MULTI-VIEW APPROACH WITH TRANSFORMER MODELS AND AUGMENTED EMBEDDINGS FOR TACKLING IMBALANCED MULTI-LABEL DATASETS

Michael Abobor and Darsana P. Josyula

Department of Computer Science, Bowie State University, Bowie, USA

ABSTRACT

Imbalanced datasets present significant challenges in machine learning. The disproportionate distribution of labels in imbalanced multi-label datasets is a result of the low datapoints of the minority class. This leads to biases in model predictions as algorithms tend to favor the majority class, resulting in poor generalization for the minority class. Any effort to balance the inequality within each class can inadvertently create issues across the other classes. This paper introduces the multi-view learning approach that combines pre- trained large language models and embeddings augmented with techniques such as SMOTE, MLeNN, MLSMOTE, MLSOL, and MLTL. This helps address the issue of imbalanced multi-label datasets in classification. This dual input model combines the original tokenized text, and the augmented embeddings extracted from the penultimate layer of the transformer, giving the model the ability to learn from both sources of information.

This approach conserves the contextual significance of the input text and makes it possible for training transformers with the augmented embeddings thereby tackling the issue of imbalance multi-class datasets.

KEYWORDS

Imbalanced datasets, Multi-label, Transformer, Augmented Embeddings, Machine Learning

1. INTRODUCTION

Several real-world scenarios use text classification but the number of data points in some classes are smaller than others [1]. Multi-label dataset poses a unique problem due to the complexity of handling the data points as each instance can exist in multiple classes simultaneously. Imbalanced datasets affect the predictions of models due to the fact that, the model is biased towards the majority class thereby affecting the minority class [2]. Attempts to balance the inequality within each class of labels individually can create issues affecting all the other labels. Several data augmentation techniques such as SMOTE and MLeNN have been proposed to solve this issue but they are typically applied to the raw vectors rather than deep learning embeddings.

Our approach in this research paper consists of fine-tuning a distilBERT model on the embeddings extracted from the penultimate layer. Imbalanced dataset of social engineering

David C. Wyld et al. (Eds): CSITAI, MATHCS, ITCAU, COMSAP, ACINT - 2024
pp. 01-12, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142501

features annotated by undergraduate students was used in this research. These extracted embeddings were processed by applying augmentation techniques such as SMOTE (Syn- thetic Minority Over-sampling Technique), MLeNN (Multi- label Edited Nearest Neighbors), MLSMOTE (Multi-label SMOTE), MLSOL (Multi-label Synthetic Over- sampling via label correlation), and MLTL (Multi-label Transfer Learning). We propose a multi-view learning framework that combines the output embeddings of a distilBERT model with synthetically generated augmented embeddings. Our motivation behind this approach is to increase the minority class by synthesizing new data points at the embedding level which can ultimately be fed into a deep learning classifier. This double input hybrid architecture takes advantage of both the rich context captured by the pre-trained transformer model and the diversity added by the synthetic augmentation techniques.

A key challenge in using augmentation techniques on embeddings is that transformer models like DistilBERT require tokenized text to train them and not embeddings as input. This helped us in harnessing the power of the transformer model as well as performing augmentation on our imbalanced dataset. By leveraging trans- former distilBERT model to pro- cess raw text alongside augmented embeddings, our multi- view framework tackles this issue of imbalance multi-class datasets. Through experiments, we have demonstrated the positive implications of performing data augmentation techniques on extracted embeddings coupled with that of embeddings of raw SMS text. Dimensionality reduction was incorporated and T-SNE graphs plotted. By leveraging trans- former distilBERT model to process raw text input tokens alongside augmented embeddings, we provide a multi-view framework that tackles this issue of imbalance multi-class datasets.

2. RELATED WORK

Data augmentation techniques play a very important role in solving the issue of imbalanced multi-label classification. Nooten et. al. utilized the GPT-3.5 large language model SMOTE is popular and it takes all the samples belonging to the minority class, selects one of the nearest neighbors randomly within the nearest neighbors of each sample and produces a new sample with the same minority class. In other words, the features of the synthetic instances are obtained by interpolation of values belonging to nearest neighbors. Adjustments can be made to both the number of nearest neighbors as well as the quantity of synthetic instances used for each minority sample[6].

Multi-label Synthetic Minority Over-sampling Technique (MLSMOTE) is an algorithm that produces synthetic rows of imbalanced multi-label dataset. Among the existent resampling techniques, those based on the generation of new samples (oversampling) have proved to work better than others. The new samples can be clones of existent ones, or be synthetically produced as in SMOTE (Synthetic Minority Over- sampling Technique). MLSMOTE extends SMOTE to multi- label classification problems in which one instance of the data belongs to several classes simultaneously. MLSMOTE takes into consideration label correlations when generating synthetic samples. This to ensure and preserve multi-labels relationships [6].

