Multi-label classification of dengue tweets with text augmentation using BERT

Jeremy Dale Coronia
College of Computer Studies
De La Salle University - Manila
Metro Manila, Philippines
jeremy_coronia@dlsu.edu.ph

Charibeth Cheng
College of Computer Studies
De La Salle University - Manila
Metro Manila, Philippines
charibeth.cheng@dlsu.edu.ph

***Abstract*—Utilizing text from online platforms has become invaluable in fields such as infoveillance, where the monitoring of such can help in the public health domain. In this paper, we build upon previous research work and present our BERT classifier model, using Twitter tweets to help in the early detection of dengue. We use the pre-trained base multilingual BERT model to account for the mix of English-Filipino tweets and perform multi-label classification of tweets. We also use contextual word embeddings via BERT to perform text augmentation to account for low representation in one of the labels. Experiments show very high classification accuracy and low loss as more synthetic samples are added. We confirm that the use of BERT in public health infoveillance is a promising avenue for future work.**

***Index Terms*—natural language processing, classification, dengue, deep learning**

I. Introduction

The wealth of online textual data available in social media platforms has become invaluable in several fields, including infoveillance. Designing and improving upon text classifiers removes the need for manual processing of data given the massive amounts of publicly available text online. In this case, we are referring to Twitter tweets about dengue.

Dengue is a mosquito-born viral disease especially prevalent in tropical countries like the Philippines, and much more in urban areas like Metro Manila. The World Health Organization lists four (4) distinct variations of the dengue virus, meaning it is possible to be infected 4 times. An estimated 100 - 400 million infections are recorded worldwide annually, but early detection and medical accessibility lowers chances of fatalities from severe dengue [1].

One avenue to help in early detection is through posts from social media platforms like Twitter. A report by [2] writes that, as of January 2021, 80.7% of its population of 110.3 million are social media users, which is 21.9% more between 2020 and 2021. Filipinos spend an average time of 4 hours, 15 minutes on social media, with Twitter being the fifth most used platform. For the Philippines which has topped social media usage worldwide for the 6th year, utilizing these online resources may be of considerable value for infoveillance.

There is already established research for using machine learning techniques and exploring feature extraction methods for social media text classification, particularly with deep learning, such as the one by [3] who trained a gated recurrent neural network from a set of human-labelled tweets. This paper aims to build upon their text classifier designed for their public health agent model by using the current state-of-the-art architecture in deep learning, as well as attempt to address the class imbalance present in the data used in training the model. We train a BERT model using the same dataset of tweets and perform text augmentation on the train set to generate synthetic data and increase samples for the under-represented labels. We evaluate and compare our results with a baseline (i.e., no text augmentation) to see any improvements.

II. Related Works

Infoveillance in the public health domain for a number of cases has been experimented with using different types of online sources, such as articles and personal [social media] posts, as well as with different machine learning techniques. For instance, the work of [4] also used online media reports to classify whether or not a given document is about disease activity or not. The goal was to explore efficient information extraction of disease activity from these reports for what is called event-based surveillance, which is disease surveillance based on online reports and other digital information sources. They obtained 0.88 and 0.86 recall and precision, respectively, with an F1 score of 0.87.

On the one hand, [5] conducted a pilot study and used two (2) machine learning algorithms, a convolutional neural network (ConvNet) and a bidirectional long short-term memory (BiLSTM), as well as two (2) classification methods, document-level learning (DocClass) and sentence-level learning (SenClass), to determine occurrence of infectious diseases as a binary classification task. In particular, DocClass takes an entire document as an input while SenClass takes a given sentence as input to the model. Their data used online documents from various websites cited as being important sources of public health information, such as NCDC or WHO-Disease Outbreak News (WHO-DON). Both models with either techniques achieved favorable results with accuracies higher than the baseline accuracy of 73%. SenClass was found to have better performance because of more training data, and BiLSTM had a higher accuracy with 92.9% than ConvNet (89.8%). The BiLSTM-SenClass also had the highest precision and recall, having a score of 92.9% for both.

