Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 757–765

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Imbalanced aspect categorization using bidirectional encoder representation from transformers

Ashok Kumar Jayaraman[a], Abirami Murugappan[a], Tina Esther Trueman[b], Gayathri Ananthakrishnan[c,*], Ashish Ghosh[d]

[a]*Department of Information Science and Technology, Anna University, Chennai, 600025, India*
[b]*Department of Computer Science, University of the People, United States*
[c]*Department of Information Technology, VIT University, Vellore, 632014, India*
[d]*Machine Intelligence Unit, Indian Statistical Institute, Kolkata, 700108, India*

Abstract

Sentiment analysis (also called opinion mining) is one of the widely used research fields of natural language processing. E-commerce service providers use this technique to analyze the sentiment of a product or a service in texts, posts, and comments. In particular, the service providers and users want to understand the sentiment on product aspect categories rather than the overall sentiment of a product. These aspect categories encounter the class imbalance problem. Therefore, the BERT (Bidirectional Encoder Representation from Transformers) based fine-tuning model is presented to deal with the imbalanced aspect categorization task. Specifically, this paper studies various data sampling techniques such as stratified random sampling (SRS), random undersampling (RUS), and random oversampling (ROS) for reducing the class imbalance problem. Empirically, the results show that the proposed BERT fine-tuning model with the SRS technique achieves better results. In particular, the model achieves 96.21% for the validation and 96.47% for testing using the news aggregator data. Similarly, the SMS spam collection data achieves 99.20% for the validation and 99.10% for testing.

Keywords: Deep learning; aspect category detection; BERT; class Imbalance; transformers; data sampling techniques.

1. Introduction

The advancement of the internet has increased the E-commerce web service providers across the globe. These services influence internet users to express their likes and dislikes on a product or a service in online forums, blogs, or business websites [1]. Service providers use this information to learn more about their products. Specifically, sentiment

* Corresponding author. Tel.: +91-9444618421
E-mail address: gayathri.a@vit.ac.in

1877-0509
10.1016/j.procs.2023.01.056

analysis plays an important role to analyze users' interests such as positive sentiment or negative sentiment [2] which are expressed in terms of the text. These sentiments can be detected in the document, sentence, and aspect or entity levels. The document and sentence levels compute overall sentiment about a product or a service. These levels are not expressing a specific sentiment about a product entity. The aspect-level or aspect-based sentiment analysis (ABSA) computes a sentiment of an entity or an attribute of a product. However, the ABSA has various sub-tasks, namely, aspect sentiment detection (ASD), aspect category detection (ACD), and aspect-term extraction (ATE). First, the ASD task detects a sentiment of a given text with respect to the predefined aspect categories. Second, the ACD task classifies a given text or a sentence into one of the user-predefined aspect categories [1, 3]. Third, the ATE task identifies all the aspect-term in a sentence or document. In this paper, the ACD task is focused with class imbalance, where the aspects are highly skewed or not having an equal distribution [4].

In general, the class imbalance problem is encountered in text classification tasks either in binary, multi-class, or multi-label settings [4, 5, 9, 6]. Most of the learning algorithms over-classify the majority class due to their higher prior probability. Also, this leads to the problem of misclassification towards the minority class. There are many open challenges in the task of imbalanced categorization like small example size, overlapping, dataset shift. The first challenge arises due to the lack of information. Especially, the models are not able to generalize well in the high dimensional data. In this scenario, the model leads to the problem of overfitting. In overlapping, there some data contains similar training instances in each class. In this scenario, the model difficult to find differences between classes. The dataset shift has different distributions in the training and testing data. It affects the problem of classification due to the bias issues [7].

