Identification of Informative Tweet during Disasters

Prof. R.B. Murumkar[1], Vinay M. Harwani[2], Namita Bhalerao[2], Nilambari K. Rathi[2], Rutuja D. Mahajan[2]

[1]Professor, Dept. of Information Technology Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India
[2]Student, Dept. of Information Technology Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India

---***---

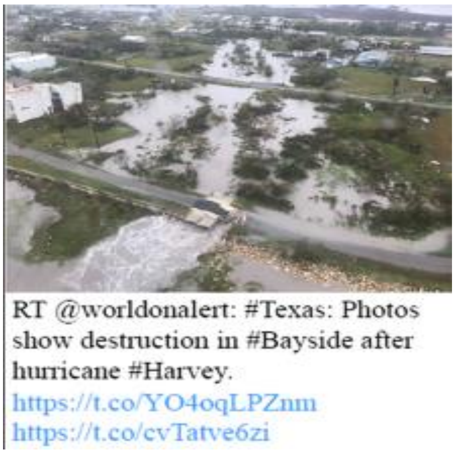
***Abstract** - During the time of crisis, people often post countless instructive and informative tweets on various social media platforms such as Twitter. Recognizing informative tweets is a difficult errand during the fiasco from such a vast pool of tweets. These tweets can be textual as well as photographic in nature. As an answer to this issue of sorting out enlightening tweets, we present a way to perceive the distinguishing calamity related informative tweets from the Twitter streams utilizing the textual content and the images together. Our objective is to construct a model by combining Bi-directional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN) to categorize the textual content for the tweets. The primarily image-based classification model will utilize the adjusted VGG-16 design to extricate the features from the picture and will classify the image accordingly. The output of the text-based model and the image-based model will be consolidated using the late fusion technique to predict the tweet label.*

Key Words: Bi-LSTM, CNN, VGG-16, Tweets, Disaster, Informative, Image Classification, Twitter.

1. INTRODUCTION

A natural calamity makes critical environmental disruptions requiring broad endeavors from society to survive and cope with them. In such cases of disasters, a considerable amount of data is circulated and posted on various micro-blogging platforms like Twitter. Therefore the need of the situation is to implement certain tools that will help in analyzing this huge amount of tweets and will classify them among the informative and non-informative categories for the rapid early damage assessment during disasters. The public authority or humanitarian association needs data on the circumstance updates of the number of individuals have been wounded and died, the number of structures that are imploded, and so forth, to make the quick move and take immediate action. The factual analysis and evaluation of these crisis related data can help humanitarian organizations and rescue squads in efficient decision-making and dynamic prioritization of tasks. Similarly, affected people need data like where clinical assets are accessible, food and shelter assets,etc. to save their lives. Tweets of this kind are grouped under informative tweets. Among the huge volume of tweets identified with a certain calamity, some of them may be simply expressing gratitude towards Twitter or neighborhood bunches for their assistance. Also some people post tweets which are not directly related to the disaster, for example, feelings, sentiments, emotions etc. These tweets are not helpful in the rescue work. Hence these kinds of tweets are named as non-informative tweets. Institutional and Volunteer rescue efforts save a lot of people during a catastrophe. However, Individual volunteers have time limitations and lack of assets. Also it is not possible for the people from the rescue squad to reach out to every tweet on their own because of the huge volume and fast pace at which tweets are getting posted. This generates an urgent need to construct certain frameworks that can consequently segregate useful information through an enormous volume of social media content. The auto-programmed classification of messages, particularly tweet messages, is a difficult task because of their restriction in character length (280 characters), non-standard shortened forms, and linguistic blunders.

Instance of Informative tweets with images during crisis:



RT @worldonalert: #Texas: Photos show destruction in #Bayside after hurricane #Harvey. https://t.co/YO4oqLPZnm https://t.co/cvTatve6zi

In this paper, we propose a method by highlighting both pictorial as well as textual nature of the tweets. The proposed strategy classifies the images using VGG-16 model consisting of CNN layers and classifies textual content using Bi-LSTM and CNN model. Their combination with the help of late fusion technique will help in generating and predicting the tweet label as informative or non-informative tweet.