Regarding social media posts, [7] developed a multi-label classifier for incident-related English tweets. Geo-tagged tweets were datamined from November 2012 until February 2013, and only two (2) cities were included. They identify four (4) possible labels under which tweets can be classified: (1) Fire; (2) Shooting; (3) Crash; and (4) Injury. They decomposed the multi-label problem into an ensemble of binary classification problems using three (3) techniques: (1) Binary Relevance; (2) Label Powerset; and (3) Classifier Chains. Results indicate that Classifier Chains show the best precision and recall for individual labels, with average scores of 93% and 76%, respectively. An SVM implementing Classifier Chains also achieved the best observed scores with an exact match of 84.35%.

Moving to Philippine data and also the basis of this paper, [3] constructed a dengue infoveillance agent in the Philippines, and their multi-label classification model is a Gated Recurrent Neural Network that obtained semantic understanding of tweets via pre-trained embeddings using GloVe [12]. Using hamming loss as an evaluation metric, they observed 4.9053% hamming loss, showing good performance in being able to classify tweets with high confidence.

III. Dataset

The dataset contains 5015 annotated tweets. The tweets are either in English, Filipino, or a combination of both. A tweet has five (5) possible labels, and any tweet can be classified in any of them, or none at all. Pertinent columns include the tweet and the aforementioned labels. The labels are described below.

1) Absent tweets are about not being able to attend an event or to a responsibility .
2) Dengue tweets discuss being sick from dengue fever.
3) Sick tweets discuss being sick in general, which may or may not be from dengue.
4) Health tweets revolve around any health- or medically-related topic.
5) Mosquito tweets involve any mention of mosquitoes.

It should be noted that Dengue tweets are placed under Sick tweets, which falls under Health tweets. This semi-hierarchy among the labels imply some correlation among the three (3) labels, and a sample generated heatmap for the training and validation set during exploratory data analysis in Figure 1 confirms this, although only to some degree.

It is observed here that the Health-Sick pair have the strongest positive correlation (0.65), followed by weak correlations for the Dengue-Sick (0.19), Dengue-Health (0.12) pairs. On the one hand, the strongest negative correlation is with Health-Absent (-0.43).

Further analysis also shows imbalance among the labels. Figure 2 shows this with Health predictably showing the most representation with roughly 40.62% of all tweets being tagged as Health while Dengue gets the least representation with roughly only 1.26% of tweets being tagged as such.

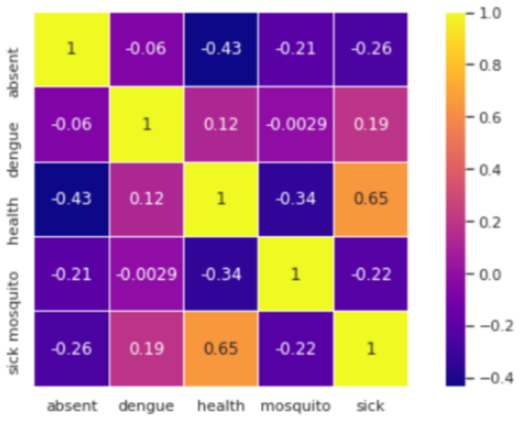

Fig. 1. Correlation of labels in the training and validation set.

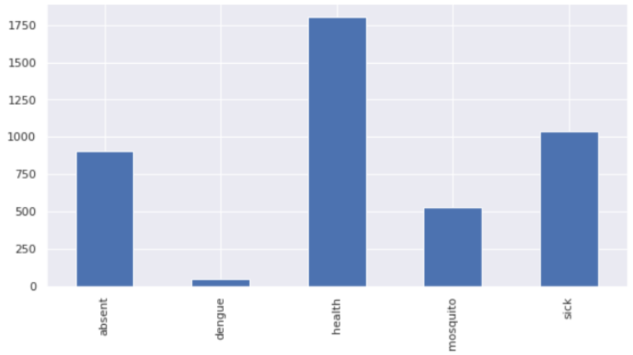

Fig. 2. Number of tweets per label in the train and validation set.

To handle this class imbalance, this paper will use text augmentation, which will be discussed further in the next section.

IV. Text Classifier

As mentioned earlier, there are five (5) labels under which a tweet can be labelled under, thus classifying it as a multi-label problem. We observe model performance through its accuracy and loss, then compare results when performing text augmentation.

A. Data Representation

We first perform text pre-processing steps to clean the tweets. The following steps are taken:

1) Remove hashtags, terms prefixed by '#', as well as user mentions, which are words pre-pended by an '@' symbol.
2) Remove web links.
3) Remove special characters.
4) Limit (or pad) to 30 word length.