However, researchers used data, algorithm, and hybrid level techniques for addressing the class imbalance problem in machine learning and deep learning. These levels use sampling methods, class weights, and both sampling and weight, respectively. Recently, recurrent neural networks (RNN) and their variants such as LSTM (long short-term memory) [10] and GRU (gated recurrent unit) [11] have shown great success in the task of text categorization. These networks read an input token at a time step. In particular, the recurrent neural network reads an input sequence either from right to left or left to right. This unidirectional information leads to the problem of capturing the semantic context between the previous and next word of a token. Therefore, to capture the semantic meaning in both directions, Devlin et al. [12] introduced a BERT (Bidirectional Encoder from Transformers) pre-trained and fine-tune model based on the concept of transformers [13]. In this paper, the BERT fine-tuning model is proposed to categorize imbalanced aspects. Specifically, it mainly contributes the following:

- Solves the imbalanced aspect categorization using context-dependent features
- Applies the stratified samplings, random undersampling, and random oversampling techniques
- Employs a BERT pre-trained model
- Outperforms the task of imbalanced aspect categorization

The chronological order of this paper is organized as follows. Section 2 describes the relevant works on the news aggregator dataset. Section 3 explains the BERT fine-tune model for the task of imbalanced aspect categorization. The empirical results and discussions are explained in Section 4. Finally, the proposed study is concluded in Section 5.

2. Related work

An imbalanced dataset consists of majority and minority classes which plays a vital role in text categorization. Generally, we cannot rely on a prediction system that is based on majority and minority classes. The majority class may influence the system to produce an inaccurate result. Therefore, most of the machine learning algorithms rely on the assumption of balanced distribution. In this paper, we present the existing research works in the News aggregator dataset. Mehta et al. [14] introduced GRUs with a multi-head self-attention (MSA). Their result indicates that the MSA is more efficient and cheaper than self-attention for the text classification task. The authors used the low-rank matrix factorization method to get multiple attention distributions and obtained attention scores by a global context vector. Pushp et al. [15] predicted a category with three different architectures using a Zero-shot learning approach. The authors studied that the models generalize well with low accuracy. Luis Bronchal [16] used the logistic regression (LR) method to categorize aspects. The author achieved an accuracy of 94.73%. Chemchem et al. [17] presented a

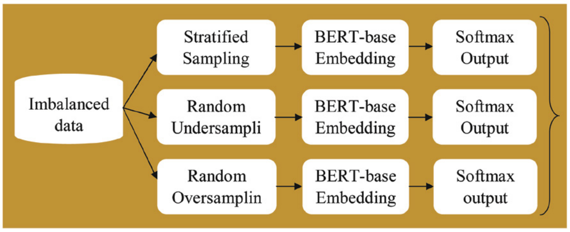

Fig. 1. The proposed BERT fine-tuning model for imbalanced aspect categorization

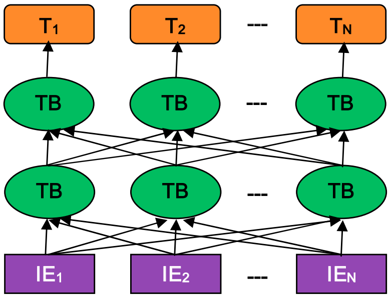

Fig. 2. BERT Model, where IE_N represents $n-th$ input token in a sequence, TB represents the Transformer Block, and T_N refers to the target embedding [12, 25]

supervised learning approach based on knowledge mining and meta-models. The authors extended MLDM algorithms (NB, MNB, Linear SVM, MLNN, and CNN) with induction rules for discovering the knowledge very quickly. Their results suggested that the NB algorithm achieved 86.2% accuracy.

Moreover, Akritidis et al. [18] introduced a supervised learning method to analyze news titles, and to construct variable-length tokens. Later, they reduced the dimensions based on the token scores. The authors show that the logistic regression method achieves 95% by extracting unigram, bigram, and trigram tokens. Bikki [19] performed an automatic text news classification using LR, Linear SVM, MNB, and RF. The author has indicated that the LR method performs well. Lin et al. [20] studied the extension of traditional active learning imbalanced classes and query generation. Their study suggested that the Make Balanced – Cost Bound (MB-CB) algorithm outperforms than Learning-Hybrid (GL-Hybrid) algorithm for imbalanced classes. Suh et al. [21] studied the oversampling technique on imbalanced Korean news articles. Their study found that the oversampling technique generally improves the classifier performance. Based on the above observations, this research work studies the imbalanced aspect categorization using BERT with stratified sampling, undersampling, and oversampling techniques.

