DOI: http://dx.doi.org/10.14236/ewic/FDIA2017.4

Early Prediction of Public Reactions to News Events Using Microblogs

Cagri Toraman
Information Retrieval Group, Computer Engineering Department, Bilkent University, Ankara, 06800, Turkey
ctoraman@cs.bilkent.edu.tr
Big Data and Cloud Computing Group, Havelsan A.Ş., Ankara, 06800, Turkey
ctoraman@havelsan.com.tr

Microblog environments like Twitter are increasingly becoming more important to leverage people's opinion on public events. We aim to predict future public reactions to news events by exploiting related tweets. We define public reactions in terms of their dimension and direction. Our system collects and preprocesses tweets, creates an inverted index to search tweets efficiently, filters them with various methods according to news events; and then uses temporal, spatial and textual features to model predictive classifiers. We also create a public-reaction dataset, BilPredict-2017, which includes several events including terrorist attacks in Turkey from 2015 to 2017. We plan to model ensemble classifiers, and evaluate the success of our system on BilPredict-2017.

Microblogs, news, prediction, public reaction, sentiment analysis, tweets

1. INTRODUCTION

Traditional news sources like newspapers provide limited information, due to the drawbacks of slow editorship and time restrictions for reaching event sources. Tweets are recently used as a dynamic news source; a typical example is getting updated with correct information about disasters (Imran et al., 2015). Another kind of information that tweets expose is peoples opinions. With the growth of social media usage, opinions of crowds can be processed to understand mass behaviors, like riots in the Arab Spring (González-Bailón et al., 2011).

We define public reaction as peoples acts or behaviors that result from common opinions for an event occurred at a particular time and place. In today's society of multiple views, it is difficult to estimate the dimension and direction of public reactions for events. For a news article titled "Fire Department saves cat from tree", no public reaction is expected. On the other hand, mass discontent of people in social media for a news article titled "bloody balance sheet of holiday accidents" is an example for negative reaction. The reaction associated with the accidents does not involve any protests in the streets; however, workers can protest a radical change in an employment law in the streets. In addition to the dimension, its direction (negative vs. positive) is also important. An example of positive reaction is peaceful post-match celebrations of football fans after a championship.

Recent research utilize social media for prediction of consumer behaviors (Asur and Huberman, 2010), or real-time event detection (Alsaedi, Burnap and Rana, 2017). In this study, given a new article, we aim to detect future events as a result of the news event, by using microblog texts, specifically tweets. We collect and clean tweets; filter them adaptively according to a given news article; and then use temporal, spatial and textual features to model predictive classifiers. More specifically, we develop a system that classifies news articles into several classes that define resulting public reactions in terms of dimension (national, local, social media, and no reaction) and direction (positive, negative, and neutral). Early prediction of public reaction for an event supports government institutions, commercial organizations, and individuals to prepare themselves, in terms of precautions for future negative events, and advantageous responses for future positive events.

Our contributions are the following. We predict the future of an event that is published in a given news article, in terms of both dimension and direction. We create a public-reaction dataset including terrorist attacks in Turkey from 2015 to 2017. We develop a pipeline system that collects and filters tweets

Proceedings of the 7th Symposium on Future Directions in Information Access 2017

efficiently, and extracts temporal, spatial and textual features to create a prediction model.

In the next section, a summary of related work is given. We then explain our system in details, describe our dataset, and lastly conclude the paper.

2. RELATED WORK

Identifying real-time events using a text stream like tweets is called event detection. Becker, Naaman and Gravano (2011) cluster tweets to identify events on-the-fly, and then, tweets can be classified as event-related or not. Alsaedi, Burnap and Rana (2017) detect already-occurred events of public reactions like riots. Such studies are not event predictors, but detectors for already-occurred events that exploits dynamic nature of social media.

In the last decade, prediction with social media is a hot research topic. Researchers mostly try to predict future consumer behaviors like book sales (Gruhl et al., 2005). Some studies do not use features that are extracted from social media, but other sources. For instance, Bandari, Asur and Huberman (2012) exploit news features to predict news popularity in social media. In this study, we do not predict news popularity, but the dimension and direction of the public reaction to a news event.

Kallus (2014) utilizes big data of web including social media to predict crowd behavior such as significant protests. Their prediction is correct when there is a significant protest in the given country during the following three days. Muthiah et al. (2015) develop a protest-prediction system that uses several web sources, such as social media and RSS feeds. Our difference is that we define target public-reaction classes that identify the dimension and direction of a future event, and provide an early-prediction system for a given arbitrary news article.

3. OUR PREDICTION SYSTEM

The pipeline of our system is given in Figure 1. We first collect and preprocess tweets. We assume that news events are given by the user, but in the experiments, news events are obtained from the training data. Cleaned tweet contents are kept in an inverted index to have filtering efficiently. Also, sentiment scores are calculated, and kept in a hash table to fetch them for feature extraction efficiently. Our filtering mechanism is adaptive, that is, we get descriptive keywords from user to represent news article. Lastly, we extract temporal, spatial and textual features, including sentiments, to train predictive classifiers. For training, we use our newly generated public-reaction dataset, BilPredict. We use classifiers to predict future public reaction to a current news article.