In the Multi-Label Edited Nearest Neighbor (MLeNN), a heuristic multi-label undersampling algorithm based on the well-known Wilson's Edited Nearest Neighbor Rule is used. Removal of the samples is by heuristic and random selection. MLeNN is a very competitive multi-label undersampling alternative [7].

Multi-Label Synthetic Oversampling based on Local label imbalance (MLSOL) aims to generate synthetic samples by utilizing the layout of the minority class through clustering. By identifying the clusters within the minority classes, the approach generates synthetic instances that are used for the minority class. MLSOL produces these synthetic data by creating ”links” between similar minority class instances. The goal is to generate samples that retain relationships between the features and labels while ensuring diversity within the minority class.

Multi-Label Synthetic Oversampling based on Local label imbalance (MLSOL) technique considers all informative la- bels. It comes up with a more divers and effective labelled synthetic instances for difficult examples. MLSOL creates new instances near difficult to learn examples [8].

Multi-Label Synthetic Oversampling based on Local label imbalance (MLSOL), that generates synthetic instances near those instances that are suffering high local imbalance and are more difficult to be predicted correctly. MLSOL combines the local imbalance of informative labels to pick up difficult seed instances. It assigns appropriate labels to the created synthetic instances so as to improve the frequency of difficult labels, without introducing noise labels [8].

Multi-Level Threshold Learning (MLTL) focuses on threshold-based learning. It generates synthetic data by adjusting the decision thresholds of minority class instances in the dataset. MLTL - Multi-Label Tomek Link is based on the standard Tomek Link resampling algorithm. A multi- label imbalance API for the Mulan framework. Research from seven well-known datasets showed that MLTL is a great augmentation technique [9].

3. METHODOLOGY

In previous research where we used Transformer models to identify social engineering features, We had three under- graduate students annotate our dataset of 9754 SMS messages. A final dataset was obtained by voting to select annotations for traits that were differently annotated by the students. We trained the model with 7000 of the data and used 2754 for testing. We realized that this multi-label dataset was imbalanced. We utilized data augmentation techniques to balance the dataset. We chose 6 out of the 13 social engineering features. Pretexting, Baiting, Urgency, Consensus, Authority, and Familiarity were chosen out of the tactics as they had enough data that we could work with. We still believed more work could be done to balance this multi-label dataset. Familiarity had the highest amount of the positive class of Yes with a count of 3481. This was the most commonly used tactic by bad actors globally. Baiting was the next with a count of 810 positive class. Followed by consensus with a count of 454. Urgency, Authority, and Pretexting had counts of 183, 174, and 143 respectively. We realized we had to work to resolve the imbalance multi-label dataset we had created.

In this research, we propose a **Multi-view Learning Framework** coupled with data augmentation techniques. In this multi-view learning framework, data is fed to the distil- BERT model in two parallel branches:

a. **A DistilBERT Branch -** This branch takes in the original tokenized SMS messages via pre-trained distilBERT model and generates contextual embeddings which we called originally extracted embeddings. All the layers of this module were frozen except the last layer and these embeddings were collected from the penultimate layer [10].
b. **Augmented Embedding Branch -** This branch processes embeddings which have been augmented using these five augmentation techniques: SMOTE, MLSMOTE, MLeNN, MLSOL, and MLTL [11].

We concatenated the output of both branches and passed it through a fully connected classification head which completes the final classification. This framework gives the model the advantage of making use of both the sematic information from the SMS messages directly and the diversity added by synthetic embeddings from the augmentation techniques.

We Trained a distilBERT model with 7000 of the dataset and extracted the embeddings from the penultimate layer to be used for data augmentation. The techniques employed are SMOTE, MLSMOTE, MLeNN, MLSOL, and MLTL.

We utilized dimension reduction techniques and plotted T-SNE for the embeddings and labels after each augmentation technique was completed. We also trained and tested a new distil- BERT model after each augmentation approach to evaluate the progress and performance of each technique.

3.1. Data Augmentation Techniques

We Trained a distilBERT model with 7000 of the dataset and extracted the embeddings from the penultimate layer to be used for data augmentation. The techniques employed are SMOTE, MLSMOTE, MLeNN, MLSOL, and MLTL [11]. We utilized dimension reduction techniques and plotted T-SNE for the embeddings after each augmentation technique was completed. We also trained and tested new distilBERT models after each augmentation approach to evaluate the progress and performance of each technique.