2. RELATED WORK

As of late, a few works have been accounted for effectively using the crisis related data from various social networking platforms like Twitter. Finding useful data from the enormous online media information is one of the crucial tasks for helping social associations.

Abhinav Kumar et al[1] proposed a method to classify informative tweets consisting of images and text together. The framework comprises of two equal profound neural designs: LSTM network for text and VGG-16 organization for handling pictures. The 256-dimensional vector resulting from VGG-16 is then concatenated with 256-dimensional feature vector from tweet textual data which is further used to classify the tweets into instructive and non-instructive. The system achieved an average F1 score of 0.82 for the datasets of Hurricane Harvey,Mexico earthquake, SriLanka floods.

Sreenivasulu Madichetty et al[2] proposed a novel approach to classify tweets adapting image features. It used CNN-ANN dual approach to classify text. CNN is used for feature extraction while ANN is used as classifier which achieves better efficiency than SVM classifier. The images are classified using fine tuned VGG-16 model. The VGG-16 architecture consists of 16 CNN layers having tunable parameters while other max pooling layers with no tunable parameters. Around 74% accuracy was achieved and it was noted that this system delivers greater efficiency than text based classification.

Md.Yasin Kabir et al[3] proposed the deep learning approach combining Bi-LSTM and CNN to categorize the tweets. It consists of 7 modules: Input layer, Embedding layer, BLSTM layer, Attention layer, Auxiliary features input, Convolution layer, and Output layer. The system additionally extracted the location from the tweets to assist in rescue operation. The method outperformed many other classification methods based on Precision,F1 Score, Recall. They also developed an adaptive algorithm to perform efficient rescue operation scheduling according to priorities.

Gregoire Burel et al[4] proposed a semantic approach for tweet classification using Dual-CNN technique. A semantic embedding layer is added to the traditional CNN layers to better capture the context of the tweets. The preprocessed tweets are given as input to word vector initialization as well as concept vector initialization. The semantic extraction used AlchemyAPI to extract named entities and then mapped them with subtypes with multiple bases like DBpedia and Freebase. Accuracy greater than 79% is achieved but it drops to 61% while identifying exquisite event related information.

Alapan Kuila et al[5] proposed a method to extract events from newswires or social media content for Indian Languages like Tamil,Hindi,English.Event identification is employed using Bi-LSTM and CNN together.Event trigger and Event Arguments are identified using the technique and these are linked correspondingly using heuristic based approach where each argument word is matched with the nearest trigger word depending upon distance that is calculated by number of sentences between them. The system achieves an F-score of 39.71,37.42,39.91 for Tamil, Hindi and English datasets which is relatively low compared to other models.

Shuichi Hasida et al[6] proposed an multi-channel representation along with CNN for tweet classification.The proposed model consists of six main structures, an input layer, embedding layer, convolution layer, pooling layer,fully-connected layer, and output layer. The context of words in the vectors can be considered using multi-channel distribution representation. In this representation, multi-channels consisting of multiple values for each element in the vector are represented. Each word has different vectors depending upon the context of the word in tweets. Accuracy of 77% is achieved via this approach.

Agung Triayudi et al[7] proposed CNN model to categorize tweets related to emergency response phase. The preprocessed input of tweets is given to feature extraction phase to form word vectors which is further given to convolutional layer and pooling layer to extract features and reduce dimension. The output of CNN layers is given to the flattening layers. The results of these layers are concatenated into a single vector. This vector will be processed at Fully Connected Layer using softmax function to get probability for each class. The model achieved 98% accuracy for small amounts of data. However the accuracy drops for subsequent amounts of data.

Alan Aipe et al[8] proposed a multi-label classification CNN model with respect to 7 class taxonomy and contains 7 similar classifiers predicting the binary output specifying whether the label is relevant or not.88% accuracy is achieved by the deep CNN architecture.

