Identifying News from Tweets

Jesse Freitas and Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY, USA
{freitj,jih}@rpi.edu

Abstract

Informal genres such as tweets provide large quantities of data in real time, which can be exploited to obtain, through ranking and classification, a succinct summary of the events that occurred. Previous work on tweet ranking and classification mainly focused on salience and social network features or rely on web documents such as online news articles. In this paper, we exploit language independent journalism and content based features to identify news from tweets. We propose a novel newsworthiness classifier trained through active learning and investigate human assessment and automatic methods to encode it on both the tweet and trending topic levels. Our findings show that content and journalism based features proved to be effective for ranking and classifying content on Twitter.

1 Introduction

Due to the massive amount of tweets posted on a daily basis, automatic tweet ranking has become an important task to assist fast information distillation. Previous work (Inouye and Kalita, 2011; Yang et al., 2011; Liu et al., 2012; Ren et al., 2013; Shou et al., 2013; Chua and Asur, 2013; Chang et al., 2013) focused on selecting informative tweets based on a variety of criteria such as readability, author's influence and users' interest. In addition, (Štajner et al., 2013) studied selection of user's responses to news by optimizing an objective function which jointly models the messages' utility scores and their entropy. (Wei and Gao, 2014) proposed using learning-to-rank techniques to help conduct single-document summarization.

Additional work has been done to improve event detection on Twitter. Previous methods relied on metadata from tweets (Cui et al., 2012) while others focus on open-domain categorization using tools trained specifically for micro-blogging services(Ritter et al., 2012a). In addition, (Cataldi et al., 2010) incorporated temporal knowledge while formalizing content to compensate for informal and short text.

Tweet ranking is also related to the previous work concerning tweet summarization which summarized important information from tweet streams. Modern summarization approaches rank sentences or phrases from informal genres such as social media, web forums, and micro-blogging sites. Some methods determine semantic relations of manually annotated hashtags and user replies using social network and web document graphs (e.g., (Huang et al., 2012)). Our goal is to accomplish ranking of micro-blogging content using natural language features to identify sentences with the most newsworthiness and relevance to the event that occurred.

We introduce a novel *newsworthiness* model to improve topic classification, and investigate both human assessment and automatic linguistic features in-depth for this new measure. Once topics are identified, we apply these methods to an ordinal ranking model to identify the tweets that make this topic newsworthy. Compared with previous work, we focus more on analysis of text than social network features for ranking and classifying tweets. In order to determine newsworthiness, we use news values based on journalism theory to encode features in-

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 11–16,
Austin, TX, November 5, 2016.

stead of traditional methods based on social features.

2 Approach

2.1 News Values and Definition

Newsworthiness describes the amount of new information for a general audience. (Galtung and Ruge, 1965) describe news as a spike in human communication or a signal that can be measured, and trending topics on Twitter behave this way. However, trending topics on social media can also be jokes, and ongoing and commemorative events. We hypothesize that newsworthiness would be an important factor in human distillation of media regarding events because it is a subset of salience that contains only novel information that is time relevant and new compared to an existing knowledge base. We define the following novel criteria to determine the newsworthiness of content based on news values defined by Galtung and Ruge (1965):

1. The content tends to refer to a negative event more than a positive event
2. The content must be well composed
3. The content typically refers to elite nations, people, or organizations
4. The content must have human interest

The basis behind newsworthy criteria is that (1) the content must be important to the general viewer but must provide new insight to an event that occurred; (2) because news content is salient, but salient content is not always newsworthy, understanding this subset is critical for automatic news summarization; (3) negative news is typically more viewed than positive news and usually pertains to named entities that are high profile; and (4) news must also have human interest meaning it must affect many people. Using Galtung and Ruge's metaphor of a signal for news, these principles should indicate a strong signal or spike in news.

The non-syntactic features listed in Table 1 are calculated as the number of words in the tweet and the normalized features are calculated as the ratio of the number of sentiment words to the total number of words in the tweet not including stopwords. The named entities and slangs were extracted using the Twitter NLP toolkit (Ritter et al., 2011; Ritter et al., 2012b) which was designed specifically for tweet content. The syntax tree features were calculated using the Stanford parser (Manning et al., 2014) trained using the English caseless model (de Marneffe et al., 2006). The premise behind using the parse tree as a feature is that more complex speech is more likely to be newsworthy because it is well composed. Sentiment terms were determined based on lexical matching from gazetteers(Hu and Liu, 2004; Taboada and Grieve, 2004; Wiebe et al., 2004; Baccianella et al., 2010; Joshi et al., 2011) and compiled into one sentiment dictionary (Li et al., 2012). Normalized stopword usage is important for both composition and human interest particularly because of the structure of a tweet. Since tweets are short and contain few words, if a tweet uses a high proportion of stopwords, it likely doesn't have many novel terms that would contain human interest. The remaining features are encoded based on the principle that generally recognized names and events are important for detecting topically familiar and important materials.