3.2. Model

We used a transformer model distilBERT which is a distilled version of a BERT model. This model uses self-attention to capture relationship between words in sentences contextually. The transformer model has a stack of self-attention and point- wise fully connected layers present in an encoder and decoder [10].

The transformer has multiple layers of self- attention blocks which helps it weigh the significance of different words based on their context. Positional encoding to preserve the order of words because the self-attention does not preserve the order in which words are fed to the model [12].

We fine-tuned a distilBERT transformer model with our annotated social engineering SMS messages by freezing all the layers of the model and extracting the embeddings from the penultimate layer. We applied the augmentation techniques on these embeddings.

3.3. Dataset

In this research, we used a dataset that captured social engineering features in SMS messages. We had undergraduate students annotate SMS messages with 13 different classes. We used 7000 out of the 9754 rows of SMS text messages for this research which originally had 13 labels with one column being SMS. The social engineering features spanning the 13 columns include Pretexting, Baiting, Blackmailing, Quid Pro Quo, Reciprocity, Commitment, Intimidation, Urgency, Consensus, Authority, Familiarity, Scarcity, and Hyperlink. During the annotation, whenever there were different annotations, a voting scheme was used to finalize what the annotation had to be.

We realized there was data imbalance as some of the social engineering features were not popular in the dataset. As a result, we scaled down the number of social engineering features that our model would predict. We then reduced the classes from 13 to 6. The 6 features identified from the text messages are: Pretexting, Baiting, Urgency, Consensus, Authority and Familiarity.

3.4. Baseline and Models

We plotted a T-SNE plot of the originally extracted embed- dings from the penultimate layer of the distilBERT model after freezing all the layers except the last layer. The plots showed less clustering of the minority class depicted by green which is the presence of social engineering features. The counts after applying each augmentation technique to the extracted embeddings was capture for comparison. We also plotted T-SNEs after each augmentation technique to compare. We compared the following:

1. **Baseline (DistiBERT): DistilBERT fine-tuned on the original dataset without any augmentation.**
2. **DistilBERT and SMOTE Augmented Embeddings.**
3. **DistilBERT and MLeNN Augmented Embeddings.**
4. **DistilBERT and MLSMOTE Augmented Embeddings.**
5. **DistilBERT and MLSOL Augmented Embeddings.**
6. **DistilBERT and MLTL Augmented Embeddings.**

4. Results

The Abstract section begins with the word, "Abstract" in 13 pt. Times New Roman, bold italics, "Small Caps" font with a 6pt. spacing following. The abstract must not exceed 150 words in length in 10 pt. Times New Roman italics. The text must be fully justified, with a 12 pt. paragraph spacing following the last line.

The results show T-SNE plots of the augmented embed- dings. We also showed a count of the minority and majority class before and after the augmentations were conducted. A comparison of the accuracy and loss of the Multi-view learning framework on imbalanced multi-class dataset showed improvement with the augmented embeddings.

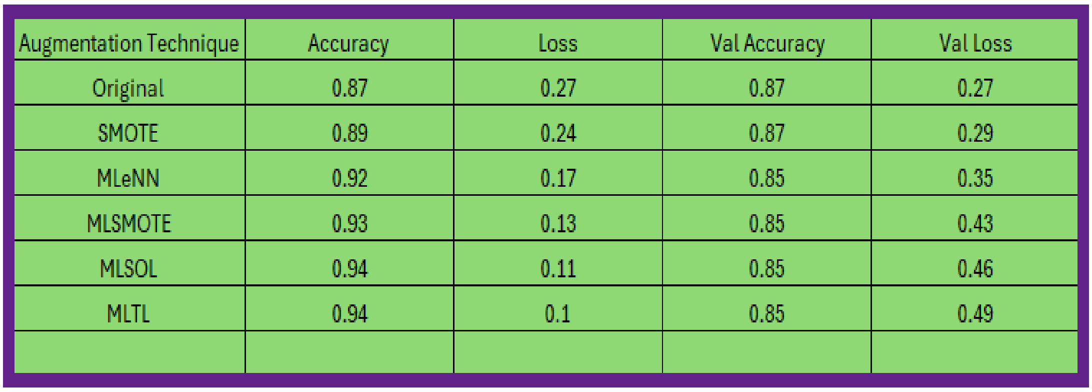
After applying augmentation techniques SMOTE, MLeNN, MLSMOTE, MLSOL, and MLTL on the extracted embed- dings, we represented the clustering of the newly augmented dataset using T-SNE plots. Counts of the original data points before and after were also captured.

