Exploiting News to Categorize Tweets: Quantifying The Impact of Different News Collections

Marco Pavan
University of Udine, Udine, Italy
marco.pavan@uniud.it

Stefano Mizzaro
University of Udine, Udine, Italy
mizzaro@uniud.it

Matteo Bernardon
University of Udine, Udine, Italy
matteo.bernardon@gmail.com

Ivan Scagnetto
University of Udine, Udine, Italy
ivan.scagnetto@uniud.it

Abstract

Short texts, due to their nature which makes them full of abbreviations and new coined acronyms, are not easy to classify. Text enrichment is emerging in the literature as a potentially useful tool. This paper is a part of a longer term research that aims at understanding the effectiveness of tweet enrichment by means of news, instead of the whole web as a knowledge source. Since the choice of a news collection may contribute to produce very different outcomes in the enrichment process, we compare the impact of three features of such collections: *volume*, *variety*, and *freshness*. We show that all three features have a significant impact on categorization accuracy.

1 Introduction

Social Network contents are analyzed for several purposes: identifying trends [MK10], categorizing and filtering news [JG13, SSTW14], measuring their importance, spread etc. [NGKA11]. Other researchers try to categorize short texts posted on social networks (e.g., tweets), using contents taken from the WWW, to understand user interests, to build user models etc. However, platforms like Twitter limit the text length, and users tend to use abbreviations and acronyms to write even faster. In a lot of cases the posted texts have a very low number of characters[1]; therefore, an automatic categorization process with topic extraction methodologies could be not enough reliable. In these cases, exploiting an additional source of information could help, providing additional text to analyze. Since short texts posted by users are often related to recent events (sharing their opinions and thoughts with friends), our approach is to use news collections instead of generic web contents in the categorization process.

On this basis, we study how the choice of the news collection affects the results: in particular, how different news collections with different properties impact the categorization effectiveness. More specifically, we analyze, by means of three experiments, three features of news collections: (i) *Volume*, to see how different numbers of news provide different sets of terms for the enrichment phase and, consequently, affect the categorizations; (ii) *Variety*, to see how news of different nature impact the enrichment process; and (iii) *Freshness*, to highlight the different effectiveness by using news from different time windows (i.e., same temporal context, 1 year old, 2 years old etc.). We exploit the methodology proposed in [MPSV14], based on a text enrichment with new set of words, extracted from news on webpages of the same temporal context,[2] and a categorization by querying the Wikipedia category tree as external knowledge base.

2 Related work

All the works in the literature addressing the problem of classifying tweets recognize that "data sparseness" and ambiguity represent a serious issue. For instance,

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

[1] Several surveys show that the mode of characters is 28 [twi16a].

[2] A set of news published in the same period of the short text.

in [HH15] the authors use the "bag-of-words" approach, adopting dimensionality reduction techniques, to reduce accuracy and performance problems.

In [AGHT11] the authors introduce several enrichment strategies (i.e., entity-based, topic-based, tweet-based and news-based) to relate tweets and news articles belonging to the same temporal context, in order to assign a semantic meaning to short messages. In [YPF10] another enrichment-based approach is proposed to classify generic online text documents, by adding a semantic context and structure, using Wikipedia as a knowledge source. In [GLJD13] the authors define a framework to enrich and relate Twitter feeds to other tweets and news speaking about the same topics. Hashtags (for tweets) and *named entities* (for news) are used to achieve such goal. A cluster-based representation enrichment method (CREST) is introduced in [DSL13]: such system enriches short texts by incorporating a vector of topical relevances (besides the commonly adopted tf-idf representation). Finally, topics are extracted using a hierarchical clustering algorithm with purity control. Enrichment techniques can also be quite sophisticated like, e.g., in [WZX+14] where a short texts are classified exploiting link analysis on topic-keyword graphs. In particular, after the initial topic modeling phase, each topic is associated to a set of related keywords. Afterwards, link analysis on a subsequent topic-keyword bipartite graph is carried out, to select the keywords most related to the analyzed short text.

Machine learning can play a fundamental role in classifying short texts: for instance, in [DDZC13] supervised SVM (Support Vector Machine) techniques are used to classify tweets into 12 predefined groups tailored for the online community of Sri Lanka. In [ZCH15] a completely automatized unsupervised bayesian model is used. In particular only tweets related to events are selected, exploiting a lexicon built from news articles published in the same period.

