2019-05-19

Comparative Study of Feature Representations for Disaster Tweet Classification

Pallavi Jain
pallavi.jain@mydit.ie

Bianca Schoen-Phelan
Technological University Dublin, bianca.phelan@tudublin.ie

Robert J. Ross
Technological University Dublin, robert.ross@tudublin.ie

Recommended Citation

Jain, Pallavi; Schoen-Phelan, Bianca; and Ross, Robert J., "Comparative Study of Feature Representations for Disaster Tweet Classification" (2019). *Other resources*. 16.
https://arrow.tudublin.ie/scschcomoth/16

Comparative Study of Feature Representations for Disaster Tweet Classification

Pallavi Jain, Bianca Schoen-Phelan, Robert Ross
School of Computer Science
Technological University Dublin, Kevin Street, Dublin, Ireland
{pallavi.jain, bianca.phelan, robert.ross}@dit.ie

INTRODUCTION

Twitter is a popular social media platform where users publicly broadcast short messages on a myriad of topics. In recent years it has enjoyed an increased usage around disaster events due to availability of information in near real time. Additionally, enhanced information representations to facilitate the classification of social media in terms of relevancy and type of information is currently a highly active research area (Ashktorab et al., 2014, Imran et al., 2014, Win et al., 2018). In this work we consider the usefulness and reliability of a range of representation models in the analysis of disaster related social media.

EXPERIMENT DESIGN

In order to assess the effectiveness of individual feature representations, we examine the performance of a classification task when applied to labelled tweets on two levels: first, on the basis of informativeness, and then on the basis of type of information. We used a total of 15 twitter datasets of different disasters where 6 datasets are taken from CrisisLex (Olteanu et al., 2014) and 9 datasets are taken from Crisis NLP (Imran et al., 2016). Each dataset consist of data labelled according to informativeness of the tweet on the particular event, and according to type of information of tweet. Pre-processing was applied to the dataset. Hashtags, URLs, punctuation, emoticons, special characters, and stop words, such as 'the' and 'a', were all stripped from the data. Also we used unigram, hybrid unigram and bigram features to add word to word relation features and part of speech (POS) tagging, which results in syntactic behavior of words (Schütze, 1995). We used the Gensim phrase model for bigrams, whereby all bigrams whose count is above three are considered and combined with tokenized unigrams of each tweet. For POS tagging CMU ARK Tweet NLP has been used. Furthermore, Spacy POS tagging removed location information as it contributes noise to a conversation. Additionally, the least 20 frequent words and words consisting of less than three characters were removed from the dataset before lemmatization was applied. Following this, duplicate tweets were removed using cosine similarity for those tweets having more than 90% similarity.

Feature representation that were investigated were BoW, TF-IDF, Word2Vec and Doc2Vec. The pre-trained Google word2vec model "GoogleNews-vectors-negative300.bin.gz" has been used for Word2Vec, for this POS tagging feature has not been considered as it is not supported by pre trained model. For Doc2Vec, each tweet is considered as a document and adds a Paragraph ID to each tweet with the tag of 'train' or 'test' for the respective data. The tagged data was then trained using the Genism Doc2Vec library (Rehurek et al., 2010) using both Distributed Bag of Words (DBOW) and Distributed Memory (DM) individually, in order to get 300 vector size document vectors from each model, which are then combined to get the document vector size of 600. Text categorization usually has high dimensionality. Thus, in order to reduce the dimensionality of data, Information Grain was used to limit features to the top 3,000 words features in the data.

RESULTS AND CONCLUSION

Our results are based on two types of data split; the first is leave one out (LOO), that is model tested on one dataset, and trains on the remaining 14 datasets. The second is cross-disaster training where a model is trained on one type of disaster, and then tested on another type of disaster dataset. As the dataset has unbalanced classes, the F1 score has been used as the evaluation metric. The results presented in Table 1 indicate that word2vec outperforms all other types of representation in leave one out as well as cross-disaster training approach, while there is very minute difference between hybrid n-gram and unigram. This could be due to the use of a pre-trained model, which does not consist of words related to hashtags and bigrams. The result of cross-disaster training shows that disaster related tweets appear to be independent of the actual type of disaster. This could be due to similar vocabulary shared across the type of disaster in terms of donation, sympathy, needs, etc. The results strongly suggest that a pre-trained word2vec model gives better results than BoW, TF-IDF and Doc2Vec. This is due to the fact that a pre-trained model is trained on a comparatively large corpus, which creates a much improved context similarity representation. However, word2vec struggles with out of vocabulary words, which is more prevalent in text from a Twitter feed, as it contains human created hashtags and slang words.

Data trained on:	**Earthquake**		**Flood**		**Storm**		**LOO Avg F1**	
Feature	**a**	**b**	**a**	**b**	**a**	**b**	**a**	**b**
BoW Unigram	0.67	0.63	0.68	0.71	0.67	0.66	0.76	0.74
BoW Hybrid	0.66	0.64	0.70	0.69	0.71	0.66	0.77	0.73
BoW POS	0.67	0.63	0.69	0.71	0.75	0.67	0.76	0.74
Doc2Vec Unigram	0.71	0.61	0.68	0.66	0.65	0.66	0.79	0.71
Doc2Vec Hybrid	0.72	0.62	0.68	0.66	0.67	0.66	0.78	0.71
Doc2Vec POS	0.72	0.62	0.67	0.67	0.63	0.66	0.78	0.71
TF-IDF Unigram	0.72	0.58	0.73	0.69	0.63	0.65	0.79	0.75
TF-IDF Hybrid	0.69	0.54	0.72	0.62	0.62	0.60	0.79	0.74
TF-IDF POS	0.72	0.58	0.72	0.70	0.72	0.65	0.72	0.75
W2Vec Unigram	**0.78**	**0.70**	**0.75**	**0.73**	**0.81**	**0.71**	**0.81**	**0.76**
W2Vec Hybrid	0.76	0.68	**0.75**	**0.73**	**0.81**	0.70	0.80	**0.76**
W2Vec POS	NA	NA	NA	NA	NA	NA	NA	NA

Table 1: **Average F1 score for a = Informativeness Classification Result, and b = Information Type Classification**

REFERENCES

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014, May). Tweedr: Mining twitter to inform disaster response. In *ISCRAM.*

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014, April). AIDR: Artificial intelligence for disaster response. *In Proceedings of the 23rd International Conference on World Wide Web* (pp. 159-162). ACM.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894.*

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. *In Eighth International AAAI Conference on Weblogs and Social Media.*

Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora.* In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Schütze, H. (1995). Distributional part-of-speech tagging. *arXiv preprint cmp-lg/9503009.*

Win, Si Si Mar, and Than Nwe Aung. 2018. "Automated Text Annotation for Social Media Data during Natural Disasters." *Advances in Science, Technology and Engineering Systems Journal* 3(2): 119–27. https://astesj.com/v03/i02/p14/.

Yang, Y., & Pedersen, J. O. (1997, July). *A comparative study on feature selection in text categorization.* In Icml (Vol. 97, No. 412-420, p. 35).