WISC at MediaEval 2017: Multimedia Satellite Task

Nataliya Tkachenko[1,2], Arkaitz Zubiaga[2], Rob Procter[1,2,3]
[1] Warwick Institute for the Science of Cities, University of Warwick, Coventry CV4 7AL, UK
[2] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[3] The Alan Turing Institute, The British Library, London NW1 2DB, UK
{nataliya.tkachenko,a.zubiaga,rob.procter}@warwick.ac.uk

ABSTRACT

This working note describes the work of the WISC team on the Multimedia Satellite Task at MediaEval 2017. We describe the runs that our team submitted to both the DIRSM and FDSI subtasks, as well as our evaluations on the development set. Our results demonstrate high accuracy in the detection of flooded areas from user-generated content in social media. In the first subtask consisting of disaster image retrieval from social media, we found that tags defined by users to describe the images are very helpful for achieving high accuracy classification. In the second subtask consisting of detecting flood in satellite images, we found that social media can increase the precision in analyses when combined with satellite images by taking advantage of spatial and temporal overlaps between data sources.

1 INTRODUCTION

Accurate and timely designation of flooded areas is beneficial to help build and maintain situational awareness and to estimate the impact of natural hazards [4]. When it comes to the estimation of impact, there is no consistency across experts as to the different methods used to measure impact [3, 11]. Assessing and comparing disaster impact has traditionally been deemed a very challenging task as systematic data or studies are hard to obtain. Moreover, historical data gathered from different sources covering different regions cannot be effectively used for new regions or at different points in time. Examples of valuation techniques used for impact assessments include market based techniques, such as property destruction, reduction in income and sales, and non-market based techniques, such as loss of life, various environmental consequences and psychological effects suffered by the affected individuals [3, 6, 11].

Owing to the importance of furthering impact estimation techniques, interest in the development of computational approaches has increased [2, 12]. Current methods for flood detection and flood impact estimation make use of contemporary, open data sources such as social media [7, 9, 10]. The objective of this shared task and the contribution by the Warwick Institute for the Science of Cities (WISC) team is to assess the extent to which these techniques approximate the results obtained through traditional methods of flood detection, such as local sensor networks and satellite images [1]. This working note presents our efforts and achievements towards this objective.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2 RELATED WORK

Recent work has proposed to combine traditional data sources such as sensor networks (e.g., river gauges and pluviometers) with user-generated data from social media such as Twitter and Flickr (Tkachenko et al., under review). To the best of our knowledge, however, the combination of satellite images and georeferenced UGC has not been tackled in scientific research, potentially because it may be of limited use in areas with high percentage of cloud coverage (e.g., Northern Europe) or for being very expensive due to the need of sufficiently high spatio-temporal resolution.

With the growing availability of the free or inexpensive multi- and hyperspectral image tiles, it is becoming increasingly important to understand how such data sources perform alongside new methods and how their combined use can help overcome each other's limitations when used independently. With our participation in the Multimedia Satellite Task, we aimed to analyse how social media can be used to identify flooded areas, as well as to identify the best classification approaches.

3 APPROACH

3.1 DIRSM Subtask

3.1.1 Experiment Settings. We performed 10-fold cross-validation experiments. We used two different ways of evaluating our approaches: (1) precision, recall and F1 score over the positive (flood) class, and (2) Average Precision at X (AP@X) at various cutoffs, X={50, 100, 200, 300, 400, 500}. Since the official evaluation relies on the latter, we ended up choosing our best submissions based on AP@X, especially looking at X={50, 100, 200}, as the other values were rather high for our smaller test sets.

3.1.2 Features. We use combinations of these features:

- **Visual features:** having performed leave-one-out tests of combinations of the visual features provided by the organisers, we found the best combination to be that including CEDD, CL and GABOR.
- **Metadata:** we combined three of the metadata provided with the dataset, namely description, title and tags. The features were all represented using a bag-of-words approach, however, we built three separate vectors, one for each metadata, which were then concatenated into a single vector. With all three features, we lowercased the texts, and tokenised them by the space character. We also tokenised multi-word user tags.
- **Word embeddings:** we trained word embeddings from a large collection of titles, descriptions and user tags. We used the entire YFCC100m dataset [8] to get overall 215 million input texts combining all three types of features, which were fed into a Gensim word embedding with 300

dimensions [5]. These word embeddings were then used to create vectors of 300 dimensions for each of title, description and user tags of each image. To create word vectors for each sentence, we averaged the word vectors of the words composing the sentence, as in [13].

- **Machine translation:** we used the Bing machine translation API to translate user tags into English, where a user tag was not originally in English. We used the translation package for Python[1] to achieve this.

3.1.3 Classifiers. We tested different classifiers, including a Logistic Regression classifier, Random Forests, Multinomial Naive Bayes and Multi-layer Perceptron. We opted to build our system using a Logistic Regression classifier based on the performance observed on the development set. We used confidence scores provided by the classifier to rank the images.

