Journal of Computer & Robotics 18 (1), Winter and Spring 2025, 25-32

Computer
& Robotics

MSDSA: Imbalanced Data Sentiment Analysis using Manifold Smoothness Satisfied Data

Shima Rashidi [a.b,] , Jafar Tanha [c, *], Arash Sharifi [a], Mehdi Hosseinzadeh[D]

[a]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
[b]University of Human Development, Sulaymaniyah, Kurdistan Region of Iraq.
[c]Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran.
[D]Pattern Recognition and Machine Learning Lab, Gachon University, Seongnam, Republic of Korea.

Received 29 February 2024, Accepted 01 May 2024

Abstract

This paper proposes a new approach to imbalanced sentiment analysis. The main goal of sentiment analysis is to understand the attitudes and preferences of the user reviews. Recently, this research area has received more attention. In this paper, we focus on imbalanced data in sentiment analysis. The proposed method has three steps. First, we learn a discriminative representation of text tweets. To do so, we fine-tune the BERT model in a supervised manner using a proposed loss function based on manifold smoothness. In this case, the goal is to find a new representation in which each sample's local neighbors belong to the same class label. Second, using the new representation, the over-sampling of the minority class has been done. To do this, we have modified the SMOTE algorithm so that only samples that satisfy the manifold smoothness should be added to the generated sample set. Third, combining the original and over-sampled data, we learn the XGBoost algorithm as a final task predictor. To evaluate the proposed model, we have applied it to the SemEval-2017 Task4 dataset. We have done considerable experiments to show the effectiveness of the proposed method. The obtained results show the strength of the proposed approach.

Keywords: *Twitter Sentiment Analysis, Imbalanced Data, Manifold Smoothness, XGBoost, BERT.*

1. Introduction

Twitter sentiment detector is one of the most important research areas, and its goal is to evaluate the quality of a product or service. Until now, many computational-based approaches have been introduced. In this case, a text tweet is fed into the model, and then the model assigns a sentiment label. This system could provide useful knowledge for other research areas like recommender systems and financial forecasting [1-3].

Recently, deep learning methods have gotten the most attention in sentiment analysis. The reason is that deep neural networks successfully extract knowledge from raw text tweets [1, 4]. However, the classic machine learning approaches perform successfully in task prediction. This area has many challenges, like low data resources and imbalanced data. In low-resource data, some approaches utilize augmentation techniques or generate new data to cope with this challenge. In [5], a generative adversarial network (GAN) [6] is utilized as a data augmentation technique. Meetei et al. [7] introduced a low-data sentiment analysis in which preprocessing techniques are used to generate additional linguistic features to deal with the low data resource.

As mentioned, one of the other challenges in this area is that the sample size of the classes is different. In other words, the dataset is imbalanced. This challenge has sparked interest in recent years. Gosh et al. [8] introduced a method for Twitter sentiment Analysis in which the imbalanced data is investigated. To do so, they have utilized the minority oversampling technique. Krawczyk et al. [9] introduced a model for sentiment

[*] Corresponding Author. Email:, tanha@tabrizu.ac.ir.

analysis in Twitter data in which multi-class imbalance is investigated. They presented a framework with three steps: First, they decomposed multiple classes into many pairs of binary classes using the one vs. one technique. Then, for each pair of classes, they reduced the dimensionality of the data using Multiple Correspondence Analysis. Next, they preprocessed each pair of classes and learned a model for each pair of classes. The final model is the weighted average of the learned binary model. Ah-Pine & Soriano-Morales [10] utilized the syntactic oversampling technique to cope with the imbalanced dataset.

This paper focuses on imbalanced data and proposes a new method to cope with this challenge. The overall schematic of the proposed approach is given in Figure 1. As it is shown, the proposed method has three main steps. In the first step, we learn a discriminative feature descriptor for each tweet. In this case, we have used a pre-trained language model and added some layers on top of the pre-trained model. We have defined a new loss function based on manifold smoothness to learn a more discriminative feature distribution. Then, we have proposed a manifold smoothness SMOTE algorithm, which generates new samples for the minority class. Finally, we use the augmented data to learn the XGBoost algorithm as a task predictor.

To recap, the main contributions of this paper are as follows:

Proposing a new loss function to train language model based on manifold smoothness

Proposing a new version of SMOTE, which generates new samples based on manifold smoothness called MS_SMOTE.