3. PROPOSED METHODOLOGY

The proposed model will take input as tweets containing photographical as well as textual content and the output will be its classification into informative or non-informative categories. The model will consist of three modules, i.e.

1. Text-based classification technique

2. Picture based classification technique and

3. Late Fusion technique for the combination of the text and image information.

3.1 Text Classification:

Text based classification for tweets will be achieved using Bi-LSTM and CNN technique. The 4 major components of the technique would be:

3.1.1. Pre-processing: Tweets are typically composed of incomplete, noisy and unstructured sentences because of the common presence of short forms, poorly phrased sentences and some informal dictionary terms. This stage will accordingly apply a progression of preprocessing steps to lessen the noise content in tweets .

3.1.2. Embedding layer: This layer will encode the input given into real-valued vectors with the help of lookup tables. Pre-trained word vectors named Crisis and GloVe will generate feature word vectors using co-occurrence based statistical models. Thus embedding applied to the words will be used to map all tokenized words in every tweet to their particular word vector tables.

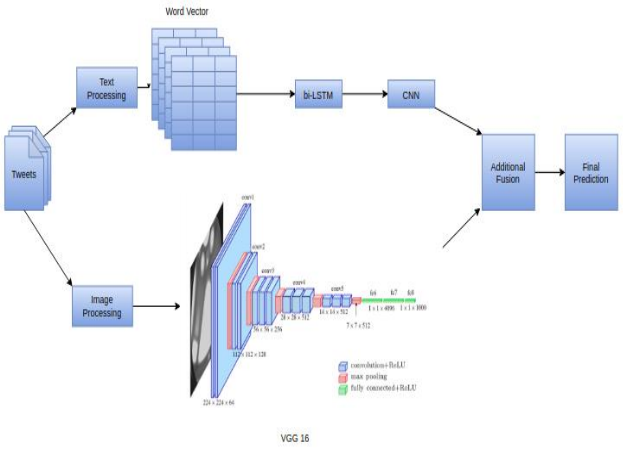
3.1.3 Bi-LSTM layer: The Long-Short Term Memory (LSM) is a specific rendition of Recurrent Neural Network (RNN) that is fit for learning long haul conditions. Although LSTM can only learn from past given inputs, Bi-LSTM has a salient feature of running the input in both forward as well as backward direction and will be critical for our application in understanding complex language.

3.1.4. Convolutional layer: The convolution layer will be performing a matrix vector operation onto the sentence level representation sequence provided as an input. The word vectors will go through all the convolution layers where all the input information will be consolidated together to produce respective feature maps.To deal with the non-linearity in the convolution layer, the Rectified Linear Unit (ReLU) will be used as an activation and will generate a rectified feature map..

3.2 Image Classification:

For the image based classification, Convolutional Neural Network (CNN) based pretrained VGG-16 model is utilized to separate features from the given image provided as an input to the model. All the weights of the VGG-16 community will be marked as non-trainable besides the weight present among the ultimate dense layer and the output layer. For the multi-modal setting, the feature vector that will come from the penultimate dense layer of VGG-16 shall be passed via other dense layers containing 256 neurons. All input pictures will be resized to '224 × 224' for the processing in accordance with VGG-16 architecture.

VGG-16 has 16 layers with tunable attributes, while different layers (max-pooling layer) have no tunable attributes. Each convolution layer has no maximum pooling layer. Each convolutional layer has different types of channels with various sizes. Hence different layers of VGG-16 will extract different features. Like lower layers will extract edges, corners, face contribution, wheels, cars and various general features. While higher layers will extract domain specific features like damaged buildings, injured people,etc.



Proposed Methodology for Tweet Classification

The resultant vectors of the image based classification model and text based classification are merged by adding them. The segregation of tweets into informative and non-informative will be done using these vector classification into class labels.

