Based on Social-Platform Text and Classification Algorithms. In: *Healthcare*, vol. 9, p. 1353, MDPI), 2021.

[22] Liu S., Liu J.: Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis, *Vaccine*, vol. 39(39), pp. 5499–5505, 2021.

[23] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Stoyanov V.: 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692., arXiv:1907.11692.

[24] Lu J., Plataniotis K., Venetsanopoulos A.: Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, *Pattern recognition letters*, vol. 26(2), pp. 181–191, 2005.

[25] Mallick R., Susan S., Agrawal V., Garg R., Rawal P.: Context-and sequence-aware convolutional recurrent encoder for neural machine translation. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pp. 853–856, 2021.

[26] Manguri K., Ramadhan R., Amin P.: Twitter sentiment analysis on worldwide COVID-19 outbreaks, *Kurdistan Journal of Applied Research*, pp. 54–65, 2020.

[27] Marcec R., Likic R.: Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines, *Postgraduate Medical Journal*, 2021.

[28] Martin L., Muller B., Suçrez P., Dupont Y., Romary L., De La Clergerie E., Sagot B.: CamemBERT: a Tasty French Language Model. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, 2020.

[29] Mohsen A., Ali Y., Al-Sorori W., Maqtary N.A., Al-Fuhaidi B., Altabeeb A.M.: A performance comparison of machine learning classifiers for Covid-19 Arabic Quarantine tweets sentiment analysis. In: *2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pp. 1–8, IEEE, 2021.

[30] Müller M., Salathä M., Kummervold P.: Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020. ArXiv preprint arXiv:2005.07503., arXiv:2005.07503.

[31] Naseem U., Razzak I., Khushi M., Eklund P., Kim J.: COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis, *IEEE Transactions on Computational Social Systems*, vol. 8(4), pp. 1003–1015, 2021.

[32] Naseem U., Razzak I., Musial K., Imran M.: Transformer based deep intelligent contextual embedding for twitter sentiment analysis, *Future Generation Computer Systems*, vol. 113, pp. 58–69, 2020.

[33] Nguyen D., Vu T., Nguyen A.: BERTweet: A pre-trained language model for English Tweets. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14, 2020.

[34] Nowak S., Chen C., Parker A., Gidengil C., Matthews L.: Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data, *PloS one*, vol. 15(10), p. 0239826, 2020.

[35] Olaleye T., Abayomi-Alli A., Adesemowo K., Arogundade O.T., Misra S., Kose U.: SCLAVOEM: hyper parameter optimization approach to predictive modelling of COVID-19 infodemic tweets using smote and classifier vote ensemble, *Soft Computing*, pp. 1–20, 2022.

[36] Pires T., Schlinger E., Garrette D.: How Multilingual is Multilingual BERT? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, 2019.

[37] Preda G.: COVID-19 All Vaccines Tweets. https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets. Last accessed on 11th Feb 2022.

[38] Saini M., Susan S.: Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-V3. In: *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 409–420, Springer, Cham, 2019.

[39] Sattar N., Arifuzzaman S.: COVID-19 Vaccination awareness and aftermath: Public sentiment analysis on Twitter data and vaccinated population prediction in the USA, *Applied Sciences*, vol. 11(13), p. 6128, 2021.

[40] Scheible R., Thomczyk F., Tippmann P., Jaravine V., Boeker M.: Gottbert: a pure german language model, 2020. ArXiv preprint arXiv:2012.02110., arXiv:2012.02110.

[41] Somasundaran S.: Two-level transformer and auxiliary coherence modeling for improved text segmentation, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34(05), pp. 7797–7804, 2020.

[42] Susan S., Kumar A.: The balancing trick: Optimized sampling of imbalanced datasets–A brief survey of the recent State of the Art, *Engineering Reports*, vol. 3(4), p. 12298, 2021.

[43] Vashishtha S., Susan S.: Inferring sentiments from supervised classification of text and speech cues using fuzzy rules, *Procedia Computer Science*, vol. 167, pp. 1370–1379, 2020.

[44] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Polosukhin I.: Attention is all you need, *Advances in neural information processing systems*, vol. 30, 2017.

[45] Wang T., Lu K., Chow K., Zhu Q.: COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model, *Ieee Access*, vol. 8, pp. 138162–138169, 2020.

[46] Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.: Xlnet: Generalized autoregressive pretraining for language understanding, 2019. Advances in neural information processing systems, 32.

Affiliations

Anmol Bansal
Delhi Technological University, e-mail: anmolbansal_2k18it025@dtu.ac.in

Arjun Choudhry
Delhi Technological University, e-mail: arjunchoudhry_2k18it031@dtu.ac.in

Anubhav Sharma
Delhi Technological University, e-mail: anubhavsharma_2k18it029@dtu.ac.in

Seba Susan
Delhi Technological University, e-mail: seba_406@yahoo.in

Received: 19.03.2022
Revised: 28.07.2022
Accepted: 08.09.2022