3. Aspect Category Detection Using BERT

In this section, the proposed imbalanced aspect categorization model is presented as shown in Fig. 1. The components of this proposed tasks are described as follows.

3.1. Data

The news aggregator data from the UCI machine learning repository [22] and SMS spam collection data [23] are used to address the class imbalance problem. The news aggregator data contains about 422,419 news documents. Each of these news documents is categorized into business (115,967), health (45,639), entertainment (152,469), and technology (108,344). Similarly, the SMS spam collection data contains about 5572 documents. Each of these documents is categorized into ham (4825) and spam (747).

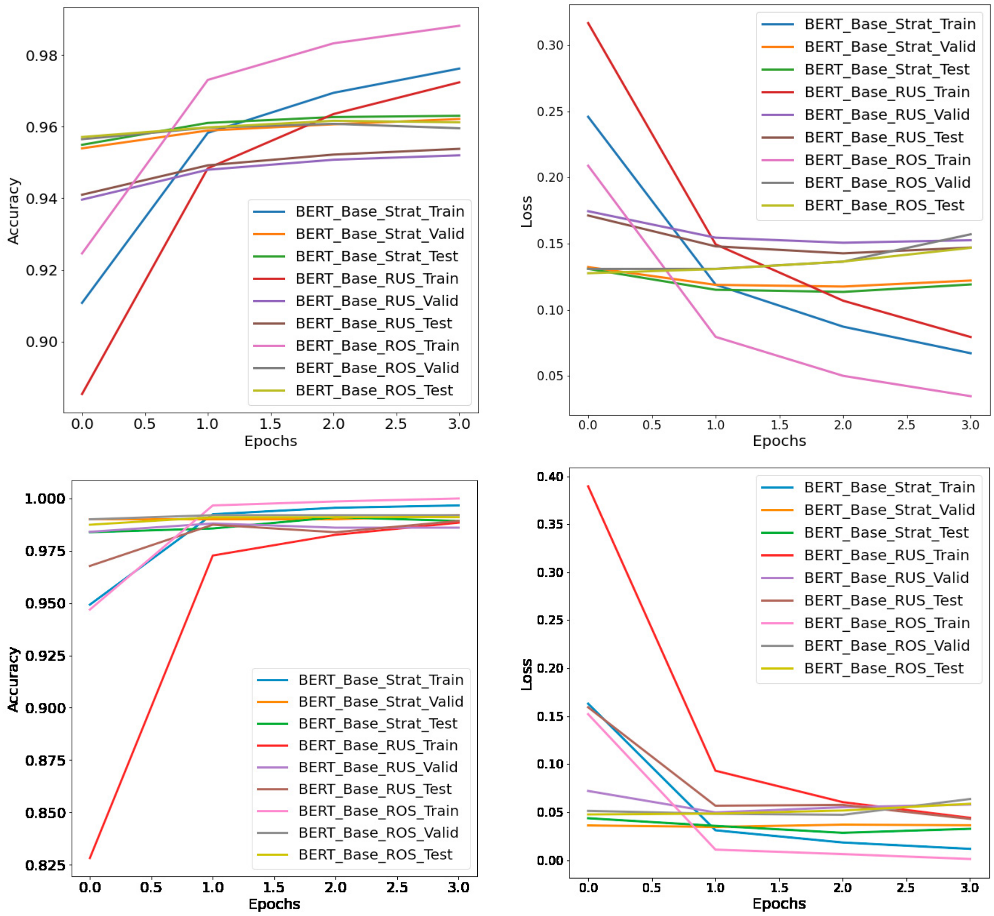

Fig. 3. The overall accuracy and loss curves for news aggregator data and spam data