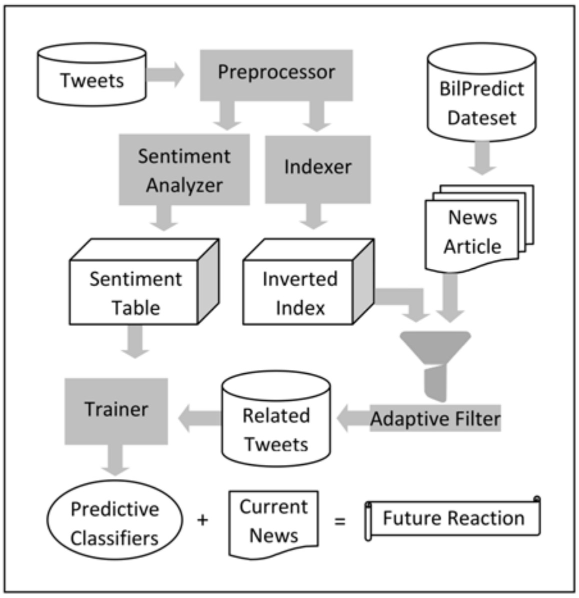

Figure 1: *Pipeline of our system for prediction of public reactions.*

3.1. Preprocessing Tweets

Given a tweet collection, we have the following preprocess operations. Turkish tweets are detected by the language attribute of Twitter API. We divide tweet content into tokens by space character. Noise in tokens, such as characters not in the dictionary, are removed. Numbers in tokens are ignored. Hashtag character is not removed to keep trend topics in inverted index. Invalid emoticon expressions are removed. We create a list of positive, neutral, and negative emoticon characters; and keep them to be exploited in the sentiment-analysis phase.

Turkish has special accent characters like 'ç' or 'ü'. In microblogs, Turkish users sometimes type the ASCII version of these accent characters, which results in missing or ambiguous words for our sentiment analyzer and indexer. Replacing ASCII characters with intended original ones is called *deasciification*. In this study, we develop a deasciification approach that recursively produces all accent versions of a token, and chooses the one that has the maximum document frequency on the given collection. The idea is the more a token version is used by microblog users, the more potential it is correct. Current approaches mostly use in-vocabulary (IV) lexicon; we do not use it for a couple of reasons. First, we make our algorithm generic to all collections in different languages. Second, we capture the cases that training sets cannot capture, such as a new phrase or slang word is produced by users.

We create a lexicon for common abbreviations, and normalize tweets accordingly. We also remove tokens that have length more than 20.

Stems of tokens are obtained by Zemberek (Akın and Akın, 2007), which is a popular Turkish stemmer. To avoid ambiguity, we consider only the first root that is found by Zemberek. We also apply the stopword list given in the study of Can et al. (2008).

3.2. Sentiment Analyzer

SentiStrength (Thelwall et al., 2010) is a popular sentiment analyser for English, which can be adapted to other languages by modifying its lexicon files. For emotions and booster words, we use the Turkish lexicon files that are constructed by Türkmenoglu and Tantug (2014). Turkish has two negation forms; having negating (1) suffix, or (2) words after verbs and nouns. We examine words morphologically with Zemberek, and add heuristic rules to detect negations. We also detect emoticons by checking our emoticon lexicon, and replace positive ones with ":)", and negative ones with ":(". We then give the processed content to SentiStrength.

3.3. Filtering and Indexing

Finding related tweets to a given news article is a challenging task due to a couple of reasons. First, we have a very huge collection of tweets, so with a lazy approach, it would take days to get a subset. Second, directing unrelated tweets to the training phase, i.e false positives, misleads the training model. Considering such concerns, we plan to examine four methods to filter tweets: Dice coefficient, cosine similarity on vector space model, boolean search on inverted index that sets the news title as query, and adaptive search on inverted index that gets query keywords from users.

3.4. Training Prediction Models

In order to model predictive classifiers, we plan to obtain temporal, spatial and textual features from the results of sentiment analyzer and filtered tweets by inverted index. Figure 2 represents our training method for public reactions. We aim to train multiple classifiers of different time windows. We believe that ensemble of time-window classifiers can reflect dynamic structure of social media.

4. EXPERIMENTS

We create a new public-reaction dataset, called BilPredict-2017 (2017) that consists of three components. First component is the ground truth that has 80 news events/articles occurred between 2015 and 2017. Each event is listed with its origin date, place, news url, public-reaction category, and reaction tags. Public reactions are labeled by experts, in terms of dimensions and directions. Dimensions are in terms of national, local, and social media. National categories represent public reactions occurred in at least two different cities. Local categories have events occurred at only a specific place. Social categories represent reactions that people share opinions only in social media, such as microblogs. Directions are either negative or positive. At the end, we have 7 categories: national positive, national negative, local positive, local negative, social positive, social negative, and no reaction.