3.2 FDSI Subtask

In this subtask, we performed the selection of the spectral images in the first instance, which were possible to construct from the 4-band spectral resolution data supplied. Selected indices in question were LWI (Land Water Index), NDVI (Normalised Difference Vegetation Index) and NDWI (Normalised Difference Water Index). For the subsequent runs we used machine learning methods for supervised classification and for unsupervised clustering machine learning. This was applied to the NDWI as the best performing spectral index in the first step of the FDSI task.

We also developed a second run, where we used KMeans to achieve binary image segmentations on the basis of the spatial distribution of the spectrally concentrated and transitioned pixels.

4 EXPERIMENTS AND RESULTS

4.1 DIRSM Subtask

Based on performance assessments, we chose these 5 submissions:

- **Run 1, visual information:** only visual features.
- **Run 2, metadata:** only metadata features.
- **Run 3, visual information and metadata:** both features by concatenating the two vectors.
- **Run 4, word vectors:** we concatenate five vectors for visual features, metadata, word vectors of titles, word vectors of user tags and word vectors of descriptions.
- **Run 5, machine translation and word vectors:** we concatenate five vectors for visual features, metadata, word vectors of titles, word vectors of machine translated user tags and word vectors of descriptions.

Run no.	X = 50	X = 100	X = 200	X = 300	X = 400
#1	0.980	0.990	0.995	0.793	0.596
#2	0.980	0.990	0.988	0.676	0.507
#3	0.980	0.990	0.979	0.671	0.504
#4	0.980	0.990	0.975	0.666	0.500
#5	0.980	0.990	0.974	0.665	0.500

Table 1: DIRSM results on the development set.

	Avg. over	
Run no.	X = {50, 100, 250, 480}	X = 480
#1	0.6275	0.5095
#2	0.7437	0.6678
#3	0.8087	0.7226
#4	0.8161	0.7197
#5	0.8199	0.7210

Table 2: DIRSM results on the test set.

Tables 1 and 2 show our results on the development and test sets, respectively. While results are similar over the development set, we observe remarkable differences in the test set. The metadata classifier (#2) performs better than that based on visual features (#1), however, the combination of both leads to substantial improvements (#3). There is still a considerable improvement when we used deep learning to represent the features using word vectors (#4), and a further slight improvement when we used machine translation to have all tags consistently in English (#5).

4.2 FDSI Subtask

Run no.	Jaccard (Dev. Set)
#1	0.83
#2	0.87

Table 3: FDSI results on the development set.

Run no.	Jaccard (Test Set 1)	Jaccard (Test Set 2)
#1	0.80	0.83
#2	0.81	0.77

Table 4: FDSI results on the test set.

Tables 3 and 4 show our results on the development and test sets, respectively. Our results show the benefit of leveraging social media features (#2) over not using them (#1) when training and testing data overlap spatially and temporally (Test Set 1). This is, however, not the case for the Test Set 2 where the test data includes new locations, which we aim to explore further in future work.

5 CONCLUSION

We have explored the use of classifiers to identify social media images of flooded areas. In the DIRSM task we have found that combining both visual features and social metadata can be beneficial, and that the use of external resources to train word embeddings and translate the metadata into English can lead to even further improvements. In the FDSI task, our results showed higher accuracy detection for the flooded areas with help of the social media classifiers. Social media can boost precision in combined analyses, where training and test data overlap spatially and temporally.

ACKNOWLEDGMENTS

We wish to thank the Alan Turing Institute for its support.

[1] https://pypi.python.org/pypi/translation

REFERENCES

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

[2] J Fohringer, D Dransch, H Kreibich, and K Schröter. 2015. Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences* 15, 12 (2015), 2725–2738.

[3] Valentina Gallina, Silvia Torresan, Andrea Critto, Anna Sperotto, Thomas Glade, and Antonio Marcomini. 2016. A review of multi-risk methodologies for natural hazards: Consequences and challenges for a climate change impact assessment. *Journal of environmental management* 168 (2016), 123–132.

[4] Benjamin Herfort, João Porto de Albuquerque, Svend-Jonas Schelhorn, and Alexander Zipf. 2014. Exploring the geographical relations between social media and flood phenomena to improve situational awareness. In *Connecting a digital Europe through location and place*. Springer, 55–71.

[5] 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.

[6] OECD. 2013. OECD: New Data for Understanding the Human Condition. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. (2013).

[7] Luke Smith, Qiuhua Liang, Phil James, and Wen Lin. 2015. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management* (2015).

[8] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[9] Nataliya Tkachenko, Stephen Jarvis, and Rob Procter. 2017. Predicting floods with Flickr tags. *PloS one* 12, 2 (2017), e0172870.

[10] Nataliya Tkachenko, Rob Procter, and Stephen Jarvis. 2016. Predicting the impact of urban flooding using open data. *Open Science* 3, 5 (2016), 160013.

[11] Gisela Wachinger, Ortwin Renn, Chloe Begg, and Christian Kuhlicke. 2013. The risk perception paradoxâĂŤimplications for governance and communication of natural hazards. *Risk analysis* 33, 6 (2013), 1049–1065.

[12] Huan Wu, Robert F Adler, Yudong Tian, George J Huffman, Hongyi Li, and JianJian Wang. 2014. Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model. *Water Resources Research* 50, 3 (2014), 2693–2717.

[13] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of the 26th International Conference on Computational Linguistics*.