Propose a framework to utilize them in sentiment analysis.

This paper is organized as follows. Section 2 gives a detailed explanation of the proposed method. The experiment and the results are presented in section 3. The advantages and shortcomings of the proposed method are discussed in section 4.

2. Proposed Method

In this section, the proposed approach is given. In Figure 1, the overall schematic of the proposed method is given. The proposed method has three subnetworks, which this section explains in detail.

Problem Formulation

Given $\{(t^{(i)}, l^{(i)})\}_{i=1}^{N}$ shows the whole training set where $t^{(i)}$ shows the ith tweet and $l^{(i)}$ shows its corresponding label. It is assumed that $l^{(i)}$ ∈{positive, neutral, negative}. Also, N shows the number of training samples. The number of samples in the classes is imbalanced in the training set. The main goal of this paper is to design a new approach for sentiment analysis of these tweets in the presence of imbalanced data. In this paper, the imbalanced ratio is defined as the fraction of the number of positive data to the remaining ones.

The proposed method has three steps:

1) learning feature representation to ensure the manifold smoothness.

2) oversampling the minority class by proposing SMOTE+ manifold smoothness.

3) learning a boosting algorithm based on the XGBoost.

In the following, we explain each of these steps in detail.

Feature Representation

In this step, we embed each tweet into a discriminative feature vector. Word embedding is a crucial step in analyzing text data in natural language processing. Several powerful networks have recently been developed for this purpose, including large language-based (LLM) models like BERT. Our study leverages the BERT model due to its superior performance compared to other models. It's worth mentioning that ChatGPT has gained significant attention recently. Although both ChatGPT and BERT rely on deep learning techniques and massive amounts of unlabeled data, but their architectures and training objectives differ. The main factor behind our choice of BERT is its tailoring for specific tasks, such as sentiment analysis, whereas ChatGPT is designed for conversational AI.