Figures 2 through 7 show the T-SNE plots of the social engineering class labels without any augmentation techniques performed. Figures 10 through 14 show the T-SNE plots after augmentations were performed on the embeddings. Figure

1 shows the accuracies and losses of training our multi- view hybrid transformer with the original and augmented embeddings. Figure 8 and 9 show the counts of the labels before and after the augmentations were performed.

4.1. Figures and Tables

Table 1. Accuracies of the models after training with the augmented datasets.



Augmentation Technique	Accuracy	Loss	Val Accuracy	Val Loss
Original	0.87	0.27	0.87	0.27
SMOTE	0.89	0.24	0.87	0.29
MLeNN	0.92	0.17	0.85	0.35
MLSMOTE	0.93	0.13	0.85	0.43
MLSOL	0.94	0.11	0.85	0.46
MLTL	0.94	0.1	0.85	0.49

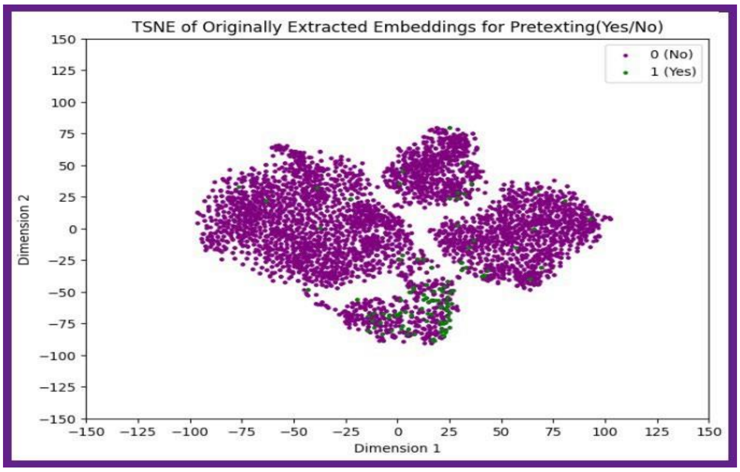


Figure 1. T-SNE Plot of Pretexting Class – Original Embeddings

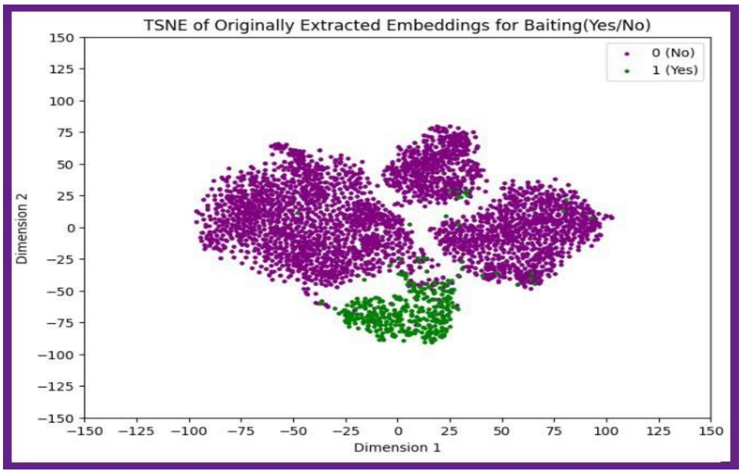


Figure 2. T-SNE Plot of Baiting Class – Original Embeddings

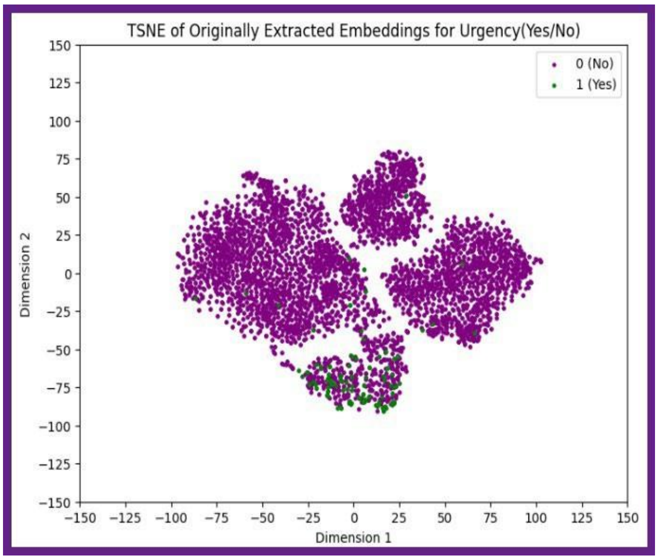


Figure 3. T-SNE Plot of Urgency Class – Original Embeddings

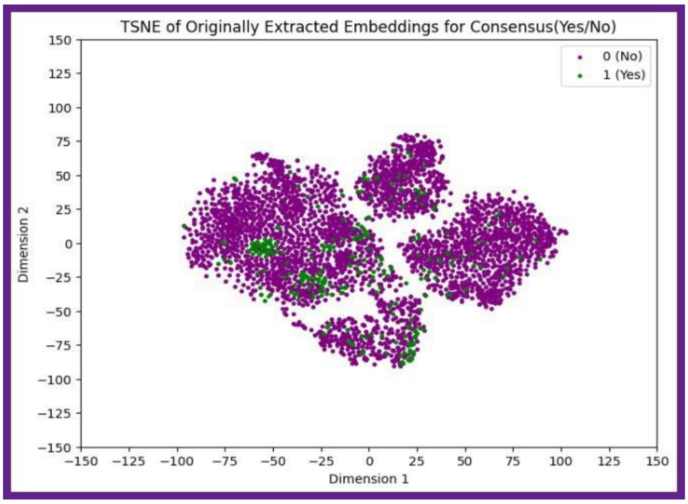


Figure 4. T-SNE Plot of Consensus Class – Original Embeddings

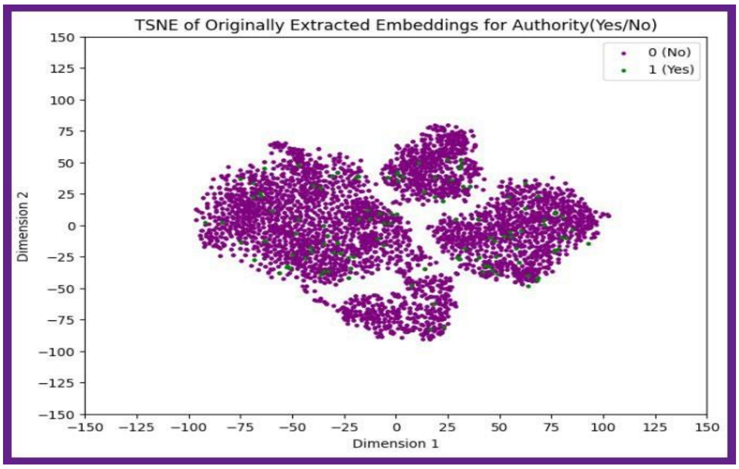


Figure 5. T-SNE Plot of Authority Class – Original Embeddings

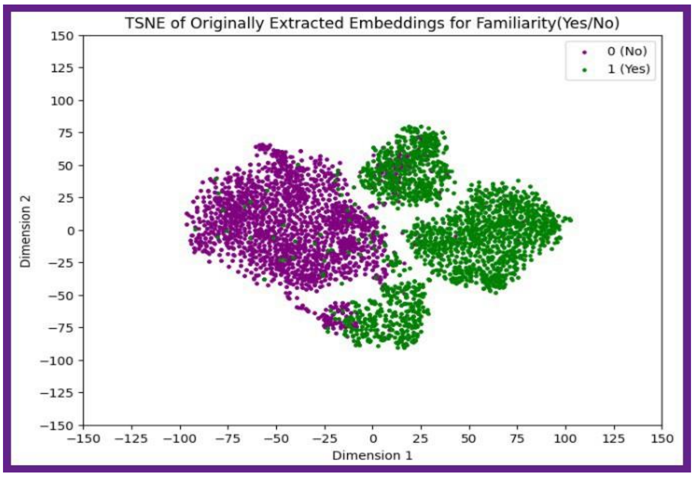


Figure 6. T-SNE Plot of Familiarity Class – Original Embeddings

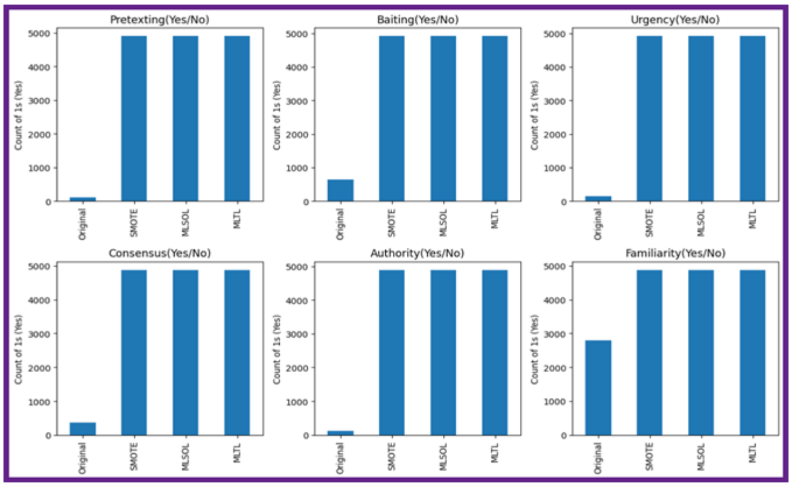


Figure 7. Counts of minority class “Yes” after data augmentation.