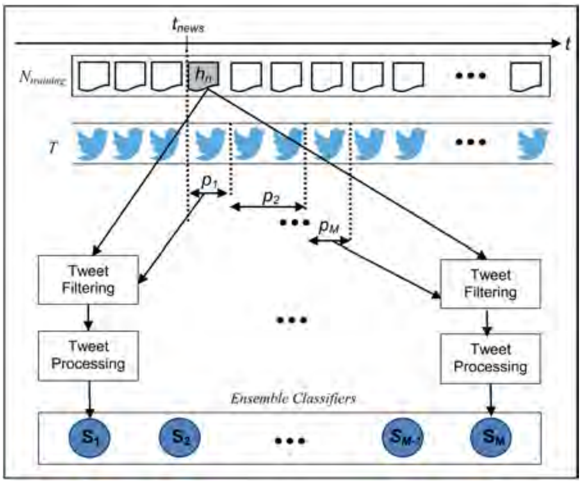

Figure 2: *Ensemble training of predictive models for public reactions. Input news article is h_n. We construct multiple sets of tweets that are obtained after the origin date of the input event. Tweet sets have M time windows, labeled with p_i, where $1 < i < M$. For each set, we get related tweets by filtering, and then process them to get useful features. Lastly, we exploit those features to train ensemble classifiers for public-reaction prediction.*

The second component is the html file contents of 80 news events. They can be used as input to our prediction system. The last component is the tweets for 80 events. For the next 10 days after the origin date of each news event in BilPredict-2017, we collect approximately 1.3 billions tweets. These tweets can be exploited to create features for prediction models.

Evaluation metrics are the accuracy of prediction, and noise time. Accuracy is given by the number of correctly predicted news events, divided by the number of all predictions. Noise time is the time that our prediction system correctly indicates noise from the signal, which is the public relevance of news events.

We plan to discuss language independency of our system, and also the effect of using sentiment analysis, stemming, deasciification, and the type of classification algorithm on effectiveness. Baseline methods to be compared are logistic regression and density estimator.

5. CONCLUSION

This study aims to predict public reactions to news events. Given a news article as input, our pipeline starts with fetching at most 10 days of tweets after the origin date of news event. We then preprocess tweets that include various cleaning, normalization, and stemming steps. Processed tweets are ready to be exploited by sentiment analyzer and inverted indexer. Related tweets to input news event are then found by inverted index efficiently. Using textual, spatial, and temporal features of those related tweets with also sentiment scores, our aim is to model an ensemble of classifiers of different time windows. Varying time windows would reflect the dynamic environment of social media. We also create a new dataset, called BilPredict-2017, for public-reaction prediction that includes news events between 2015 and 2017. In future work, we plan to model ensemble classifiers, and evaluate the success of our system on BilPredict-2017.

ACKNOWLEDGEMENT

This study is supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) grant no. 215E169.

REFERENCES

Akın, A.A. and Akın, M.D. (2007) Zemberek, an open source NLP framework for Turkic languages. *Structure*, vol. 10, pp. 1-5.

Alsaedi, N., Burnap, P. and Rana, O. (2017) Can we predict a riot? Disruptive event detection using twitter. *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 2, p. 18.

Asur, S. and Huberman, B.A.. (2010) Predicting the future with social media. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, pp. 492-499.

Bandari, R., Asur, S. and Huberman, B.A. (2012) The pulse of news in social media: Forecasting popularity. *ICWSM'12*, pp. 26-33.

Becker, H., Naaman, M. and Gravano, L. (2011) Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM'11*, pp. 438-441.

Bilkent IR Group. (2017) BilkentInformationRetrieval-Group/ TUBITAK215E169 [Online]. Available at: https://github.com/BilkentInformationRetrievalGroup/ TUBITAK215E169 [Accessed: 28 August 2017].

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H.C. and Vursavas, O.M. (2008) Information retrieval on Turkish texts. *Journal of the Association for Information Science and Technology*, vol. 59, no. 3, pp. 407-421.

González-Bailón, S., Borge-Holthoefer, J., Rivero, A. and Moreno, Y. (2011) The dynamics of protest recruitment through an online network. *Scientific Reports*, vol. 1, pp. 197.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005) The predictive power of online chatter. *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 78-87.

Imran, M., Castillo, C., Diaz, F. and Vieweg, S. (2015) Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, pp. 67.

Kallus, N. (2014) Predicting crowd behavior with big public data. *Proceedings of the 23rd International Conference on World Wide Web*, pp. 625-630.

Muthiah, S., Huang, B., Arredondo, J., Mares, D., Getoor, L., Katz, G. and Ramakrishnan, N. (2015) Planned protest modeling in news and social media. *AAAI'15* pp. 3920-3927.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010) Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, vol. 61, no. 12, pp. 2544-2558.

Türkmenoglu, C. and Tantug, A.C. (2014) Sentiment analysis in Turkish media. *International Conference on Machine Learning (ICML'14).*