So far, it is clear that the problem of classifying short texts (whatever the related semantic domain) must rely on some forms of background knowledge, to fill the gaps and lack of information of the original messages. Such knowledge base can be found in external semantic platforms like, e.g., Wikipedia (as in some of the above mentioned works, and in the INEX Tweet Contextualization Track [ine13]), the WWW or other, possibly more focused, archives/structures. Hence, it is of utmost importance to study how the choice of the external collection influences the accuracy of the short text categorization process.

3 Features of News Collections

To run a set of experiments to analyze the collections features, we use two different open source document

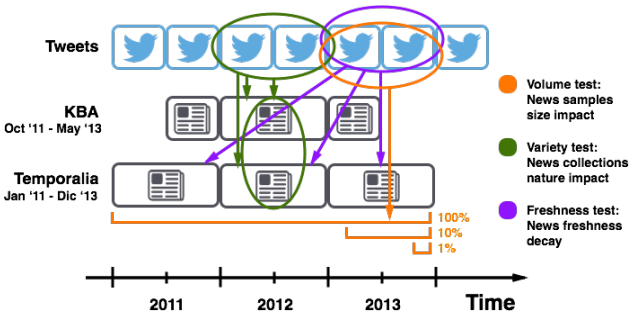


Figure 1: News collections distribution with features based tests

collections, which differ in number and kind of documents included, have different sizes, span from 2011 to 2013, and also have some temporal overlaps to allow several comparisons. They are shown in Table 1 and allow us to analyze the following three key features:

- *Volume:* we want to see the impact of news samples with different cardinality, extracted from the same collection in different percentages. With this test we aim to measure how the amount increment correlates to the final enrichment effectiveness.
- *Variety:* news are often different in nature, such as texts from blogs, forums, online newspapers etc., and different variety of texts could have different impact on the text enrichment. We want to measure how the news variety affects the results.
- *Freshness:* short texts are often related to recent events, therefore, it is interesting to study how important is to have the publishing time of the news close to the publishing time of the short text being enriched, and how the enrichment effectiveness changes using increasingly older news.

Figure 1 shows a representation of the two collections distributed over time and tweets as short texts to analyze. The *Volume test*, highlighted in orange, aims to compare the categorization results with samples of news from the same collection but with different sizes; the *Variety test*, in green, compares results among news samples with same cardinality but with different kinds of news; and the *Freshness test*, in purple, exploits news from the same collection but in different years. The figure shows only some examples; the details of all the experiments are described in the next section.

4 Experimental evaluation

4.1 Experimental design

To evaluate the impact of each news collection on the categorization process we selected a set of 5 popular Twitter account famous in different fields. In particular, David Cameron (@*David_Cameron*)

Table 1: The two news collections used in the experiments

Acronym	Name	# of docs/ size	kind of docs	Timespan
Temporalia	NTCIR Temporal Information Access 2012[a]	~2M / ~20GB	blogs news	Jan2011 – Dec2013
KBA	Knowledge Base Acceleration 2012[b]	~20M / ~930GB[c]	blogs, news, forums, social	Oct2011 – May2013

[a] `http://ntcirtemporalia.github.io/NTCIR-12/collection.html`

[b] `http://trec-kba.org/`

[c] Data extracted from the 3rd stream corpora `http://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html`

for Politics, Harry Kane (@*HKane*) for Sport, Bill Gates (@*BillGates*) for Technology, Neil Patrick Harris (@*ActuallyNPH*) for Cinema and Rihanna (@*rihanna*) for Music. We extracted a set of tweets from each account in a specific time window, according to the test we planned to run, in order to have a sufficient amount of short texts to enrich and categorize. We used a Python wrapper [pyt16] around the official Twitter API [twi16b] to retrieve tweets. We repeated this process to have a sample of 1000 tweets for each test which involves a large temporal window (e.g., six months or one year). Instead, for tests focused on one month, we built samples of 250 tweets. We then defined the benchmarks as follows in the next sections.

