Deep Learning Based Framework for Classification of Water Quality in Social Media Data

Muhammad Hanif, Ammar Khawer, Muhammad Atif Tahir, Muhammad Rafi
National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
{hanif.soomro,k201414,atif.tahir,muhammad.rafi}@nu.edu.pk

ABSTRACT

This paper describes the method proposed by team FAST-NU-DS, for the task of WaterMM: Water Quality in Social Multimedia at MediaEval, 2021. The task aims to analyze water security, safety, and quality of water and build a classifier that differentiates whether the tweet is discussing water quality issues. The task includes a dataset in the form of tweets containing the tweet's text, their metadata, and a few tweets also contain images. The proposed method has performed pre-processing steps on the text and tags of the dataset and applied Bidirectional Encoder Representations from Transformers (BERT). The proposed method has applied Visual Geometry Group (VGG16) pre-trained on the ImageNet dataset for the binary classification of images. The proposed method has achieved a 0.31 F1 score for text-only content. Moreover, the combination of text and images provided a 0.24 F1 score.

1 INTRODUCTION

The enormous amount of data generated by social media is being investigated for the solution of various problems. Various social media platforms, including Twitter, allow users to share text and image content, which can be used for situational awareness at any time. The task of "WaterMM: Water Quality in Social Multimedia" at MediaEval, 2021 [1], has focused on examining water safety, quality and security by using social media data. The task is aimed to assist with the complaints regarding the quality and conditions of drinking water through social media data, which will help the water utility and protection agencies to better serve the communities at large.

2 LITERATURE REVIEW

The research effort utilized BERT and various competitors for the representation of disaster-related tweets [4]. The method has experimentally proved that the BERT has surpassed various embedding methods, including Glove [8] and FastText [3]. Another research effort has taken Two different datasets in English and Italian languages and applied BERT. The research has focused on avoidance of noise and managing various web-related noisy objects, including emoticons, emojis, mentions, hashtags, and so on [9]. Researchers [11] have performed multi-label classification on disaster-based tweets. The method has produced state-of-the-art results on the dataset by using two variants of BERT. Another research framework [7] has been proposed to investigate the flooding situation. The framework collects real-time images and text based data and shows its relevancy or irrelevancy with flooding disasters. The framework classifies tweets based on their text and checks if the tweet contains an image, then image features are also considered to make a strong prediction.

MediaEval'21, December 13-15 2021, Online

3 PROPOSED APPROACH

The proposed method for WaterMM: Water Quality in Social Multimedia at MediaEval, 2021 [1], has utilized a bilingual text-based dataset, which contains tweets in either Italian language or English language. At the next stage, images are added along with text to perform binary classification based on either the tweet is discussing water quality-related issues or not.

3.1 Approach for text data

For the first sub-task, only text contents are utilized to binary tweets and predict whether the tweet discusses water quality. For the processing of text extracted from tweets, the description of tweets and tags are considered for the binary classification task. As the dataset for "WaterMM: Water Quality in Social Multimedia" at MediaEval, 2021 [1] contains tweets in the English language and in the Italian language. Therefore, googletrans [6] library is utilized for the translation of each tweet from the Italian language to the English language. So that, all the data is available in one language (English).

After translation of train and test data, various pre-processing steps are performed to clean text contents. Initially, the Uniform Resource Locator (URLs) are removed from the description of tweets. Moreover, hash symbols and punctuations from tweets are also removed from each tweet of training and testing sets. The pre-processing step also removed smileys and emoticons from the text of tweets and contents converted into the lowercase. Finally, numbers and symbols are eliminated from the data to make the contents more meaningful. The description part of the tweet also contains stop words, which have less importance for the binary classification task. So, stop words are also removed from the tweets.

It has been observed that the dataset of WaterMM: Water Quality in Social Multimedia at MediaEval, 2021 [1] is highly imbalanced. The minority class of the dataset contains 1140 tweets, which shows discussion related to water quality. However, the majority class has 4248 tweets, which are not discussing the quality of water. Therefore, the minority of majority class ratio is almost 1:4. Oversampling technique has been used to reduce the class imbalance. The minority class is oversampled three times to decrease the imbalance between classes.

Later, the Bidirectional Encoder Representations from Transformers (BERT) is trained by using a train-set of the dataset. Each instance of the training set is created by combining text and tags of the tweet, which are then converted into tokens. The 'bert-base-uncased' model is selected for the processing, which lowercased the

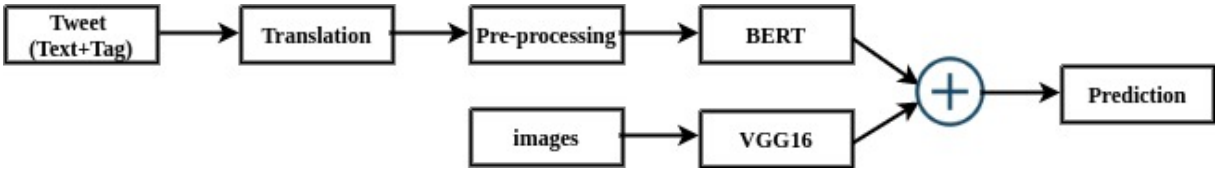

Figure 1: Data flow diagram showing processing of visual and text data

contents and then converted them into tokens. For further processing, the [CLS] and [SEP] keywords are added to separate the dataset instances. The maximum length for a single text-based instance of the dataset is set as 256 tokens. The training set is divided into train and valid sets by allocating 10% of training data to the validation set, and the remaining 90% is allocated for training. Furthermore, the AdamW optimizer is used along with the learning rate set as 2e-5, and the epsilon is set as 1e-8. The model has been trained for three epochs. Finally, The trained model is then used to predict 1920 test set instances. The prediction in the form of 0 or 1 is collected and stored in a comma-separated format for test-set.