Table 1. News Aggregator Data Distribution

Categories	#documents	Stratified Training	Undersampling Training	Oversampling Training	Validation	Testing
Business	115967	93933	36967	123500	10437	11597
Entertainment	152469	123500	36967	123500	13722	15247
Health	45639	36967	36967	123500	4108	4564
Technology	108344	87759	36967	123500	9751	10834
Total	422419	342159	147868	494000	38018	42242

Table 2. SMS Spam Collection Data Distribution

Categories	#documents	Stratified Training	Undersampling Training	Oversampling Training	Validation	Testing
Ham	4825	3907	605	3907	435	483
Spam	747	605	605	3907	67	75
Total	5572	4512	1210	7814	502	558

Table 3. The Model Performance for Training

Data	Categories	Stratified			Undersampling			Oversampling		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
News Aggregator	Business	0.9771	0.9817	0.9794	0.9761	0.9800	0.9781	0.9891	0.9933	0.9912
	Entertainment	0.9969	0.9968	0.9968	0.9962	0.9961	0.9961	0.9986	0.9969	0.9978
	Health	0.9928	0.9896	0.9912	0.9951	0.9931	0.9941	0.9982	0.9984	0.9983
	Technology	0.9825	0.9790	0.9808	0.9824	0.9806	0.9815	0.9941	0.9914	0.9928
	Macro-score	0.9873	0.9868	0.9870	0.9874	0.9874	0.9874	0.9950	0.9950	0.9950
	Micro-score	0.9873	0.9873	0.9873	0.9874	0.9874	0.9874	0.9950	0.9950	0.9950
	Weighted-score	0.9873	0.9873	0.9873	0.9874	0.9874	0.9874	0.9950	0.9950	0.9950
SMS Spam Collection	Ham	0.9985	1.0000	0.9992	0.9902	0.9983	0.9942	1.0000	1.0000	1.0000
	Spam	1.0000	0.9901	0.9950	0.9983	0.9901	0.9942	1.0000	1.0000	1.0000
	Macro-score	0.9992	0.9950	0.9971	0.9942	0.9942	0.9942	1.0000	1.0000	1.0000
	Micro-score	0.9987	0.9987	0.9987	0.9942	0.9942	0.9942	1.0000	1.0000	1.0000
	Weighted-score	0.9987	0.9987	0.9987	0.9942	0.9942	0.9942	1.0000	1.0000	1.0000

* P-Precision, R-Recall, F-F1-measure

Table 4. The Model Performance for the Validation Data

Data	Categories	Stratified			Undersampling			Oversampling		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
News Aggregator	Business	0.9440	0.9494	0.9467	0.9369	0.9336	0.9353	0.9422	0.9454	0.9438
	Entertainment	0.9834	0.9822	0.9828	0.9830	0.9734	0.9782	0.9819	0.9830	0.9825
	Health	0.9651	0.9547	0.9599	0.9236	0.9623	0.9425	0.9565	0.9630	0.9597
	Technology	0.9506	0.9506	0.9506	0.9373	0.9372	0.9373	0.9518	0.9442	0.9480
	Macro-score	0.9607	0.9592	0.9600	0.9452	0.9516	0.9483	0.9581	0.9589	0.9585
	Micro-score	0.9621	0.9621	0.9621	0.9520	0.9520	0.9520	0.9606	0.9606	0.9606
	Weighted-score	0.9622	0.9621	0.9621	0.9522	0.9520	0.9521	0.9606	0.9606	0.9606
SMS Spam Collection	Ham	0.9931	0.9977	0.9954	1.0000	0.9839	0.9919	0.9909	1.0000	0.9954
	Spam	0.9846	0.9552	0.9697	0.9054	1.0000	0.9504	1.0000	0.9403	0.9692
	Macro-score	0.9889	0.9765	0.9826	0.9527	0.9920	0.9711	0.9954	0.9701	0.9823
	Micro-score	0.9920	0.9920	0.9920	0.9861	0.9861	0.9861	0.9920	0.9920	0.9920
	Weighted-score	0.9920	0.9920	0.9920	0.9874	0.9861	0.9863	0.9921	0.9920	0.9919