Moreover, stop words were also not removed. This is to standardize the input and remove noise or those that contain no semantic meaning in the textual data.

The cleaned tweets set is further pre-processed similar to BERT's pre-processing treatment. For that, we use the HuggingFace transformers[1] library, which is available on GitHub.

B. Text Augmentation

To address class imbalance, specifically with the Dengue label having little to no representation, we perform text augmentation to generate synthetic data. To accomplish that, we perform insertion or substitution using contextualized word embeddings via a BERT multilingual model. This ensures that there is more data to learn from. It is important to note that text augmentation be performed strictly on the training data to lessen chances of overfitting. Considering that tweets can be a mix of both Filipino (or predominantly Tagalog) and English, the pre-trained $BERT_{\text{BASE}}$ multilingual model (cased) is used.

An example is shown below of how insertion and substitution will work:

- Original: the quick brown fox jumps over the lazy dog
- **Insertion**: the *lazy* quick brown fox *always* jumps over the lazy dog
- **Substitution**: the quick *thinking* fox jumps over the lazy dog

Insertion and substitution is randomly applied on the text to-be-augmented. They are not applied sequentially. Text augmentation is implemented prior to text cleaning via the NLPAug python library.[2]

C. Model

The widespread success of the Transformer architecture can be attributed to its attention mechanism that handles input-output dependencies and removes the need of sequential dependency on prior words, enabling it to get richer information from text in its entire surrounding. BERT [8] is a state-of-the-art implementation of the encoder stack of a transformer model. Building on top of the transformer model, the Bidirectional Encoder Representations from Transformers (BERT) is an unsupervised bidirectional encoder model, capable of reading in both directions at once.

BERT is optimized by minimizing loss through two (2) tasks it does. The first is the masked language model (or Cloze task), where it randomly masks 15% of an input into a [MASK] token prior to feeding it to the transformer, and trains itself to predict said masked tokens from the context of the non-masked words. At other times, it swaps out a random word and performs prediction of the correct word. The second, for better handling of sentence relationships, tasks BERT to predict sentence B given a sentence A, distinguished by end-of-sentence tokens ([SEP]). The first input is prepended by a [CLS] token.

We hypothesize that BERT's bidirectional training can have a deeper sense of language context and flow. To avoid overfitting of data and in the interest of limited computational resources, we use the pre-trained $BERT_{\text{BASE}}$ model [9]. The pre-trained model reduces the need for substantial architectural modifications as well. It has 12 transformer blocks, 768 hidden layers, and 110 million parameters. It is comparable to a 12-layer encoder stack for a richer semantic understanding of the text.

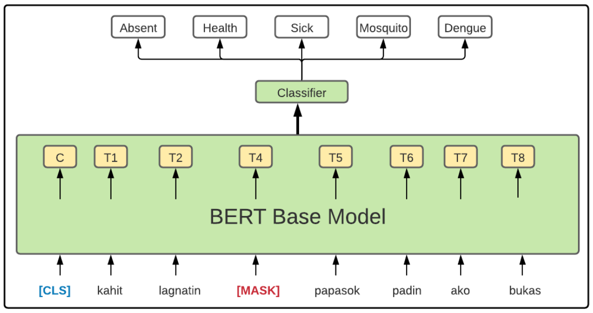

Fig. 3. Overview of the approach. We feed the pre-processed tweets set into the $BERT_{\text{BASE}}$ model where we set the max sentence length to 30. [CLS] is a token prepended at the start of every input. We use a 5-neuron single output layer to perform multi-label classification.

Figure 3 shows an overview of our approach. Regarding model building, we use the BERT encoder stack and its output is fed into 5-neuron output layer as our classifier. The training set will be fed to the network while the validation set is for evaluation during model training. The remaining 20% will be used as holdout data for experimentation and hyperparameter tuning.

V. Results and Discussion

We tested our BERT model without text augmentation, as well as with incrementally added synthetic data for the Dengue label. We explicitly set the other labels to be False for those synthetic data. This is to avoid 'bleeding' of information. Due to the small number of Dengue-label instances, we set the number of synthetic samples to be 10, 25, 50, 75, 100, and evaluate the loss and accuracy of each one against the no-augmentation baseline.

