Supporting Experts to Handle Tweet Collections About Significant Events

Ali Hürriyetoğlu[1](✉), Nelleke Oostdijk[2], Mustafa Erkan Başar[2], and Antal van den Bosch[2,3]

[1] Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands
a.hurriyetoglu@cbs.nl
[2] Centre for Language Studies, Radboud University,
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
n.oostdijk@let.ru.nl, ebasar@science.ru.nl
[3] Meertens Institute, Amsterdam, The Netherlands
antal.van.den.bosch@meertens.knaw.nl

Abstract. We introduce Relevancer that processes a tweet set and enables generating an automatic classifier from it. Relevancer satisfies information needs of experts during significant events. Enabling experts to combine automatic procedures with expertise is the main contribution of our approach and the added value of the tool. Even a small amount of feedback enables the tool to distinguish between relevant and irrelevant information effectively. Thus, Relevancer facilitates the quick understanding of and proper reaction to events presented on Twitter.

Keywords: Social media · Text mining · Machine learning · Twitter

1 Introduction

Tweet collections comprise a relatively new document type. The form, motivation behind their generation, and intended function of the tweets are more diverse than traditional document types such as essays or news articles. Understanding this diversity is essential for extracting relevant information from tweets. However, the Twitter platform does not provide any kind of infrastructure other than hashtags, which is limited by the number of users who know about it, to organize tweets. Thus, collecting and analyzing tweets introduces various challenges [4]. Using one or more key terms and/or a geographical area to collect tweets is prone to cause the final collection to be incomplete or unbalanced [5], which in turn decreases our ability to leverage the available information.

In order to alleviate some of the problems that users experience, we developed Relevancer which aims to support experts in analyzing tweet sets collected via imprecise queries in the context of high-impact events.[1] Experts can define their information need with up-to-date information in terms of automatically detected information threads [1], and use these annotations to organize unlabeled tweets.

[1] https://bitbucket.org/hurrial/relevancer.

F. Frasincar et al. (Eds.): NLDB 2017, LNCS 10260, pp. 138–141, 2017.
DOI: 10.1007/978-3-319-59569-6_14

The main observations and contributions that are integrated in Relevancer to help experts to come to grips with event-related event collections are: (i) almost every event-related tweet collection comprises tweets about similar but irrelevant events [1]; by taking into account the temporal distribution of the tweets about an event it is possible to achieve an increase in the quality of the information thread detection and a decrease in computation time [2]; and the use of inflection-aware key terms can decrease the degree of ambiguity [3].

Section 2 outlines the main processing stages and Sect. 3 describes a use case that demonstrates the analysis steps and the performance of the system. Finally, Sect. 4 concludes this paper with a brief summary, some remarks on the system's performance, and directions for future development.

2 System Architecture

Relevancer consists of the following components:

Filtering. We handle frequent hashtags and users separately. Upon presenting them to the expert for annotation, each of these is represented by five related sample tweets. If the expert decides that a user or a hashtag is irrelevant, tweets that contain this hashtag or were posted by this particular user are kept apart. The remaining tweets are passed on to the next step.

Pre-processing. The aim of the pre-processing is to convert the collection to more standard text without loosing any information. An expert may choose to apply all or only some of the following steps. As a result, the expert is in control of any bias may arise due to preprocessing.

RT Elimination. Any tweet that starts with a 'RT @' or that in its meta-information has an indication of it being a retweet is eliminated.

Normalization. User names and URLs in a tweet text are converted to 'usrusr' and 'urlurl' respectively.

Text cleaning. Text parts that are auto-generated, meta-informative, and immediate repetitions of the same word(s) are removed.

Duplicate elimination. Tweets that after normalization have an exact equivalent in terms of tweet text are excluded from the collection.

Near-duplicate elimination. A tweet is excluded, if it is similar to another tweet, i.e. above a certain threshold in terms of cosine-similarity.

Clustering. We run KMeans on the collection recursively in search of coherent clusters using tri-, four- and five-gram characters. Tweets that are in the automatically identified coherent clusters are kept apart from the subsequent iterations of the clustering. The iterations continue by relaxing the coherency criteria until the requested number of coherent clusters is obtained.

Annotation. Coherent clusters are presented to an expert in order to distinguish between the available information threads and decide on a label for them. Next, if the expert finds the cluster coherent, she attaches the relevant label to it. Otherwise she marks it as incoherent[2]. The cluster annotation speeds the labeling process in comparison to individual tweet labeling.

Classifier Generation. For training, 90% of the labeled tweets are used, while the rest is used to validate a Support Vector Machine (SVM) classifier.