4.1.1 Volume test

To measure the impact of collections volume we defined 2 tests, "Test 1a" based on Temporalia and "Test 1b" on KBA. We analyzed samples using news subsets with different cardinality. With these tests we can see how changing the amount of news affects the results, and also if the results will generalize across different collections. The 2 tests are defined as follows:

Test 1a: Tweets posted in whole 2013, categorized with Temporalia 1%, Temporalia 10% and Temporalia 100%.

Test 1b: Tweets posted in whole 2013, categorized with KBA 1%, KBA 10% and KBA 100%.

4.1.2 Variety test

We defined "Test 2a" and "Test 2b" to measure how the variety of news inside a collection could impact the enrichment phase and consequently the categorization process. We selected news samples with the same cardinality from different collections and from different time windows, in order to see the effects of changing news varieties, and also if on a wider time window of 6 months we have the same effects we get on only 1 month. The 2 tests are defined as follows:

Test 2a: Tweets posted in January 2013, categorized with Temporalia Jan 2013 (60K news sample), KBA Jan 2013 (60K news sample) and Temporalia+KBA Jan 2013 (30K+30K news sample).

Test 2b: Tweets posted in the second half of 2012, categorized with Temporalia Jul-Dec 2012 (400K news sample), KBA Jul-Dec 2012 (400K news sample) and Temporalia+KBA Jul-Dec 2012 (200K+ 200K news sample).

4.1.3 Freshness test

To benchmark how the news freshness is important we defined 3 tests, "Test 3a", "Test 3b", based on different news "aging", and "Test 3c", based on a different collection. For the first test we want to see the difference between enriching the tweets with news extracted from the same temporal context (i.e., at most 1 month before the publishing date) and news in the same year of publishing (i.e., at most 1 year before the publishing date). In the second test we want to extend this analysis to more than 1 year before the publishing date, in particular we benchmark the results using news related to event of the same year of the tweets, 1 year old and 2 years old. The third test aims to compare the same "aging effect" with a different collection. The 3 tests are defined as follows:

Test 3a: Tweets posted in whole 2013, categorized with Temporalia 2013 - *contextualized*[3] and Temporalia Jan 2013 (both samples are composed of 60K news).

Test 3b: Tweets posted in whole 2013, categorized with Temporalia 2013, Temporalia 2012 and Temporalia 2011 (both samples are composed of 90K news).

Test 3c: Tweets posted in whole 2012, categorized with KBA 2012 - *contextualized*, KBA Jan 2012 and KBA 2012 (both samples are composed of 100K news).

4.2 Measures

To evaluate the experiments and to benchmark the collections effectiveness we carried out an expert evaluation to assess each analyzed feature over short texts samples composed of either all tweets for one month based tests (250) or a set of 250 randomly extracted tweets for tests based on larger temporal windows.

We used a categorization prototype system [MPSV14] for the categorization of short texts which

[3] Only news from the same month when the tweet has been posted.

provides, as final outcome, a list of labels extracted from Wikipedia category tree. The system includes a module which analyzes text, searches related documents into a news collection, and extracts a set of words used to enrich the original short text.

The texts have been submitted to the categorization system with different news collections according to the three tests described in Section 4.1. For each test, in order to assess the news impact over the enrichment process, the set of categories yielded by the system has been evaluated by expert users. The latter assigned a rating, i.e., a number between 1 and 5 (1=lowest value, 5=highest value) indicating how the categories properly represent the topic discussed in the tweet.

In particular for the Volume test, we run the evaluation several times, with news samples randomly rebuilt each time, where we used only a portion of the entire collection. We kept the average ratings obtained with different sub-collections, avoiding bias due to the random set of news. Specifically for samples with 10% or 1% of news we run respectively the evaluation 3 or 5 times, approximating the average ratings to the nearest integer value.

4.3 Results

Results are reported in the following charts, which show distribution functions of ratings obtained by each test with the different experiment settings. In particular, we display the cumulative distribution function (CDF), the inverted complementary cumulative distribution function (I-CCDF), and a table reporting the mean ratings. The I-CCDF is provided for an easier reading, showing the data in ascending order and thus highlighting the news collection performing better as the line at the top of the chart.