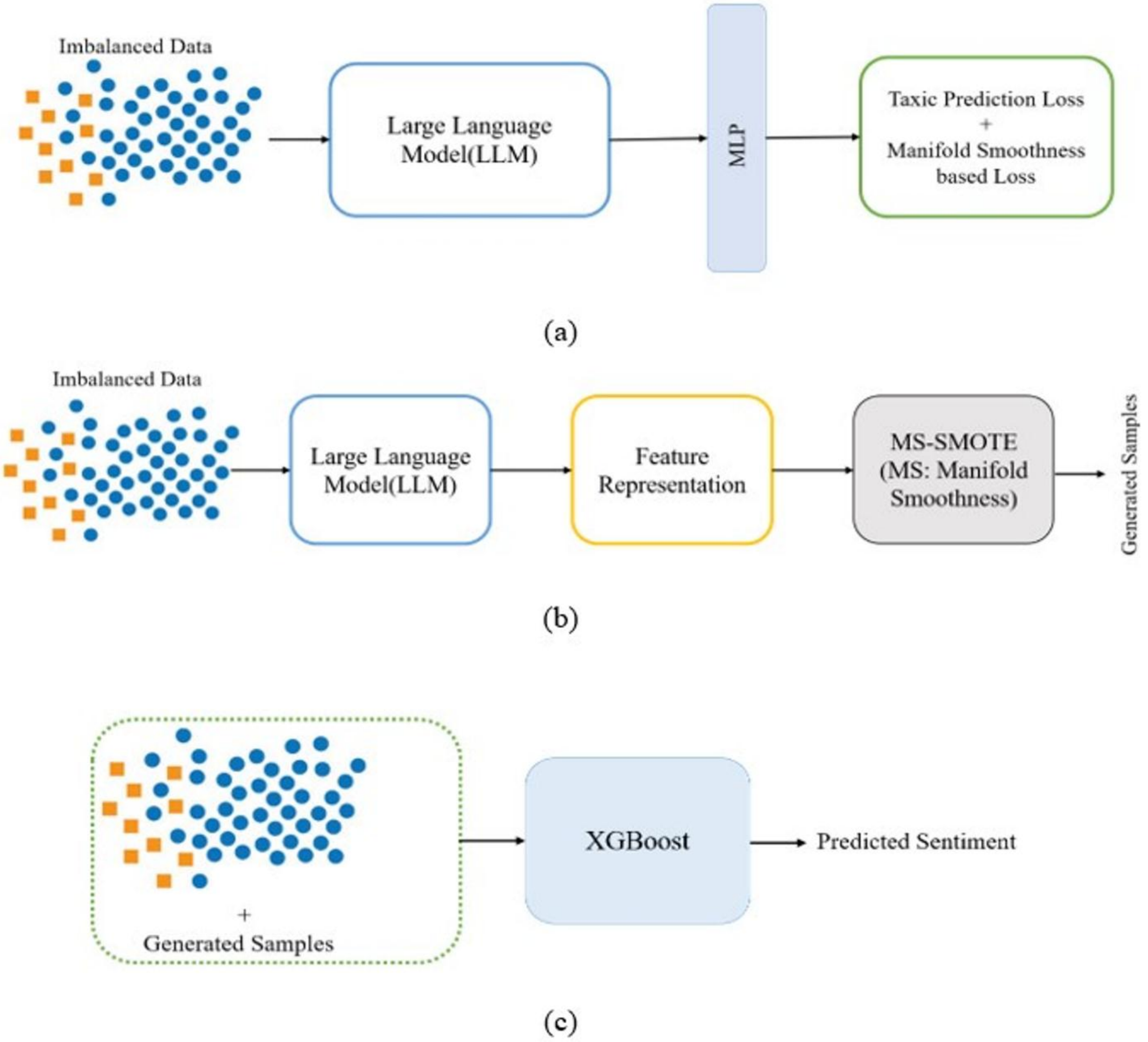

Figure 1- The overall schematic of the proposed method. (a) learn a tweet feature extractor using large language models. The network is fine-tuned using the proposed manifold smoothness-based loss. (b) generate samples for the minority class using the oversampling technique. We have proposed to modify the SMOTE algorithm by incorporating manifold smoothness. (c) using the original data and the generated samples, learn the XGBoost algorithm.

The overall schematic of the subnetwork of this step is shown in Figure 1. As shown, we have used the pre-trained BERT model and then added an MLP on top of it. Then, we fine-tune the model using our data. It should be noted that the layers of BERT are frozen, and only MLP layers are tuned. To train the network, we have defined a new loss function that tries to ensure the manifold smoothness. Hence, it is defined as follows:

$$L_{ms} = 1 - \exp\left(-\left(k - count\left(top\left(N_F\left(t^{(i)}\right); k\right), l^{(i)}\right)\right)\right) \qquad (1)$$

Where $top\left(N_F\left(t^{(i)}\right); k\right)$ retrieve the labels of the top k nearest samples to $t^{(i)}$, based on the feature representation provided by $N_F\left(t^{(i)}\right)$. It should be noted that k is a hyperparameter of the proposed model. Also, $count\left(top\left(N_F\left(t^{(i)}\right); k\right), l^{(i)}\right)$ count the number of retrieved samples that have the same class label with i^{th}sample. The proposed loss (L_{ms}) is minimized if all k nearest retrieved samples have the same class label with i^{th}sample ($l^{(i)}$) and it is maximized if the labels of all k nearest retrieved samples are different with $l^{(i)}$.

Manifold smoothing helps the model learn a model with more discriminative power by expanding the decision boundaries and increasing the distinctiveness of the class representations. It should be noted that after training the model with the proposed loss function, we expect the distribution of the samples of different class labels in the learned space to be somewhat discriminative.

To implement the proposed loss function; we should select the discriminative batches and then feed them into the network to help the model optimize the loss function effectively. Hence, in this case, batch generation plays an important role.

2.1 SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is an imbalanced data-handling method in machine learning [11]. SMOTE is commonly used in classification tasks where the number of samples of different class labels are too different. In other words, there is a minority class that has a significantly smaller number of samples than the other classes (all of them are considered as majority class). In SMOTE, the goal is to generate many synthetic samples for the minority class by interpolating technique. In the following, we first explain how an original SMOTE algorithm performs and then give the proposed version of this algorithm, which utilizes the manifold smoothness to generate samples.

2.1.1 Original SMOTE Algorithm

In this section, we explain the original SMOTE algorithm. This algorithm has five steps, which are explained in the following:

Step 1: SMOTE identifies the minority class and selects a random instance from it. Assume that c is the minority class and c' is the majority class.