3 Experiments and Evaluation

We demonstrate the use and performance of the Relevancer on a set of 229,494 Dutch tweets posted between December 16, 2010 and June 30th, 2013 and containing the key term 'griep' (EN: flu). After the preprocessing, retweets (24,019), tweets that were posted from outside the Netherlands (1,736), clearly irrelevant (158), and exact duplicates (8,156) were eliminated.

Our use case aims at finding personal flu experiences. Tweets in which users are reporting symptoms or declaring that they actually have or suspect having the flu are considered as relevant. Irrelevant tweets are mostly tweets containing general information and news about the flu, tweets about some celebrity suffering from the flu, or tweets in which users empathize or joke about the flu.

The remaining 195,425 tweets were clustered in two steps. First we split the tweet set in buckets of ten days each. The bucketing method increases the performance of and decreases the time spend by the clustering algorithm [2]. We set the clustering algorithm to search for ten coherent clusters in each bucket. Then, we extract the tweets that are in a coherent cluster and search for ten other clusters in the remaining, un-clustered, tweets. In total, 10,473 tweets were put in 1,001 clusters. Since the temporal distribution of the tweets in the clusters has a correlation of 0.56 with the whole data, aggregated at a day level, we consider that the clustered part is weakly representative of the whole set.

We labeled 306 of the clusters: 238 were found to be relevant (2,306 tweets), 196 irrelevant (2,189 tweets), and 101 incoherent (985 tweets). The tweets that are in the relevant or irrelevant clusters were used to train the SVM classifier. The performance of the classifier on the held-out 10% and unclustered part are illustrated in the Table 1.

The baseline of the classifier is the prediction of the majority class in the test set, in which the precision and recall of the minority class is undefined. Therefore, we compare the generated classifier and the baseline, which have 0.67 and 0.56 accuracy respectively.

The results show that Relevancer can support an expert to manage a tweet set under rather poor conditions. The expert can create a tweet set using any key word, label automatically detected information threads, and have a classifier

[2] The label definition affects the coherence judgment. Specificity of the labels determines the required level of the tweet similarity in a cluster.

Table 1. Classifier performance on the validation and 255 unclustured tweets

	Validation set				Unclustered			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Relevant	.95	.96	.96	237	.66	.90	.76	144
Irrelevant	.96	.94	.95	213	.75	.39	.51	111
Avg/Total	.95	.95	.95	255	.70	.68	.65	255

that can classify the remaining (unclustered) tweets or new tweets. A classifier generated as a result of this procedure will perform between 0.67 and 0.95 accuracy when applied straightforwardly.

4 Conclusion

We described our processing tool that enables an expert to explore a tweet set, to define labels for groups of tweets, and to generate a classifier. At the end of the analysis process, the experts understands the data and is able to use his understanding in an operational setting to classify new tweets. The worst case performance of the classifier is significantly better than a majority class based baseline. The tool is supported by a web interface that can potentially be used to monitor and improve the performance in a particular use-case in real time.

In further research, we will integrate visualization of the tweet and cluster distributions, add additional interaction possibilities between the tweet set and the expert, refine clusters based on their inter-cluster distance distribution from the cluster center, and improve the flexibility of the bucketing for the clustering.

Acknowledgments. COMMIT, Statistics Netherlands, and Floodtags supported our work.

References

1. Hürriyetoğlu, A., Gudehus, C., Oostdijk, N., Bosch, A.: Relevancer: finding and labeling relevant information in tweet collections. In: Spiro, E., Ahn, Y.-Y. (eds.) SocInfo 2016. LNCS, vol. 10047, pp. 210–224. Springer, Cham (2016). doi:10.1007/978-3-319-47874-6_15
2. Hürriyetoğlu, A., van den Bosch, A., Oostdijk, N.: Using relevancer to detect relevant tweets: the Nepal earthquake case. In: Working Notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India. December, 2016. http://ceur-ws.org/Vol-1737/T2-6.pdf
3. Hürriyetoğlu, A., van den Bosch, J.W.A., Oostdijk, N.: Analysing the role of key term inflections in knowledge discovery on twitter. In: Proceedings of the 1st International Workshop on Knowledge Discovery on the WEB, Cagliari, Italy, September, 2016. http://www.iascgroup.it/kdweb2016-program/accepted-papers.html
4. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Comput. Surv. (CSUR) **47**(4), 1–38 (2015). http://doi.acm.org/10.1145/2771588
5. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: CrisisLex: A lexicon for collecting and filtering Microblogged communications in crises, pp. 376–385. The AAAI Press (2014)