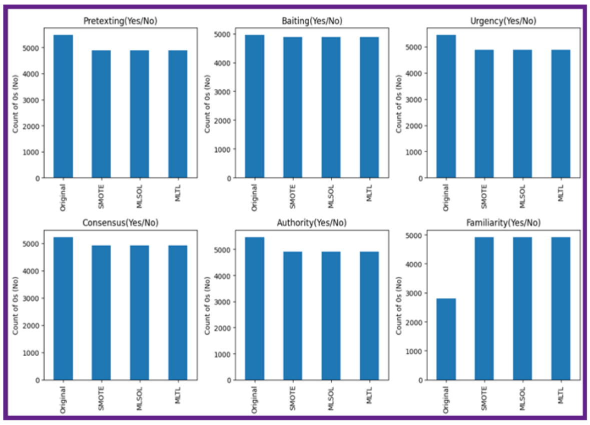


Figure 8. Counts of Majority class “Yes” after data augmentation.

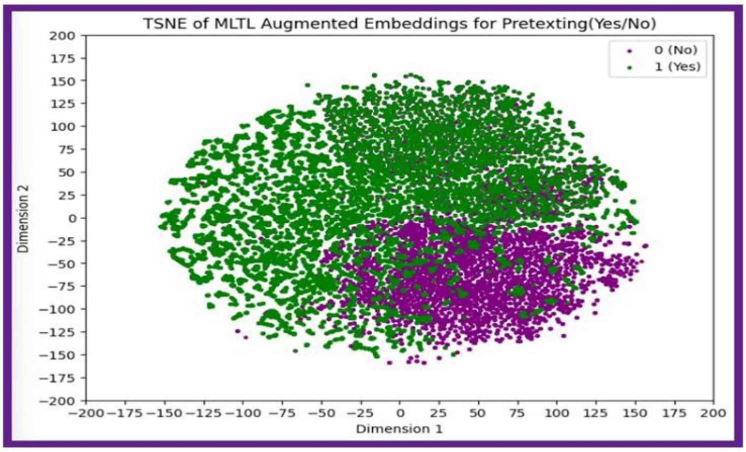


Figure 9. T-SNE Plot of Pretexting after data augmentation applied on embeddings - MLTL.

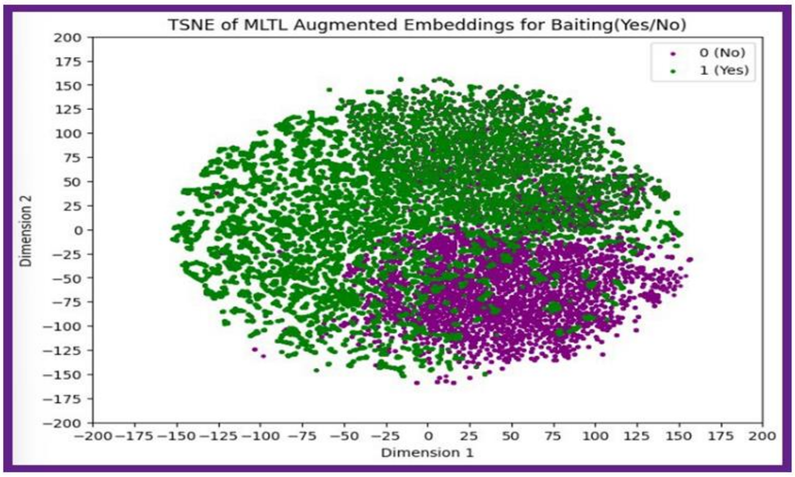


Figure 10. T-SNE Plot of Baiting after data augmentation applied on embeddings - MLTL.