Step 2: select one instance from the minority class. This selected instance is called anchor and is shown by a. Choose k nearest neighbors (typically k=5) to the anchor from the minority class. Also, the chosen k nearest neighbor samples are shown by $\{r_i^a\}_{i=1}^k$.

Step 3: For each anchor, SMOTE creates a set of new instances by perturbing the features of the anchor instance. The perturbation is the difference between the anchor and one of the nearest retrieved ones:

$$g_i^a = a + \tau \times (a - r_i^a) \qquad (2)$$

The amount of perturbation is controlled by a parameter shown by τ, which determines the range of values that can be added or subtracted from each feature.

Step 4: the generated instances are labeled as the minority class.

Step 5: Steps 3 and 4 are repeated multiple times until a desired number of synthetic samples have been generated for the minority class. Finally, the synthetic samples are combined with the original training dataset, and the model is trained on this augmented dataset.

2.1.2 Manifold Smoothness in SMOTE

In this section, we have modified the SMOTE algorithm to utilize manifold smoothness to generate more qualified samples for the minority class. In this modified version, steps 2 and 3 differ from the original version. In this case, we only explain how these steps are different from the original version:

Step 2: It chooses k nearest neighbors (typically k=5) to the selected instance from the minority class. In the following sub-steps, these neighbors are investigated to check whether they satisfy the manifold smoothness. Based on this condition, they are selected for the next step.

Step 3: For each anchor, SMOTE creates a set of new instances by perturbing the features of the anchor instance. The perturbation is the difference between the anchor and one of its nearest retrieved ones:

$$g_i^a = a + \tau \times (a - r_i^a) \qquad (3)$$

For the generated sample (g_i^a), we should find its k'-nearest neighbors. If all of these neighbors belong to the same label, we accept this sample; otherwise, it is not accepted, and we should try with the new perturbation coefficient. It should be noted that this procedure is repeated three times. After three times, if we could not generate a sample, we should ignore the anchor instance and go to the next one.

2.2 eXtreme Gradient Boosting (XGBoost)

In this step, the generated samples belong to the minority class, and all the training samples are fed into the XGBoost algorithm [12] to train a model. XGBoost is a powerful machine learning algorithm that has recently gained widespread popularity due to its exceptional performance in handling complex data sets and delivering accurate predictions.

3. Results

In this section, the experiment results are demonstrated and analyzed. To evaluate the proposed method, we have applied it to SemEval-2017 Task 4. SemEval-2017 was first introduced in an international workshop on semantic evaluation, held in conjunction with the 15th Annual Conference of the North American Association for Computational Linguistics (NAACL). This dataset includes many tasks. We focus on task 4, which aims to accurately identify the sentiment expressed in tweets, which can be challenging due to difficulties in understanding the text. SemEval-2017 includes a set of retrieved tweets from the Twitter social network. Each tweet is preprocessed, such as removing emojis and weblinks. In this task, an overall sentiment, including positive, negative, or neutral, is assigned to each tweet. The train set of SemEval 2017 includes 50,333 samples for subtask A. The number of positive, neutral, and negative in the training set are 19902, 22591, and 7840, respectively. Also, the test set contains 12284 samples, including 2375 positive samples, 5937 neutral samples, and 3972 negative samples.

In this paper, we have designed two experiments to assess the proposed method. In the first setting, an ablation study is done to show the effectiveness of the different modules of the proposed method. In the second setting, we designed an experiment in which the model was trained for the different imbalanced ratios and then applied to the test set. Finally, it is compared with state-of-the-art methods.

We need evaluation metrics to assess and compare the model with the base models. In sentiment analysis research, the following metrics are commonly used:

AvgRec, F_1^{PN} and accuracy. For imbalanced data, the AvgRec measure, which computes the average of the recall on the different classes in the problem, is a good choice. It should be noted that in our case, we have three classes: positive, negative, and neutral. The AvgRec measure is defined as follows:

$$AvgRec = \frac{1}{3}(R_p + R_N + R_U) \qquad (4)$$

Where $R_p, R_N,$ and R_U respectively denote positive, negative, and neutral recall measures. It is shown that this measure is invariant to class imbalance [13, 14]. Also, F_1^{PN} computes the average of positive F1-measure and negative F1-measure. Also, accuracy computes the fraction of the truly predicted samples to all samples.