* P-Precision, R-Recall, F-F1-measure

Table 5. The Model Performance for Testing

Data	Categories	Stratified			Undersampling			Oversampling		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
News Aggregator	Business	0.9454	0.9521	0.9487	0.9365	0.9340	0.9353	0.9467	0.9469	0.9468
	Entertainment	0.9861	0.9829	0.9845	0.9828	0.9735	0.9781	0.9809	0.9856	0.9833
	Health	0.9617	0.9628	0.9622	0.9279	0.9671	0.9471	0.9643	0.9586	0.9614
	Technology	0.9567	0.9533	0.9550	0.9409	0.9394	0.9401	0.9509	0.9466	0.9487
	Macro-score	0.9625	0.9628	0.9626	0.9470	0.9535	0.9502	0.9607	0.9594	0.9600
	Micro-score	0.9647	0.9647	0.9647	0.9532	0.9532	0.9532	0.9621	0.9621	0.9621
	Weighted-score	0.9647	0.9647	0.9647	0.9534	0.9532	0.9533	0.9620	0.9621	0.9620
SMS Spam Collection	Ham	0.9918	0.9979	0.9948	0.9979	0.9896	0.9938	0.9898	1.0000	0.9949
	Spam	0.9861	0.9467	0.9660	0.9367	0.9867	0.9610	1.0000	0.9333	0.9655
	Macro-score	0.9889	0.9723	0.9804	0.9673	0.9882	0.9774	0.9949	0.9667	0.9802
	Micro-score	0.9910	0.9910	0.9910	0.9892	0.9892	0.9892	0.9910	0.9910	0.9910
	Weighted-score	0.9910	0.9910	0.9910	0.9897	0.9892	0.9894	0.9911	0.9910	0.9909

* P-Precision, R-Recall, F-F1-measure

3.2. Sampling Methods

Sampling methods help the researchers to infer or investigate information from a subset rather than investigating the whole dataset. Specifically, it reduces the bias in the dataset to have a balanced distribution [5, 24]. In this paper, the stratified random sampling, random undersampling, and random oversampling methods are studied for class imbalance problems.

3.2.1. Stratified Random Sampling

The stratified random sampling classifies or separates the instances or documents into groups or strata based on some characteristics such as business, health, technology, or entertainment related news. These groups are referred to as subsets or subgroups. Random sampling is applied to each group to represent subgroups based upon the percentage. It helps to reduce the bias in the document selection. Therefore, the stratified random sampling method is more accurate than simple random sampling [5, 24].

3.2.2. Random Undersampling

Random undersampling technique randomly deletes the majority class examples in the training data until to have a balanced distribution in the minority and majority classes [**?** 5, 24]. In this technique, there is a loss of information from the majority class that information may be important or critical to fit the decision boundary. Also, one cannot detect or preserve what type of information is thrown in the majority class. However, the RUS technique outperforms in some of the empirical studies.

3.2.3. Random Oversampling

This sampling technique randomly increases the training data based on the minority examples until to have a balanced distribution in the minority and majority classes [5, 8, 24]. In this sampling technique, there is no loss of information in the training data. Moreover, it increases the training time and memory power due to its time and space complexity. In particular, one can choose examples from the training data and one can over-sample the examples with replacement.