4.3.1 Volume Test

Figure 2 shows the results related to Test 1a and 1b, highlighting how for both collections the number of news is an important feature to consider. We can observe a noticeable improvement with Temporalia 100% compared to smaller samples. Increasing the volume allows us to include a large number of both relevant and not relevant news: the first ones yield a global improvement, while the second ones have a low overall impact. The general improvement is also confirmed by the Wilcoxon test. Then, we notice only a slight difference between Temporalia 1% and 10%, where the news increase in number from an order of magnitude 10K to 100K. The Wilcoxon test, over the latter couple of rating distributions, confirmed a non statistically significant difference between those samples, with a p-value>0.05. On the other hand, with KBA we already have a noticeable difference between KBA 1% and KBA 10%, due to order of magnitude from 100K to 1M, and even better using KBA 100% (10M). This fact emphasizes how increasing the sample sizes has considerable effects on the results only when a certain amount of news is reached. The diverse impact of Temporalia and KBA is probably also due to other factors than the only difference in size. Of course the same percentage, applied to collections with very different sizes, yields sets of extracted documents whose cardinality is very different; whence we can also expect a different variety of such sets. Moreover, for instance, KBA does not fully cover year 2013, whence the effectiveness could be affected by the publishing date of the analyzed short texts. Such aspects are taken into consideration in the remaining experiments.

4.3.2 Variety Test

Figure 3 shows how the variety of news inside the analyzed samples affects the enrichment effectiveness. Continuous lines represent the results over 1 month of news (Test 2a), and dotted lines over 6 months (Test 2b). For both experiments there is a noticeable difference among the samples which highlights how increasing the variety of news allows to improve the final categorization also on different time windows. The Wilcoxon test over the sample pairs of each test confirms the statistically significant difference between all the rating distributions. This fact highlights how important is to increase the variety of news in order to improve the set of words to use as text enrichment.

4.3.3 Freshness Test

The chart in Figure 4 shows the results related to Test 3a, 3b and 3c, and it is possible to notice how the news freshness affected the results especially when the news get older. Collections with contextualized news got the best effectiveness due to the news publishing time close to the tweets (same month), therefore they allow to have more relevant additional text to exploit. The system has worsened the categorization process with tweets randomly selected from whole 2013, and using collections of news extracted from the same year, either equally distributed over all months or only in January. The effectiveness decreases drastically when the news get older in previous years. In particular we can notice how we got the same lowest effectiveness with Temporalia 2012 and Temporalia 2011, highlighting how 1 (or more) year old news are poor of information for these purposes.

Test 3a results, related to Temporalia 2013, show how large is the difference between news distant only some months in time, and Test 3b results, where we analyzed three years of Temporalia news, highlight how going back to 1 year is crucial for the categorization