Our BERT classifier has the following hyperparameters:

1) RMSprop optimizer.
2) Learning rate of 1e-5.
3) ReLU as the activation function.
4) 10 epochs and a batch size of 64.

Model	**Loss**	**Accuracy**
BERT-NoAug	0.2021	95%
BERT-10Aug	0.2004	95%
BERT-25Aug	0.2012	95%
BERT-50Aug	0.1929	96%
**BERT-75Aug*	0.1807	96%
BERT-100Aug	0.1991	96%

TABLE I
Comparison of results of the different BERT models trained over 10 epochs.

Table I shows results of the incremental text augmentation when compared against our BERT baseline. It can be seen

[1] https://github.com/huggingface/transformers

[2] https://github.com/makcedward/nlpaug

here that there was only minimal improvement in loss by incrementally adding samples to the Dengue label. Bert-NoAug had the highest loss at 0.2021, while BERT-75Aug performed best with 96% accuracy and a low 0.1807 loss value. At Bert-100Aug, it can be seen that the loss value decreased, implying some upper bound that is being approached in which the model will perform slightly worse as more synthetic samples are added.

Our results overall show that a base BERT classifier trained can accurately classify English-Filipino tweets with relative confidence, although further parameter tuning needs to be done to optimize the loss value and to properly deploy a model like this for public health infoveillance. It should also be noted that the accuracy was capped at 96%, and that there is only minimal loss difference per augmentation, though this may more be attributed to the incremental (over)sampling performed.

VI. Conclusion

We presented a BERT model here in this paper that performs multi-label classification on tweets about dengue. We used a base multilingual model to account for English-Filipino tweets and also used contextual word embeddings using the same pre-trained BERT to perform text augmentation. Results from our experiments show that the baseline model has very high accuracy and relatively low loss, but that incremental oversampling of synthetic text data decreases loss. It was observed, however, that only minimal improvements were found when applying text augmentation, though it may more be attributed to the incremental oversampling. As more increments were added, however, the model started performing worse, which can be explained by the generated synthetic samples not reflecting the 'real' Dengue-label tweets or the model overfitting to the training data.

For future work, more in-depth analysis on the BERT model's performance on tweets as well as other sampling techniques to address class imbalance can be undertaken to optimize the loss value for proper deployment. The possibility of training a BERT model purely on Filipino text (or tweets) should also be explored so it could better understand the language. Lastly, the applications of BERT in other viral diseases should be explored, such as with COVID-19 infoveillance.

References

[1] World Health Organization (WHO), *"Dengue and severe dengue"*, Jun. 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[2] S. Kemp, *"Digital 2021: The Philippines"*, Feb. 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-philippines

[3] E. D. Livelo and C. Cheng, "Intelligent Dengue Infoveillance Using Gated Recurrent Neural Learning and Cross-Label Frequencies," In *2018 IEEE International Conference on Agents (ICA)*, 2018, pp. 2-7, doi: 10.1109/AGENTS.2018.8459963.

[4] J. Feldman *et al.*, "Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise," *Journal of the American Medical Informatics Association: JAMIA*, vol. 26, pp. 1355-1359, Nov. 2019. [Online]. Available: doi:10.1093/jamia/ocz112

[5] M. Kim, K. Chae, S. Lee, H. Jang and S. Kim, "Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches," *International Journal of Environmental Research and Public Health*, vol. 17, Dec. 2020. [Online]. Available: doi:10.3390/ijerph17249467

[6] S. E. Jordan *et al.*, "Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response," *Data*, vol. 4, Dec. 2018. [Online]. Available: doi:10.3390/data4010006

[7] A. Schulz, E. L. Mencía, T. T. Dang and B. Schmidt, "Evaluating Multi-Label Classification of Incident-related Tweets," In *Proceedings of the Making Sense of Microposts (#Microposts)*, 2014, pp. 26-33.

[8] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], May 2019.

[9] I. Turc, M. Chang, K. Lee and K. Toutanova "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models," arXiv:1908.08962 [cs], Aug. 2019.

[10] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[11] Bird, S., Loper, E., & Klein, E., *Natural Language Processing with Python.* O'Reilly Media Inc.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP).* pp. 1532–1543. http://www.aclweb.org/anthology/D14-1162