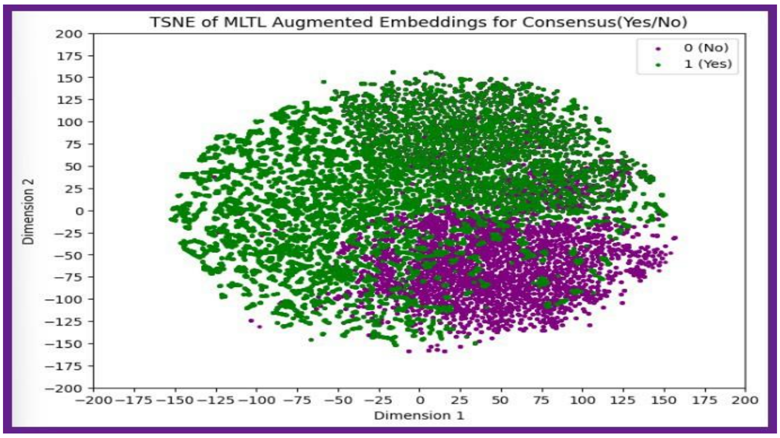


Figure 11. T-SNE Plot of Consensus after data augmentation applied on embeddings - MLTL.

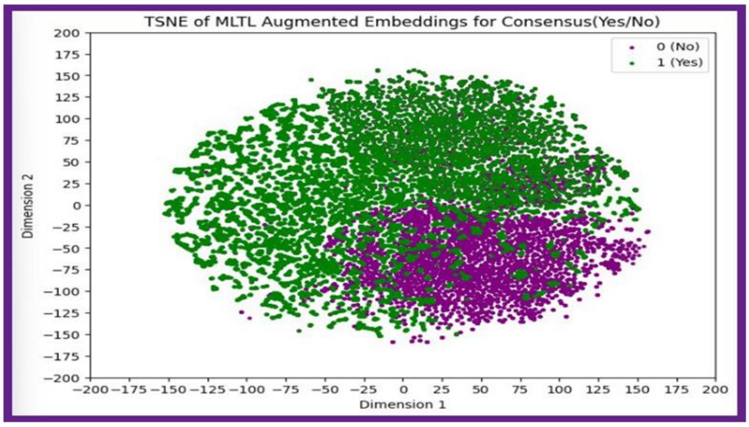


Figure 12. T-SNE Plot of Consensus after data augmentation applied on embeddings - MLTL.

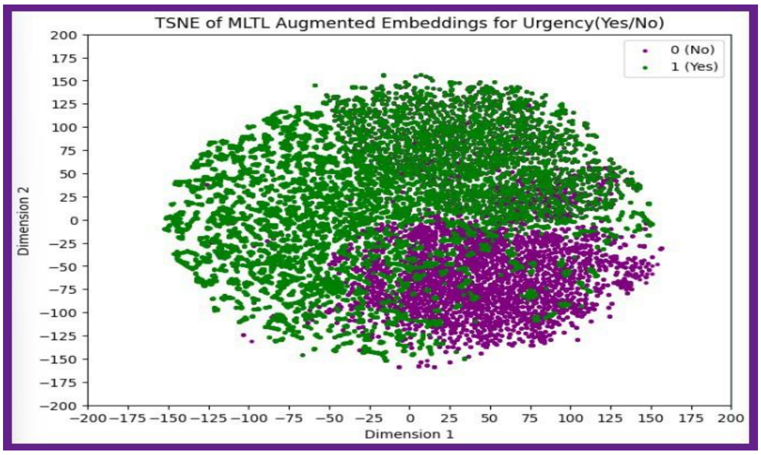


Figure 13. T-SNE Plot of Urgency after data augmentation applied on embeddings - MLTL

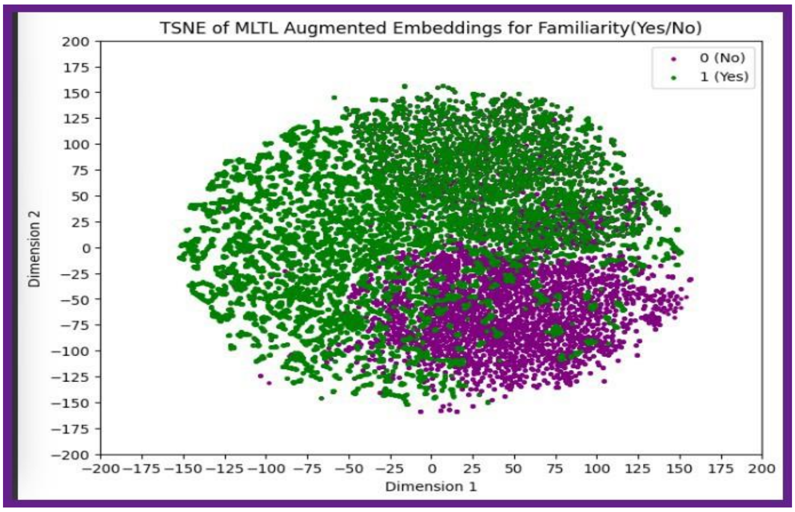


Figure 14. T-SNE Plot of Familiarity after data augmentation applied on embeddings - MLTL.