3.1 Ablation Study

In this section, we have done an ablation experiment to show the effectiveness of the different modules of the proposed method. In this experiment, we have created three versions of the proposed method:

1) v1: the whole model is similar, except that the learning of the feature extractor network is done without considering manifold smoothness; 2) v2: The whole model is similar except that the original SMOTE is utilized, and 3) it is the full version of the proposed method. The obtained results are given in Table 1. As shown, the fully proposed method could perform better than the other approaches in three metrics. It should be noted that all three measures are necessary to evaluate the method, and the accuracy measure alone is not good for the imbalanced dataset. The results show that utilizing the oversampling technique to get better results is crucial. In version v2, which does not use the oversampling technique, the obtained F_1^{PN} is much lower than the other versions.

Table 1- The results of the ablation study on the SemEval 2017-Task4 dataset. Also, v1 and v2 are two versions created to show the effectiveness of the proposed loss function and the modified SMOTE algorithm on the final performance.

Approach	**AvgRec**	$\boldsymbol{F_1^{PN}}$	**Accuracy**
v1	69.3	61.8	70.1
v2	72.2	50.5	63.5
Our approach	**74.9**	**73.0**	**76.2**

3.2 Comparison with SOTA

In this section, we have compared the proposed method with baselines and show how it works. We have chosen SVM, Naïve Bayes, Random Forest, and XGBoost as baselines to do so. In this experiment, we have assumed that we have two labels {positive, negative}, and we have set the labeled ratio to different values {0.۰1, 0.05, 0.1, 0.2, 0.4, 0.6}. In this case, 0.01 means that only 1% of the positive samples in the training set are used to train the model. The other ratios are defined similarly. It should be noted that in this experiment, all positive samples are considered as minority class, and the neutral and the negative samples are considered as majority class. The obtained results are shown in Table 2. As shown, the proposed method performs significantly better than the other methods. It means that the proposed method is effective in handling imbalanced data.

Table 2- The impact of the imbalanced ratio on SemEval-2017 Task 4. The results are compared with base approaches.

Approach		**AvgRec**	$\boldsymbol{F_1^{PN}}$	**Accuracy**
SVM	**1%**	0.494	0.497	**0.988**
	5%	0.471	0.485	0.943
	10%	0.446	0.471	0.892
	20%	0.413	0.452	0.826
	40%	0.816	0.498	0.832
	60%	0.693	0.535	0.829
Naïve Bayes	**1%**	0.494	0.497	**0.988**
	5%	0.541	0.498	0.682
	10%	0.574	0.530	0.636
	20%	0.595	0.559	0.617
	40%	0.592	0.553	0.609

	60%	0.587	0.543	0.596
Random Forest	**1%**	0.494	0.497	**0.988**
	5%	0.471	0.485	0.943
	10%	0.446	0.471	0.892
	20%	0.413	0.452	0.823
	40%	0.413	0.452	0.826
	60%	0.413	0.452	0.826
XGBoost	**1%**	0.494	0.494	0.975
	5%	0.616	0.546	0.936
	10%	0.561	0.523	0.868
	20%	0.556	0.520	0.796
	40%	0.601	0.572	0.797
	60%	0.624	0.595	0.803
Our Approach	**1%**	0.559	0.503	0.935
	5%	0.597	0.552	0.926
	10%	0.649	0.567	0.897
	20%	0.663	0.581	0.798
	40%	0.710	0.599	0.765
	60%	**0.722**	**0.603**	0.828

As it is shown in Table 2, the proposed method generally outperforms the other approaches. Among the base approaches, the random forest has the worst result, and XGBoost could perform better than SVM, Random Forest, and NaiveBayes.

Also, to compare the proposed method with the state-of-the-art approaches, we used the same training and test sets (i.e., standard split) and then learned the model. The obtained results are given in Table 3. As shown, the proposed method performs better than the other comparing approaches. We increase the AvgRec measure by 1.7% compared to the best-comparing approach.

Table 3- The comparison of the proposed method with the recent successful approaches.