Feature	News Value
Slang Usage	2
First Person Usage	4
Geo-Political Entities	3
People Recognized	3
Companies Recognized	3
Sentiment Usage	1, 4
Normalized Sentiment Usage	1, 4
Normalized Stopwords Usage	2, 4
Max Parse Tree Height	2
Max NP Parse Tree Height	2
Max VP Parse Tree Height	2

Table 1: Newsworthiness features and news values they encode.

2.2 Newsworthiness Identification

There are two tasks to complete to identify newsworthy, salient content. The first is to identify the tweets within a topic that make the trending topic newsworthy. The second task is to identify trending topics on Twitter that meet the criteria for news values. To accomplish these tasks we use two Support Vector Machine based methods to perform news classification on trending topics and ordinal regression for ranking tweets in the topic. The models are trained using an eighth order polynomial kernel with default parameters and we tune the cost parameter based on the task. In order to train these models, we use the same 11 features in both tasks based on news criteria journalists use to write articles.

For identifying trending topics, our goal was to improve the precision and recall of existing systems so the model was tuned to maximize F-score performance using three fold cross validation to maintain consistency with the validation used by Huang et al. (2012). The ordinal regression model for ranking tweets was tuned using the same cross validation method to minimize the squared error from ground truth ranking.

We also evaluate an *actively trained* model for classification similar to the committee method used by Zubiaga et al. (2015). We choose candidates using *Query-By-Committee (QBC)* (Settles, 2009) where multiple models are trained using the same data and predict the class of each Twitter trend. For our committee we use our journalism based model and Zubiaga's Twitter based model. The contradicting predictions from the two models are used to choose the new training data. We use one iteration for creating candidates and our journalism model is then retrained using the new training data subset selected by the committee.

3 Experiments

3.1 Data

Our first ground truth dataset for classifying tweets was collected using CrowdFlower[1] to annotate the newsworthiness of 3,482 topically related tweets in English about Hurricane Irene. The dataset was collected during three separate hours during three different days shown in Table 2. We hypothesize that the subevents related to the topic will affect the amount of newsworthy content we are attempting to rank and may affect the performance of the ordinal regression model.

The ordinal regression dataset is composed of the same tweets used by Huang et al. (2012). Five annotators labeled the tweets' newsworthiness from one to three where three is most newsworthy. Annotators were given a brief summary of the events that occurred regarding Hurricane Irene and the areas that were impacted by flooding and power outages. They were provided a guideline for newsworthiness and asked to score the tweets on whether they contained new information at the time it was tweeted and would be noteworthy content in the media. The tweets were filtered to remove annotations if the CrowdFlower site's annotator confidence score was below 50 percent.

Date	Event
Aug. 27th, 2011	Irene landfall in NC
Aug. 28th, 2011	Irene landfall in NYC
Sept. 1st, 2011	Irene dissipates

Table 2: Tweets were collected for one hour on each day during the storm

The second dataset is comprised of 2,593 trending topics used by Zubiaga (2013). The topics were collected from February 1 to 28, 2012, and five topics were randomly sampled per hour. Each topic contains at most 1,500 tweets and the topic is labeled newsworthy or not newsworthy. The tweets are in multiple languages including English, Spanish, and Portuguese and were translated to English by both human and machine translators. This dataset was selected to evaluate the news features on a set of more diverse events than Hurricane Irene.

In order to demonstrate the effectiveness of our approaches, we evaluated the features on both the tweet ranking and trend classification tasks by comparing them to the performance of other approaches.

3.2 Evaluation

The ordinal regression and classification models were evaluated separately from each other to determine individual performance.

To compare our ranking method we used *Normalized Discounted Cumulative Gain* (nDCG) (Järvelin and Kekäläinen, 2002) and evaluated the results on the three individual hours of data.

The classification task is evaluated using precision, recall, and F-score and is compared using a baseline approach for classifying news trends (Zubiaga et al., 2015). The baseline approach classifies trending topics as news, or not news using so-

[1] http://www.crowdflower.com/

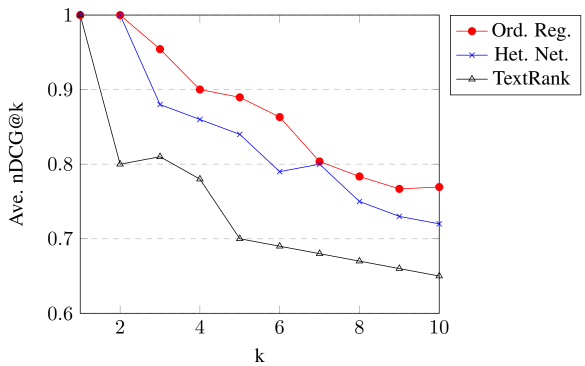

Figure 1: nDCG@k for Ordinal Regression, Heterogeneous Networks, TextRank

cial features such as user diversity, hashtag usage, and retweets about the topic, but does not consider as many language and content features.