3.3. BERT

In this section, the BERT model is presented in the task of imbalanced aspect categorization. Traditionally, the RNNs like LSTM [10] and GRU [11] process an input sequence step by step (or token by token). These networks fail to take the entire input sequence at a time. Therefore, a transformer-based BERT model is introduced as a big improvement for text classification. Recently, this model has shown the better performance in various NLP tasks [12]. Specifically, the BERT architecture is designed as a deep or multi-layer bidirectional encoder representation based on transformers [13]. It mainly trains a large unsupervised (or unlabeled) text by considering the previous and next context information of a word in all layers [25]. The BERT architecture involves two steps, namely, the pre-training language model and the fine-tuning model. Each of these steps explained as follows.

3.3.1. Pre-trained Language Model

The BERT architecture is pre-trained for two different tasks, namely, the masked language (ML) model and the next sentence prediction (NSP) model. The ML model predicts randomly masked or replaced words within the sequence itself. Specifically, this model learns the relationship between words. The NSP model predicts the next sentence in a sentence pair. This model learns the relationships between sentences [25, 26].

3.3.2. BERT Fine-tuning

The BERT pre-trained model is used to train the model on smaller data. This fine-tuning process can be implemented with different techniques such as training the whole architecture, training only on some layers, or freezing the whole architecture [26]. The researchers indicated that the BERT-based fine-tuning model outperforms the existing state-of-the-art results in various NLP tasks. Therefore, this paper explores the BERT-based fine-tuning model for the task of imbalanced aspect categorization.

3.4. BERT for Imbalanced Aspect Categorization

The BERT model takes the whole sequence as an input at once, and it enables all input words or tokens in a parallel way. Specifically, the BERT model is built with two variants, namely, BERT-Base and BERT-large. These models were individually trained on the upper-cased and lower-cased English text. First, the BERT-base model is built with 12 blocks of transformer, 768 units of hidden state, and 12 heads of self-attention, and 110 million of trainable

Table 6. Result Comparison for the news aggregator data

Author	Method	Performance
Mehta et al. [14]	BiGRU	90.5
	CNN	91.4
	TE	89.9
	BERT	92.0
	LAMA	92.2
	LAMA+Ctx	92.3
Pushp et al. [15]	Arch1_FCL	61.7
	Arch2_LSTM_FCL	63.0
	Arch2_LSTM_TE	64.2
Luis Bronchal [16]	LR	94.7
Chemchem et al. [17]	NB	86.2
	MNB	84.7
	Linear SVM	87.3
	MLNN	84.4
	CNN	76.2
Akritidis et al. [18]	LR	95.0
Poojitha Bikki [19]	LR	94.2
	Linear SVM	95.0
	MNB	93.4
	RF	79.2
Proposed	BERT_Base_Strat_FT	96.5
	BERT_Base_RUS_FT	95.3
	BERT_Base_ROS_FT	96.2

parameters. Second, the BERT-large fine-tuning model is built with 24 blocks of transformer, 1024 units of hidden states, and 16 heads of self-attention, and 340 million of trainable parameters. The architecture of the BERT layers [12, 25] is shown in Fig. 2. Each layer of the BERT architecture consists of two-sub layers, namely, an MSA layer and a fully connected feed-forward neural network (FCFFNN). The MSA mechanism is learned information at different positions. The FCFFNN is applied at different positions with ReLU activation function for controlling the information in one direction. Moreover, a layer normalization is employed around each of the transformer block as a residual connection [12, 13].

In particular, BERT models take 512 tokens as input sequence length to output the input sequence representation. The input sequence to BERT can be a single or two-sentence pair. Each input sequence always starts with a special classification token $[CLS]$ and the sentence pairs are differentiated with $[SEP]$ token. However, the input representation of a given token can be constructed by adding three different embedding such as a token, segment, and position. In this paper, a BERT-base model is adopted to deal with the imbalanced aspect categorization problem. Finally, a softmax layer is applied to predict the probability of aspect categories on the top of the BERT-base model as in equation (1).

$$p(c/h) = softmax(Wh) \tag{1}$$

Where c refers to the aspect categories and W represents the parameter matrix of a specific-task.