4. CONCLUSION

Through this research, we got acquainted with different perspectives to classify the tweets on the basis of text as well as images it contains during disastrous situations into informative and non-informative categories.Various novel approaches that we got over are CNN-ANN, LSTM, Dual-CNN and many more. After evaluation of methods, we concluded that CNN approach has better efficiency compared to other approaches for text classification. Pre-trained VGG-16 model gives more accurate results for image classification. This research paper proposes a model that specifies Bi-LSTM and CNN combined approach for text based classification while VGG-16 architecture for image classification. Thus the approach will prove beneficial to segregate the important information from the tweets during crisis conditions.In future, the scope can be extended for multi-linguality that is classification based on different languages.

REFERENCES

[1] Kumar, Abhinav & Singh, Jyoti & Dwivedi, Yogesh & Rana, Nripendra. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. Annals of Operations Research. 10.1007/s10479-020-03514-x.

[2] Madichetty, Sreenivasulu & M, Sridevi. (2020). Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. Multimedia Tools and Applications. 79. 10.1007/s11042-020-09343-1.

[3] Kabir, Md Yasin and Sanjay Madria. “A Deep Learning Approach for Tweet Classification and Rescue Scheduling for Effective Disaster Management.” Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances inGeographic Information Systems (2019): n.pag.

[4] Burel, G., Saif, H., Fernández, M., & Alani, H. (2017). On Semantics and Deep Learning for Event Detection in Crisis Situations.

[5] Kuila, A., Bussa, S.C., & Sarkar, S. (2018). A Neural Network based Event Extraction System for Indian Languages. *FIRE*.

[6] Hashida, S., Tamura, K., & Sakai, T. (2018). Classifying Tweets using Convolutional Neural Networks with Multi-Channel Distributed Representation.

[7] Journal of Software Engineering, J., Systems, I., & agung triayudi. (2019). CONVOLUTIONAL NEURAL NETWORK FOR TEXT CLASSIFICATION ON TWITTER. Journal of Software Engineering & Intelligent SYstems, 4(3), 123–131.

[8] Aipe, A., Ekbal, A., Sundararaman, M.N., & Kurohashi, S. (2018). Deep Learning Approach towards Multi-label Classification of Crisis Related Tweets. *ISCRAM*.

[9] X. She and D. Zhang, "Text Classification Based on Hybrid CNN-LSTM Hybrid Model," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 185-189, doi: 10.1109/ISCID.2018.10144.

[10] J. Zhang, Y. Li, J. Tian and T. Li, "LSTM-CNN Hybrid Model for Text Classification," 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, 2018, pp. 1675-1680, doi: 10.1109/IAEAC.2018.8577620.

[11] Singh, Jyoti & Dwivedi, Yogesh & Rana, Nripendra & Kumar, Abhinav & Kapoor, Kawaljeet. (2019). Event Classification and Location Prediction from Tweets during Disasters.. Annals of Operations Research. 283. 10.1007/s10479-017-2522-3.

[12] Nguyen, Dat & Mannai, Kamela & Joty, Shafiq & Sajjad, Hassan & Imran, Muhammad & Mitra, Prasenjit. (2016). Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks.

[13] Nguyen, Dat & Joty, Shafiq & Imran, Muhammad & Sajjad, Hassan & Mitra, Prasenjit. (2016). Applications of Online Deep Learning for Crisis Response Using Social Media Information.

[14] S. Madichetty and M. Sridevi, "Detecting Informative Tweets during Disaster using Deep Neural Networks," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 2019, pp.709-713,doi:10.1109/COMSNETS.2019.8711095.

[15] Jain, Pallavi & Ross, Robert & Schoen-Phelan, Bianca. (2019). Estimating Distributed Representation Performance in Disaster-Related Social Media Classification. 10.1145/3341161.3343680.

[16] Zhou, Chunting & Sun, Chonglin & Liu, Zhiyuan & Lau, Francis. (2015). A C-LSTM Neural Network for Text Classification.

[17] Jun Li, Guimin Huang, Jianheng Chen, Yabing Wang, and Carmen De Maio. 2019. Dual CNN for Relation Extraction with Knowledge-Based Attention and Word Embeddings. Intell. Neuroscience 2019 (2019). DOI:https://doi.org/10.1155/2019/6789520