Flood Detection via Twitter Streams Using Textual and Visual Features

Firoj Alam[1], Zohaib Hassan [2], Kashif Ahmad [3], Asma Gul [4], Michael Alexander Riegler [5], Nicola Conci [2], Ala Al-Fuqaha [3]

[1]Qatar Computing Research Institute, Doha, Qatar, [2] University of Trento, Italy

[3] Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar,[4] Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan,[5] SimulaMet, Norway

ABSTRACT

The paper presents our proposed solutions for the MediaEval 2020 Flood-Related Multimedia Task, which aims to analyze and detect flooding events in multimedia content shared over Twitter. In total, we proposed four different solutions including a multi-modal solution combining textual and visual information for the mandatory run, and three single modal image and text-based solutions as optional runs. In the multi-modal method, we rely on a supervised multimodal bitransformer model that combines textual and visual features in an early fusion, achieving a micro F1-score of .859 on the development data set. For the text-based flood events detection, we use a transformer network (i.e., pretrained Italian BERT model) achieving an F1-score of .853. For image-based solutions, we employed multiple deep models, pre-trained on both, the ImageNet and Places data sets, individually and combined in an early fusion achieving F1-scores of .816 and .805 on the development set, respectively.

1 INTRODUCTION

Floods are the most frequent and devastating type of natural disaster causing a significant loss in terms of human lives and infrastructure worldwide, every year. According to a recent report[1], around 80-90% of natural disasters worldwide over the last decade are caused by floods, and more than two billion people worldwide were affected between 1998-2017. The damage of flood can be significantly mitigated if timely and accurate information about the location, scale, and most affected areas is available [3, 22]. However, several challenges, such as the availability of reporters and other resources, etc., are associated with the gathering of such information during floods [22]. On the other hand, social media has been proved very effective in information dissemination in such events [4, 6, 8, 21, 22].

Similar to previous years of the challenge [9–11], the MediaEval 2020 flood-related multimedia task [7] aims to analyze tweets from Twitter for flood events detection. The participants were provided a collection of tweets with associated images, and were asked to propose a framework able to automatically identify flood-related tweets relevant to a particular area. This paper provides the details of the methods proposed by team HBKU_UNITN_SIMULA for the task. In total, we proposed four different solutions including a multimodal one for the mandatory run, a textual information based solution, and a couple of image-based solutions for flood events detection in Twitter images.

2 PROPOSED APPROACHES

Transfer learning has become mainstream in computer vision and natural language processing (NLP). For example, in computer vision models (e.g., VGG16, ResNet18) trained using ImageNet [18] or Places Database [25] have been used as pre-trained models to initialize networks for fine-tunning the task-specific models. For NLP, word-embedding [20], sentence-embedding [13], and recent BERT [15] based models have shown significant progresses in downstream tasks. For this study, we used deep CNN models for image classification and a transformer model for text classification and finally combine them to design a multimodel network. Prior work with similar approaches in this direction include [1, 2, 17, 21]. The study in [17] used a combination of a deep CNN for image and a transformer model for text in a similar manner as proposed in this work. For the disaster response task, in [21] the authors used a deep CNN (VGG16) for the image, a CNN with static embedding for text, and finally combine them in the shared representation before a softmax layer for classification. In [1], the authors propose a cross-attention module for multimodal fusion, and the study in [2] proposes different fusion approaches for combining disaster-related tweet classification tasks. Our work is in line with the study by [17], however, our work is different in how we use different pre-trained models (i.e., models trained with ImageNet and Places database). In the next section we discuss the details of the models used in this study.

2.1 Text-based Model (Run 2)

Pre-processing. For the text-based model, we first pre-process the tweet texts as they are noisy, and consist of many symbols, emoticons, URLs, usernames, and invisible characters. Prior studies like [21] show that filtering and cleaning the tweets before training a classifier helps significantly. We pre-process the tweet texts before the classification experiments. The preprocessing includes removal of invisible characters, URLs, and hashtag signs.

Transformer model. The pre-processed texts are then fed into a transformer network by adding a task-specific layer on top of the network. We use a model-specific tokenizer, which is a part of the transformer model. Currently, the pre-trained transformer models are available for monolingual and multilingual settings. Since the

[1] https://www.who.int/health-topics/floods#tab=tab_1

MediaEval'20, December 14-15 2020, Online

tweets are in the Italian language, and for Italian a monolingual model exists, namely Italian BERT[2][3], we used it for our experiments.

For the training, we used the Transformer Toolkit [24]. We fine-tune the model using a learning rate of $1e-5$ for ten epochs [15]. The training of the pre-trained models has some instability as reported in [15], therefore, we run each experiment ten times using different seeds and select the model that performs the best on the development set. Finally, we evaluate the selected model on the test set.

2.2 Image-based Model (Run 3 and 4)

For the flood image detection we employed two different methods. In the first method, we fine-tuned an existing model, namely VggNet16 [23], pre-trained on the Places dataset [25]. In the second method, we jointly utilized the models pre-trained on the ImageNet [14] and the Places dataset. The basic motivation for the joint use of the models comes from our previous experience on similar tasks [4, 5], where the fusion of object and scene-level information extracted with the models pre-trained on ImageNet and Places dataset, respectively, have been proven very effective in classification of disaster-related images.

The class distribution of the dataset for the challenge this year is very imbalanced. Therefore, we used an oversampling technique to balance the distribution of the class labels in the training. This of course comes with the risk that the test data might be imbalanced which would most probably reduce the performance. We used the Synthetic Minority Oversampling Technique (SMOTE) [12] to up-sample the minority class. We used the *imblearn* implementation [19] for our experiments. In fact, the number of samples in the minority class have been increased by a factor of three.