4. Results and discussions

We conducted experiments in Google Colaboratory using Keras API and Tensorflow libraries with P100 GPU and 24GB RAM [27]. Specifically, the UCI news aggregator and SMS spam collection datasets are used for imbalanced aspect categorization. The news aggregator dataset contains 422,419 news pages in four categories, namely, business (115,967), health (45,639), entertainment (152,469), and technology (108,344). On the other hand, the SMS spam

Table 7. Result Comparison for the SMS Spam Collection data

Author	Method	Performance
Liu et al. [29]	LR	94.7
	NB	94.3
	RF	92.1
	SVM	94.9
	LSTM	94.9
	CNN-LSTM	91.8
	Spam Transformer	96.1
Proposed	BERT_Base_Strat_FT	99.1
	BERT_Base_RUS_FT	98.9
	BERT_Base_ROS_FT	99.1

collection dataset contains 5572 news pages in two categories, namely, ham (4825) and spam (747). In these datasets, the labeled categories have occurred in an imbalanced manner. Therefore, different sampling techniques are applied to deal with these imbalanced categories. Initially, the stratified sampling technique is applied to split the given datasets into training (80%), validation (10%), and testing (10%). Each of these data has equal distribution in all categories such as 27% entertainment, 36% business, 11% technology, 26% health category in the news aggregator data, and 87% for ham and 13% for spam in the SMS collection data. Then, the random undersampling and oversampling techniques are applied to the training data to have a balanced distribution as shown in Table 1 and Table 2. Moreover, the BERT-base fine-tuning model is employed on the stratified, undersampling, and oversampling training data separately. These models were evaluated on validation and testing data. The one cycle learning rate policy [28] is used to train, validate, and test the model with a $2e^{-5}$ learning rate, 4 cycles, 20000 maximum features, unigram words, and maximum sequence length 64. The overall accuracy and loss curves for training, validation, and testing with stratified sampling, random undersampling, and random oversampling are shown in Fig.3 and Fig. 4 for both datasets. These figures indicate that random oversampling learns well in training and stratified sampling performs well in the validation and testing. In particular, the standard evaluation metrics such as precision, recall, F1-score, and macro, micro, and weighted scores were used to measure the performance of the models. Table 3-5 shows the evaluation scores for training, validation, and testing with stratified, undersampling, and oversampling. In the training data, the fine-tuning model produces almost similar results with stratified (98.73%) and undersampling (98.74%) techniques, and higher results with the oversampling (99.50%) techniques for the news aggregator data. Similarly, the fine-tuning model produces almost similar results with stratified (99.87%) and oversampling (100%) techniques, and lower results with undersampling (99.42%) techniques for SMS spam collection data. However, the proposed fine-tuned model comparatively achieves better results with the SRS technique in the validation (96.21%) and testing (96.47%) for the news aggregator data. On the other hand, the stratified and oversampling techniques achieve similar validation (99.20%) and testing (99.10%) results for the SMS spam collection data. Furthermore, the sampling-based BERT-base fine-tuning model is compared with others who have used the same dataset as shown in Table 5 and Table 6. Our results outperform than Luis Bronchal [16], Chemchem et al. [17], Akritidis et al. [18], Poojitha Bikki [19], Mehta et al. [14], Pushp et al. [15], and Liu et al. [29]. Overall, the proposed BERT-base fine-tune model outperforms others.

5. Conclusion

Online news websites identify and publishes new articles based on the internet users or visitors' interest. In this paper, the imbalanced aspect categorization task is addressed using BERT fine-tuning model. The proposed BERT fine-tuning model learns context-dependent features. Specifically, the stratified sampling, random undersampling, and random oversampling techniques are used to deal with the imbalanced aspects. Empirically, the results show that the BERT-Base fine-tuning model with a stratified sampling technique achieves a better result (96.5%) than the existing models. In the future, the proposed model is planned to study with the information fusion and gender-based imbalanced categories. Also, the model can be extended in parallel and distributed environment using graph transformer networks.