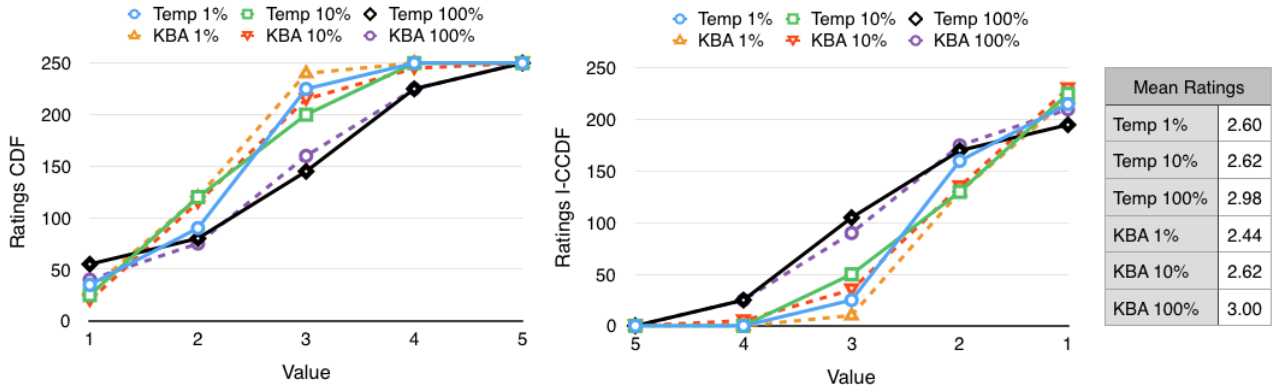


Figure 2: Volume impact CDF, I-CCDF, and mean ratings

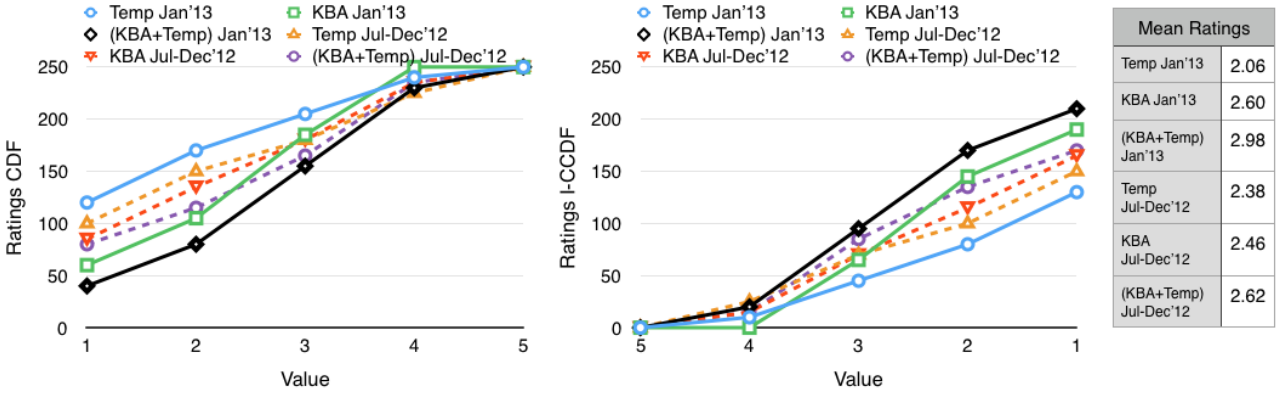


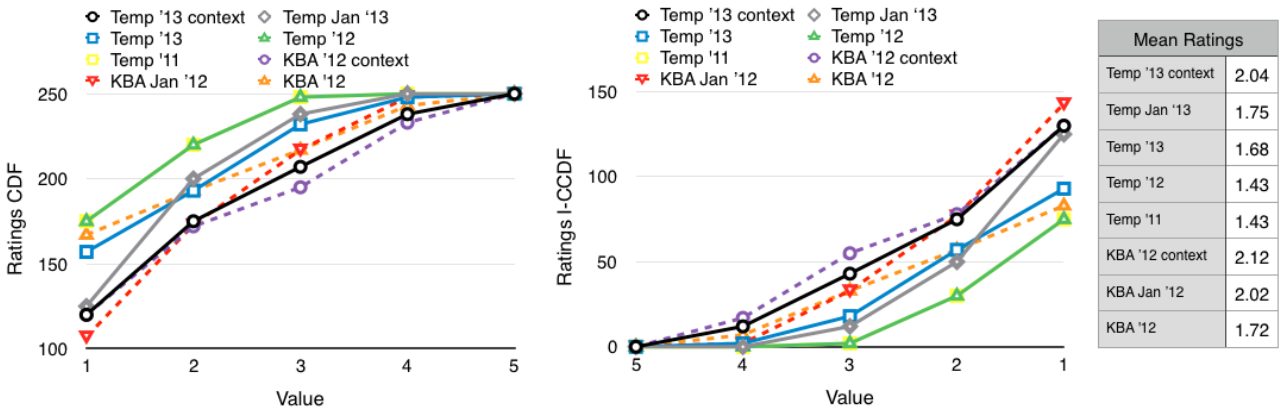
Figure 3: Variety impact CDF, I-CCDF, and mean ratings

process. With KBA collections we can notice how the results are similar and the rating distributions, represented by dotted lines, highlight better effectiveness with higher news freshness. Wilcoxon tests confirm that there is statistical significant difference among the rating distributions in both Temporalia and KBA, except for Temporalia '11 and '12 which obviously have equal values. This is a further confirmation that few months old news have a strong impact as those from previous years.