2.3 Multimodal Model (Run 1)

The multimodal network consists of a text and an image network combined to form a shared representation before the classification layer. The text network consists of BERT [15] and the image network consists of ResNet152 [16]. We used ResNet152 in multimodal network as it was shown to work well in a previous study [17]. The input to the whole network is pre-processed text and extracted features for the images. The object and scene-level features are extracted through VGGNet16 pre-trained on ImageNet and Places datasets. During the training the model jointly learns the image embeddings and token embedding spaces of BERT. We use the Adam optimizer with a minibatch size of 32 for training the model.

3 RESULTS AND ANALYSIS

In total, we submitted four runs. Our first run is based on the multimodal framework where textual and visual features are combined using early fusion. Our second run is based on an Italian version of the BERT model for text analysis. The third run is based on VGGNet-19 pre-trained on Places datasets, which is fine-tuned on the flood-related images. In run four, two versions of VGGNet-16, one pre-trained on ImageNet and the other pre-trained on the Places dataset, are combined in an early fusion manner by concatenating the features obtained from the last fully connected layer.

Table 1: Experimental results of the proposed methods.

Run #	Type of Features	Micro F1-Score
1	Textual + Visual	0.859
2	Textual	0.853
3	Visual	0.816
4	Visual	0.805

Table 1 provides the experimental results of our proposed solutions for the task on the development set. Overall better results are obtained with run 1, which shows the advantage of a multi-modal solution over the single modality in the task. On the other hand, the lowest F1-score is obtained with fusion of object and scene-level features. However, it is interesting that the results obtained with the individual model (VGGNet16) pre-trained on the Places dataset has outperformed the method combining the object and scene-level features. In order to investigate the potential causes of the reduction in the performance of the fusion framework, we also analyzed the performance of VGGNet16 pre-trained on ImageNet where an F1-score of .804 is obtained. The lower performance of the model when pre-trained on ImageNet compared to the Places data indicates that the scene-level features are more important for the task. It is to be noted that the models pre-trained on ImageNet extract object-level while the ones pre-trained on the Places datasets correspond to scene-level information. Due to the imbalanced dataset we decided not to discuss the test data set results before the test data is publicly released and we are able to perform a more detailed analysis.

4 CONCLUSIONS AND FUTURE WORK

The task aims at the multimodal analysis of floods on Twitter. The participants were provided with a collection of tweets containing text and associated images and were asked to propose a multimodal framework able to automatically determine whether a tweet represents a flood-related event relevant to a specific area or not. We proposed four different solutions to the task including a multimodal, a textual, and a couple of image-based solutions. Overall, better results are observed for the multimodal approach indicating the advantage of the joint use of textual and visual information. As far as the evaluation of textual and visual information is concerned, significantly better results are obtained with textual features compared to visual information. Moreover, we also observed that scene-level information is more critical for the task compared to object-level features extracted with models pre-trained on ImageNet.

We believe there is still room for improvement in the multimodal, textual, and image-based solutions. In the future, we aim to explore the task further by introducing more sophisticated methods to jointly combine textual and visual information in a better way. We also plan to perform a more detailed analysis of the test data once publicly released.

REFERENCES

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal Categorization of Crisis Events in Social

[2] https://huggingface.co/dbmdz/bert-base-italian-uncased

[3] Note that the model is trained on Wikipedia dump and various texts from the OPUS corpora.

Media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14679–14689.

[2] Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (Apr. 2020), 346–353. https://doi.org/10.1609/aaai.v34i01.5369

[3] Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Nicola Conci, and Pål Halvorsen. 2018. Social media and satellites. *Multimedia Tools and Applications* (2018), 1–39.

[4] Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Olga Ostroukhova, Pål Halvorsen, Nicola Conci, and Rozenn Dahyot. 2019. Automatic detection of passable roads after floods in remote sensed and social media data. *Signal Processing: Image Communication* 74 (2019), 110–118.

[5] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. 2017. Convolutional neural networks for disaster images retrieval. In *Proceedings of the MediaEval 2017 Workshop (Sept. 13–15, 2017). Dublin, Ireland.*

[6] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2020. Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. *Behaviour & Information Technology* 39, 3 (2020), 288–318.

[7] Stelios Andreadis, Ilias Gialampoukidis, Anastasios Karakostas, Stefanos Vrochidis, Ioannis Kompatsiaris, Roberto Fiorin, Daniele Norbiato, and Michele Ferri. 2020. The Flood-related Multimedia Task at MediaEval 2020. (2020).

[8] Benjamin Bischke, Damian Borth, Christian Schulze, and Andreas Dengel. 2016. Contextual enrichment of remote-sensed events with social media streams. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 1077–1081.

[9] Benjamin Bischke, Patrick Helber, Erkan Basar, Simon Brugman, Zhengyu Zhao, and Konstantin Pogorelov. The Multimedia Satellite Task at MediaEval 2019: Flood Severity Estimation. In *Proc. of the MediaEval 2019 Workshop* (Oct. 27-29, 2019). Sophia Antipolis, France.

[10] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. 2017. The Multimedia Satellite Task at MediaEval 2017: Emergence Response for Flooding Events. In *Proceedings of the MediaEval 2017 Workshop (Sept. 13-15, 2017). Dublin, Ireland.*

[11] Benjamin Bischke, Patrick Helber, Zhengyu Zhao, Jens de Bruijn, and Damian Borth. The Multimedia Satellite Task at MediaEval 2018: Emergency Response for Flooding Events. In *Proc. of the MediaEval 2018 Workshop* (Oct. 29-31, 2018). Sophia-Antipolis, France.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[13] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[19] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[21] Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. In *ISCRAM*.

[22] Naina Said, Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. 2019. Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications* (17 Jul 2019). https://doi.org/10.1007/s11042-019-07942-1

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

[24] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[25] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.