4 Results

4.1 Ranking Individual Tweets

Figure 1 illustrates the evaluation of each method on nDCG@k from 1 to 10. The results indicate that our ordinal regression model performed better in terms of nDCG than the traditional TextRank method using the standard dampening factor with filtering and heterogeneous networks without web documents (Huang et al., 2012). The edges are calculated using cosine similarity between tweets and the filtering used removed tweets that used excessive slang or punctuation. The ordinal regression curve in Figure 1 represents the average performance of our model after evaluating the model on three different time periods of data described in Section 3.1.

4.2 Trend Classification

Method	Precision	Recall	F-Score
Baseline Features	0.582	0.670	0.623
Content Features	0.604	0.743	0.663
Active Training	**0.814**	**0.745**	**0.778**

Table 3: Trend Classification

Using journalism content features we were able to achieve better performance than our baseline(Zubiaga et al., 2015) in terms of precision and F-score while maintaining recall as shown in Table 3. Further, the model performed best when actively trained using the same journalism features and achieved a final F-score of 77.8%.

5 Discussion

We determine the statistical significance of each feature for both the trend classifier and the tweet ranker. We found features in each task were highly correlated and share overlap. For the sake of clarity, we only show significant features in Table 4.

Feature	Rank	Class
Slang Usage	***	
Geo-Political Entities	**	
Normalized Stopword Usage		***
Sentiment Terms		***
Company Entities	*	***
First Person Usage	***	***
NP Parse Tree Height	.	***

Table 4: F-statistic significant features. We show only significant features (significance codes: 0 (***) 0.001 (**) 0.01 (*) 0.05 (.) 0.1 ( )). *Rank* is the significance of the features in the tweet ranking task and *Class* is the significance of the features in the trend classification task.

Newsworthiness can affect how quickly and how much novel information can be discerned respectively. One of the goals of incorporating different criteria into ranking and classification other than traditional importance ranking was to demonstrate that salience is not the only factor that users and journalists consider when digesting material from social media. Another goal is to demonstrate that content based features can perform as well as other modern approaches that rely on web documents and social media graphs in order to bypass the challenge of understanding the short context-free nature of microblog posts.

In this paper we propose and evaluate two individual tasks used in identifying news on Twitter. We find that with the use of active learning and content based features we are able to significantly improve the precision of trend classification while maintaining recall. One challenge we faced was that Zubiaga's features for trending topics did not extend well to single tweet features for ranking. Because of this, we were unable to evaluate query-by-committee methods on ordinal regression which is something we would like to explore in the future.

While the features we used are not advanced, the application of them encode Galtung and Ruge's standards of world news for journalists and news reporting agencies. Our features attempt to capture a journalism perspective instead of previous work which focused on social media networks and social features. While this approach has limitations, the application of this approach in conjunction with web documents could improve news summarization tasks using Twitter data.

6 Acknowledgements

This work was supported by the U.S. ARL NS-CTA No. W911NF-09-2-0053, DARPA DEFT No. FA8750-13-2-0041, and NSF IIS-1523198. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA. ACM.

Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. 2013. Towards twitter context summarization with user influence models. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 527–536. ACM.

Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic summarization of events from social media. In *ICWSM*. Citeseer.

Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1794–1798, New York, NY, USA. ACM.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *IN PROC. INTL CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC*, pages 449–454.

Johan Galtung and Marie Holmboe Ruge. 1965. The structure of foreign news. *The Journal of Peace Research*, 2(2):64–91.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Hongzhao Huang, Arkaitz Zubiaga, Heng Ji, Hongbo Deng, Dong Wang, Hieu Khac Le, Tarek F. Abdelzaher, Jiawei Han, Alice Leung, John P. Hancock, and others. 2012. Tweet Ranking Based on Heterogeneous Networks. In *COLING*, pages 1239–1256.

David Inouye and Jugal K Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October.

Aditya Joshi, Balamurali A. R., Pushpak Bhattacharyya, and Rajat Kumar Mohanty. 2011. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of the 49th Annual Meeting of Association for Computational Linguistics (Demo)*. ACL.

Hao Li, Yu Chen, Heng Ji, Smaranda Muresan, and Dequan Zheng. 2012. Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 127–136, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Xiaohua Liu, Yitong Li, Furu Wei, and Ming Zhou. 2012. Graph-based multi-tweet summarization using social signals. In *COLING*, pages 1699–1714.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proc. ACL2014)*.

Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. 2013. Personalized time-aware tweets summarization. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*, pages 513–522. ACM.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP*.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012a. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA. ACM.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012b. Open domain event extraction from twitter. In *KDD*.

Burr Settles. 2009. Active learning literature survey.

Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 533–542. ACM.

Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. 2013. Automatic selection of social media responses to news. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 50–58. ACM.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161.

Zhongyu Wei and Wei Gao. 2014. Utilizing microblogs for automatic news highlights extraction. COLING.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277– 308, January.

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 255–264. ACM.

Arkaitz Zubiaga, Heng Ji, and Kevin Knight. 2013. Curating and contextualizing twitter stories to assist with social newsgathering. In *Proc. International Conference on Intelligent User Interfaces (IUI2013)*.

Arkaitz Zubiaga, Damiano Spina, Raquel Martnez, and Vctor Fresno. 2015. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473.