5 Discussion and Conclusions

The experiments performed in this work have demonstrated that text enrichment is sensibly affected by the features of the news collections that we have analyzed. More precisely, there is a critical threshold for what concerns the collection Volume, that allows to have a sufficient amount of news to reach a good level of effectiveness. Moreover, such threshold seems to be dependent on the whole size of the collection taken into consideration. Our benchmarks confirm the importance of news variety, highlighting how increasing the number of available kinds yields a better enrichment both for texts selected in one month and in the larger time window. The news Freshness appears to be a sensible feature since news published close to the same period of the short text provide a better set of terms to use in the enrichment phase. Indeed, as soon as the news begin to age (even of just a few months) the effectiveness of the categorization drastically decreases.

For future work, we plan to refine and complete the experiments on the three focused features. For instance, it could be interesting to look at the impact of the number of documents extracted from the news collection and used to categorize short texts. As we pointed out in Section 4.3, a larger database will produce a higher number of elements (with the same percentage), and this fact can have subtle implications on the final outcomes. We also plan to carry on further experiments about the variety, investigating which kinds of news it is important to include in the collection, and which ones are marginal. As the freshness is concerned, we could investigate more precisely, varying the granularity of the time windows, which is the temporal threshold causing a quick decrease of the effectiveness of the enrichment process. Moreover, we plan to carry on further experiments on



Mean Ratings	
Temp '13 context	2.04
Temp Jan '13	1.75
Temp '13	1.68
Temp '12	1.43
Temp '11	1.43
KBA '12 context	2.12
KBA Jan '12	2.02
KBA '12	1.72

Figure 4: Freshness impact CDF, I-CCDF, and mean ratings

different news collections and new kinds of short texts (e.g., instant chat messages, online comments). Unfortunately we could not use the Signal Media collection available at `http://research.signalmedia.co/newsir16/signal-dataset.html`; indeed, a collection covering a one-month period is not sufficient for the kind of experiments we described in this paper (think, e.g., of the freshness test).

References

[AGHT11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer, 2011.

[DDZC13] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using SVM. In *Proc. of ICCSE'13*, pages 287–291. IEEE, 2013.

[DSL13] Zichao Dai, Aixin Sun, and Xu-Ying Liu. Crest: Cluster-based representation enrichment for short text classification. In *Advances in Knowledge Discovery and Data Mining*, pages 256–267. Springer, 2013.

[GLJD13] Weiwei Guo, Hao Li, Heng Ji, and Mona T Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249, 2013.

[HH15] Yin-Fu Huang and Chen-Ting Huang. Mining domain information from social contents based on news categories. In *Proc. of IDEAS'15*, pages 186–191. ACM, 2015.

[ine13] INEX 2013 Tweet Contextualization Track. `http://inex.mmci.uni-saarland.de/tracks/qa/`, 2013.

[JG13] Nirmal Jonnalagedda and Susan Gauch. Personalized News Recommendation Using Twitter. In *Proc. of WI-IAT'13*, pages 21–25. IEEE Computer Society, 2013.

[MK10] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the Twitter stream. In *Proc. of ACM SIGMOD'10*, pages 1155–1158. ACM, 2010.

[MPSV14] S. Mizzaro, M. Pavan, I. Scagnetto, and M. Valenti. Short text categorization exploiting contextual enrichment and external knowledge. In *SIGIR '14 Proceedings*. SoMeRA, SIGIR, July 2014.

[NGKA11] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proc. of WebSci'11*, page 8. ACM, 2011.

[pyt16] Python wrapper around the Twitter API. `https://dev.twitter.com/rest/public`, 2016.

[SSTW14] Timm O Sprenger, Philipp G Sandner, Andranik Tumasjan, and Isabell M Welpe. News or noise? using twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8):791–830, 2014.

[twi16a] The Next Web. `http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/#gref`, 2016. [Online, visited Feb-2016].

[twi16b] Twitter REST APIs. `https://dev.twitter.com/rest/public`, 2016.

[WZX+14] Peng Wang, Heng Zhang, Bo Xu, Chenglin Liu, and Hongwei Hao. Short text feature enrichment using link analysis on topic-keyword graph. In *Natural Language Processing and Chinese Computing*, pages 79–90. Springer, 2014.

[YPF10] Hiroki Yamakawa, Jing Peng, and Anna Feldman. Semantic enrichment of text representation with wikipedia for text classification. In *Proc. of SMC'10*, pages 4333–4340. IEEE, 2010.

[ZCH15] Deyu Zhou, Liangyu Chen, and Yulan He. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Proc. of AAAI'15*, 2015.