3.2 Approach for text and image data

The second sub-task has utilized text as well as images available for the tweets. Though very few tweets contain images, only 954 tweets from the train-set contain images, and 245 tweets from the test-set contain images. Due to the insufficient quantity of images, the oversampling has been performed for both minority and majority classes. The class of images, which represent the availability of water quality, has only 264 images. The quantity of minority class is oversampled by creating five different augmented samples of each image. For augmentation, python-based library "Augmentor" [2] is utilized. The random samples are created by varying different parameters, including rotate, zoop, and flip. Similarly, the majority class is added with two additional augmented copies of images for each of its instances.

The increased quantity of images is utilized for the classification by applying Visual Geometry Group (VGG16) model [10], pre-trained on ImageNet [5] dataset. The model is fine-tuned by retraining the last four layers of the model. The rest of the model is frozen to keep previous learning on the ImageNet dataset. The learning rate is set as 10^{-5} and the dropout value as 0.3. The sigmoid function has been used, and the problem is related to binary classification. The quantity of 20% instances are allocated f validation set, and the rest of the 80% is utilized for training. The model is retrained for 25 epochs, and the trained model is saved for prediction on test-set images. The model predicts test set images, and the confidence score for each of the images is stored.

On the other hand, the prediction confidence for text instances is retrieved by using BERT, and predictions are normalized between 0 and 1. Later, the prediction achieved by applying VGG16 based model is combined with BERT-based predictions, and the average is calculated. Then, the sigmoid function is applied for the final prediction. The approach is depicted in figure 1.

Table 1: Results of Proposed method on Test Data

Run#	Method	F1-Score
Run 1	BERT	31.67%
Run 3	BERT+VGG16	24.45%

4 RESULTS AND ANALYSIS

The proposed method has achieved a 31.67% F1-score for the first run. The first run has utilized descriptions and tags of the tweets for its prediction. However, the second run has achieved 24.45% F1-score. The second run has utilized descriptions and tags of tweets and images for a limited number of instances. The results for textual and combination of both visual and textual are summarized in table 1. It has been observed that the test set's score is less compared to train and validation sets. The reason for less effective results may involve a very similar type of tweets in both classes. The tweet text contains various similar words in both classes, which might have confused the algorithms, such as water and bottles. The quantity of tweets containing images is less than sufficient for deep learning models, due to which Run two has produced a low evaluation score compared to Run 1. Moreover, observations revealed that few of the images declared as a part of the class showing water quality but do not visualize anything related to water. On the other hand, it has also been observed that images in the negative class, which does not discuss water quality, also include water-related contents as water bottles. So, this has confused deep learning models to discriminate between both classes. Results may be improved using multiple deep learning based models for image classification. Text-based classification method can also be improved by increasing minority class, where instead of simple over-sampling, synonyms may be used to increase instances.

5 DISCUSSION AND OUTLOOK

The research has proposed a Bidirectional Encoder Representations from Transformers (BERT) approach for finding water-quality related tweets. The method has also utilized Visual Geometry Group (VGG16), pre-trained on ImageNet dataset to binary classify the images based on whether they contain evidence of water quality. Research can be enhanced by using the Places dataset, which describes scene-based information. Furthermore, advanced over-sampling techniques may be used as translation-based or synonym-based oversampling.

ACKNOWLEDGMENTS

This research work was funded by Higher Education Commission (HEC) Pakistan and Ministry of Planning Development and Reforms under the National Center in Big Data and Cloud Computing.

REFERENCES

[1] Stelios Andreadis, Ilias Gialampoukidis, Aristeidis Bozas, Anastasia Moumtzidou, Roberto Fiorin, Francesca Lombardo, Anastasios Karakostas, Daniele Norbiato, Stefanos Vrochidis, Michele Ferri, and Ioannis Kompatsiaris. 2021. WaterMM:Water Quality in Social Multimedia Task at MediaEval 2021. In *Proceedings of the MediaEval 2021 Workshop, Online*.

[2] Marcus D Bloice, Christof Stocker, and Andreas Holzinger. 2017. Augmentor: an image augmentation library for machine learning. *arXiv preprint arXiv:1708.04680* (2017).

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[4] Ashis Kumar Chanda. 2021. Efficacy of BERT embeddings on predicting disaster from Twitter data. *arXiv preprint arXiv:2108.10698* (2021).

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[6] Suhun Han. 2020. googletrans 3.0.0. https://pypi.org/project/googletrans/. (2020). Accessed: 2020-11-1.

[7] Anastasia Moumtzidou, Stelios Andreadis, Ilias Gialampoukidis, Anastasios Karakostas, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2018. Flood relevance estimation from visual and textual content in social media streams. In *Companion Proceedings of the The Web Conference 2018*. 1621–1627.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[9] Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. 2021. Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications* 181 (2021), 115119.

[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] Hamada M Zahera, Ibrahim A Elgendy, Rricha Jalota, Mohamed Ahmed Sherif, EM Voorhees, and A Ellis. 2019. Fine-tuned BERT Model for Multi-Label Tweets Classification.. In *TREC*. 1–7.