5. DISCUSSION

The results showed that all the augmentation techniques had improved performances when the hybrid model was trained over when trained on originally extracted embedding. MLTL augmented embeddings had the highest accuracy of 94 percent. This can also be observed from the T-SNE plots of the different features. Figures 2 through 7 shows the T-SNE plots of the embeddings which were originally extracted from the distilBERT model. Comparing the T-SNE plot of the original embeddings using the imbalanced multi-class dataset, we observe the sparse clustering of the minority class which in this case is the minority class shown in green on the plots. Once the embeddings were augmented, the population of the minority class increased and well represented in the space on the plots in figures 10 through14. The accuracies of the multi-view hybrid distilBERT model increased when trained with the augmented embeddings.

6. CONCLUSIONS

Imbalance in multi-class datasets have impacts on machine learning models by extending a favourable bias towards the majority class thereby leaving the minority class to poor generalization. Models are as good as their dataset which was use in training and as such, the issue of imbalance multi-class dataset is an important one to tackle. A key challenge in using augmentation techniques on embeddings is that transformer models like distilBERT require tokenized text to train them and not embeddings directly as input. In this paper, we introduced the dual-input model which combines the original tokenized text and the augmented embeddings which gives our models the ability to learn from both sources of information. This way, we preserve the contextual significance of the input text. This hybrid approach made it possible for us to leverage the power of a transformer model as well as performing augmentation on our imbalanced dataset. This framework combines the original raw text embeddings from distilBERT with augmented embeddings which was obtained by the following techniques, SMOTE, MLeNN, MLSMOTE, MLSOL, and MLTL. Through experiments, we have demonstrated the positive implications of performing data augmentation techniques on extracted embeddings coupled with that of embeddings of raw SMS text.

REFERENCES

[1] L. Zhuang, H. Dai, and X. Hang, “A novel field learning algorithm for dual imbalance text classification,” in *International conference on fuzzy systems and knowledge discovery*, pp. 39– 48, Springer, 2005.

[2] J. Liu, K. Huang, C. Chen, and J. Mao, “An oversampling algorithm of multi-label data based on cluster-specific samples and fuzzy rough set theory,” *Complex & Intelligent Systems*, pp. 1– 16, 2024.

[3] J. Van Nooten and W. Daelemans, “Improving dutch vaccine hesitancy monitoring via multi- label data augmentation with gpt-3.5,” in *Proceed- ings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, July 2023; Toronto, Canada*, vol. 1, pp. 251–270, 2023.

[4] A. Umparat and S. Phoomvuthisarn, “Improving pre-trained models for multi-label classification in stack overflow: A comparison of imbalanced data handling methods,” in *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 464–469, IEEE, 2023.

[5] D. Elreedy and A. F. Atiya, “A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance,” *Information Sciences*, vol. 505, pp. 32–64, 2019.

[6] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, “Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation,” Knowledge-Based Systems, vol. 89, pp. 385–397, 2015.

[7] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, “Mlenn: a first approach to heuristic multilabel undersampling,” in Intelligent Data Engineering and Automated Learning–IDEAL 2014: 15th International Conference, Salamanca, Spain, September 10-12, 2014. Proceedings 15, pp. 1–9, Springer, 2014.

[8] B. Liu, K. Blekas, and G. Tsoumakas, “Multi-label sampling based on local label imbalance,” Pattern Recognition, vol. 122, p. 108294, 2022.

[9] R. M. Pereira, Y. M. Costa, and C. N. Silla Jr, “Mltl: A multi-label approach for the tomek link undersampling algorithm,” Neurocomputing, vol. 383, pp. 95–105, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[12] Y. Lee, J. Saxe, and R. Harang, “Catbert: Context-aware tiny bert for detecting social engineering emails,” 2020.

AUTHOR

Michael, a doctoral candidate in Computer Science, specializing in Artificial Intelligence and Machine Learning, is set to complete his doctorate by December 2024. His research spans natural language processing, computer vision, reinforcement learning, data mining, robotics, predictive modelling, and anomaly detection. He works as an AI Engineer where he designs and implements machine learning models, optimizing algorithms and integrates AI technologies into products and services. He leverages various architectures, including transformers, convolutional neural networks, recurrent neural networks, and several other models to enhance predictive accuracy and performance. Driven by a passion for pushing technological boundaries, Michael aims to make a significant impact in both academia and industry, making life easier through technology for users