Acknowledgements

This work was supported by the University Grants Commission (UGC), Government of India, New Delhi, under the National Doctoral Fellowship.

References

[1] Schouten, K., Van Der Weijde, O., Frasincar, F., & Dekker, R. (2017). Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. IEEE transactions on cybernetics, 48(4), 1263-1275.
[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in information retrieval, 2(1–2), 1-135.
[3] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent systems, 28(2), 15-21.
[4] Wang, S., Minku, L. L., & Yao, X. (2014). Resampling-based ensemble methods for online class imbalance learning. IEEE Transactions on Knowledge and Data Engineering, 27(5), 1356-1368.
[5] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.
[6] Cao, B., Liu, Y., Hou, C., Fan, J., Zheng, B., & Yin, J. (2020). Expediting the accuracy-improving process of svms for class imbalance learning. IEEE Transactions on Knowledge and Data Engineering, 33(11), 3550-3567.
[7] Fernandez, A., García, S., & Herrera, F. (2011, May). Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In International conference on hybrid artificial intelligence systems (pp. 1-10). Springer, Berlin, Heidelberg.
[8] Sen, A., Islam, M. M., Murase, K., & Yao, X. (2015). Binarization with boosting and oversampling for multiclass classification. IEEE transactions on cybernetics, 46(5), 1078-1091.
[9] Wang, S., Minku, L. L., & Yao, X. (2018). A systematic study of online class imbalance learning with concept drift. IEEE transactions on neural networks and learning systems, 29(10), 4802-4821.
[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[11] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Workshop on Deep Learning, December 2014.
[12] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).
[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[14] Mehta, S., Rangwala, H., & Ramakrishnan, N. (2019). Low rank factorization for compact multi-head self-attention. arXiv preprint arXiv:1912.00835.
[15] Pushp, P. K., & Srivastava, M. M. (2017). Train once, test anywhere: Zero-shot learning for text classification. arXiv preprint arXiv:1712.05972.
[16] Luis Bronchal (2017). Classifying with Logistic Regression. Kaggle, URL: https://www.kaggle.com/lbronchal/classifying-with-logistic-regression-0-9473.
[17] Chemchem, A., Alin, F., & Krajecki, M. (2018, August). Deep learning and data mining classification through the intelligent agent reasoning. In 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (pp. 13-20). IEEE.
[18] Akritidis, L., Fevgas, A., Bozanis, P., & Alamaniotis, M. (2019, July). A Self-Pruning Classification Model for News. In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-6). IEEE.
[19] Bikki, P (2018). Machine learning for text categorization: experiments using clustering and classification.
[20] C Lin, M. (2018, January). Active learning with unbalanced classes & example-generated queries. In AAAI Conference on Human Computation.
[21] Suh, Y., Yu, J., Mo, J., Song, L., & Kim, C. (2017). A comparison of oversampling methods on imbalanced topic classification of Korean news articles. Journal of Cognitive Science, 18(4), 391-437.
[22] Gasparetti, F. (2017). Modeling user interests from web browsing activities. Data mining and knowledge discovery, 31(2), 502-547.
[23] Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011, September). Contributions to the study of SMS spam filtering: new collection and results. In Proceedings of the 11th ACM symposium on Document engineering (pp. 259-262).
[24] Rathpisey, H., & Adji, T. B. (2019, October). Handling imbalance issue in hate speech classification using sampling-based methods. In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 193-198). IEEE.
[25] Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using BERT. In 2019 Artificial Intelligence for Transforming Business and Society (AITB) (Vol. 1, pp. 1-5). IEEE.
[26] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In China national conference on Chinese computational linguistics (pp. 194-206). Springer, Cham.
[27] Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners. Apress.
[28] Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820.
[29] Liu, X., Lu, H., & Nayak, A. (2021). A spam transformer model for SMS spam detection. IEEE Access, 9, 80253-80263.