Approach	**AvgRec**	F_1^{PN}	**Accuracy**
XGBoost	58.6	57.5	58.6
BB_twtr [15]	68.1	68.5	65.8
DataStories [16]	68.1	67.7	65.1
BERTweet [17]	73.2	72.8	71.7
Our approach	**74.9**	**73.0**	**76.2**

4. Conclusion

In this paper, we have proposed a new approach for imbalanced data in sentiment analysis. The proposed method introduces a unique framework in which the tweet representation is learned using a deep-learning-based BETR model. In this case, we propose a new loss function based on manifold smoothness, which aims to learn a discriminative representation of the samples. Then, we oversample the minority class using the new modification of the SMOTE algorithm. In this new modification, we only accept those generated samples that satisfy the manifold smoothness. Finally, the original data and the generated samples are fed into the XGBoost algorithm to learn a sentiment predictor model.

One of the main advantages of the proposed method that leads the model to better performance is the learned discriminative representation of the samples. This representation helps the model train a stronger and more generalizable task predictor.

One caveat to this approach is that optimizing might be sensitive to batch generation. As explained, the proposed loss function is based on the manifold smoothness. To check it properly, we should generate proper samples. Also, the size of the local neighbor is important in satisfying manifold smoothness in steps one and two. In future work, we want to use the differentiable version of XGBoost to design an end-to-end framework.

Competing interests

The authors declare no competing financial interest.

Authors contribution statement

SR: conceptualization, data curation, result analysis, methodology, writing, review & editing. ASH&MH: result analysis, project administration, review & editing. JT: conceptualization, supervision, project administration, review & editing.

References

[1] B. AlBadani, R. Shi, and J. Dong, "A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM," *Applied System Innovation*, vol. 5, no. 1, p. 13, 2022.

[2] I. K. Gupta, K. A. A. Rana, V. Gaur, K. Sagar, D. Sharma, and A. Alkhayyat, "Low-resource language information processing using dwarf mongoose optimization with deep learning based sentiment classification," *ACM Transactions on Asian and Low-Resource Language Information Processing,* 2023.

[3] P. Balage Filho, L. Avanço, T. Pardo, and M. d. G. V. Nunes, "NILC_USP: An improved hybrid system for sentiment analysis in twitter messages," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014),* 2014, pp. 428-432.

[4] A. Tripathy, A. Anand, and V. Kadyan, "Sentiment classification of movie reviews using GA and NeuroGA," *Multimedia Tools and Applications,* vol. 82, no. 6, pp. 7991-8011, 2023.

[5] R. Gupta, "Data augmentation for low resource sentiment analysis using generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7380-7384.

[6] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[7] L. S. Meetei, T. D. Singh, S. K. Borgohain, and S. Bandyopadhyay, "Low resource language specific preprocessing and features for sentiment analysis task. Language," *Resources and Evaluation,* vol. 55, no. 4, pp. 947-969, 2021.

[8] K. Ghosh, A. Banerjee, S. Chatterjee, and S. Sen, "Imbalanced twitter sentiment analysis using minority oversampling," *IEEE 10th international conference on awareness science and technology (iCAST),* pp. 1-5, 2019

[9] B. Krawczyk, B. T. McInnes, and A. Cano, "Sentiment classification from multi-class imbalanced twitter data using binarization," *Hybrid Artificial Intelligent Systems: 12th International Conference*, pp. 26-37, 2017.

[10] J. Ah-Pine and E. P. Soriano-Morales, "A study of synthetic oversampling for twitter imbalanced sentiment analysis," *Workshop on interactions between data mining and natural language processing (DMNLP),* 2016.

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

[12] T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2,* vol. 1, no. 4, pp. 1-4, 2015.

[13] F. Sebastiani, "An axiomatically derived measure for the evaluation of classification algorithms," in *International Conference on The Theory of Information Retrieval*, 2015 pp. 11–20.

[14] P. Nakov *et al.*, "Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts," *Language Resources and Evaluation,* vol. 50, no. 1, pp. 35–65, 2016.

[15] M. Cliche, "BB twtr at SemEval-2017 Task 4: Twitter sentiment analysis with CNNs and LSTMs," *International Workshop on Semantic Evaluations,* pp. 573–580, 2017.

[16] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *International Workshop on Semantic Evaluations*, 2017, pp. 747–754.

[17] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," *arXiv preprint arXiv:2005.10200,* 2020.