TSix: A Human-involved-creation Dataset for Tweet Summarization

Minh-Tien Nguyen[1,2], **Dac Viet Lai**[1], **Huy-Tien Nguyen**[1], **Le-Minh Nguyen**[1]
[1]Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan.
{tiennm, vietld, ntienhuy, nguyenml}@jaist.ac.jp
[2] Hung Yen University of Technology and Education, Hung Yen, Vietnam.

Abstract

We present a new dataset for tweet summarization. The dataset includes six events collected from Twitter from October 10 to November 9, 2016. Our dataset features two prominent properties. Firstly, human-annotated gold-standard references allow to correctly evaluate extractive summarization methods. Secondly, tweets are assigned into sub-topics divided by consecutive days, which facilitate incremental tweet stream summarization methods. To reveal the potential usefulness of our dataset, we compare several well-known summarization methods. Experimental results indicate that among extractive approaches, hybrid term frequency – document term frequency obtains competitive results in term of ROUGE-scores. The analysis also shows that polarity is an implicit factor of tweets in our dataset, suggesting that it can be exploited as a component besides tweet content quality in the summarization process.

Keywords: Tweet summarization, hybrid TF-IDF, dataset, corpus, annotation.

1. Introduction

The growth of micro-blogging services such as Twitter encourages users sharing their viewpoints regarding an event. For example, users following `US Election` can post their tweets (short messages with a maximum of 140 characters) on their timelines. After posting, their friends can immediately update new information about this event. Those who are out of their networks can also track the event by using the keyword-search function provided by Twitter. However, search results are usually overwhelming due to millions of returned tweets, which span for weeks. Even if the filter is enabled, digging a large number of tweets for interesting contents would be a nightmare due to their noise. These demand a topic-driven system extracting high-quality tweets for user interests.

The bottleneck of social short-text summarization is the shortage of standard datasets for evaluating summarization methods whereas well-known DUC datasets have been freely published for document summarization. For tweet summarization, authors usually create their data. For example, although (Shou et al., 2013) released a dataset including events for evaluating their method, the dataset is now inaccessible. (Imran et al., 2014) published a dataset for disaster response during the Joplin tornado collected from Twitter. Although this dataset contains more than 230,000 tweets, its lack of references challenges the evaluation.

This paper leverages tweet summarization by introducing a new dataset including six events collected from Twitter. The involvement of humans in creating gold-standard references facilitates the evaluation. To show the potential usability of our dataset, we employ hybrid term frequency – inverse document frequency (TF-IDF) for extracting important tweets. Experimental results show that the hybrid model achieves competitive ROUGE scores over baselines. We also further analyze the polarity aspect of tweets. The analysis shows that extracted tweets tend to be non-sentiment. The dataset can be publicly accessible.[1]

[1]https://goo.gl/kXBof9

2. Summarization

This section presents our proposal in creating the dataset in two steps: data creation and the summarization model including tweet scoring and selection.

2.1. Data Creation

Data collection To create the dataset, we first defined a list of topics that satisfy following conditions. Firstly, trending topics are preferred in order to collect a large amount of data from various sources. Second, they are potential for last 30 days. Thirdly, they must be impressive to news providers. Once the list of trends was filed, we assigned each topic with a list of keywords. After that, those keywords were utilized to crawl data by tracking tweet streams using the public REST APIs[2] of Twitter.

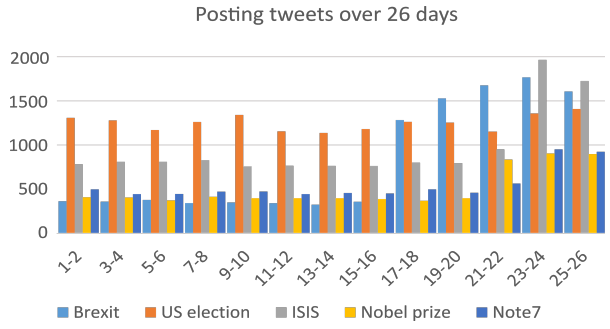


Figure 1: Tweets posted over 26 days after preprocessing.

Figure 1 plots the distribution of posting tweets over 26 days after pre-processing. Table 1 shows the statistics of the collected data and the used keywords. They can be used to estimate the quality of tweets by measuring how much important information is generated by social users in tweets.

Data segmentation Since the average number of tweets per day is not so large (around 600, Figure 1[3]), we set the period for a day. We collected tweets by their posting time to create a collection corresponding to each time

[2]https://dev.twitter.com/rest/public
[3]We plot five events due to space limitation.

Table 1: Six events, collected from Oct 10 to Nov 9, 2016.

Event	#tweets	#hashtags	Keywords
Brexit	10,978	9,705	brexit, #brexitshambles, #Brexiters, #BrexitCentral, England Europe exit.
US election	17,714	8,566	Donald Trump, Trump, Hilary Clinton, Clinton #debate, election.
ISIS	13,047	9488	ISIS, IS Syria, IS Mosul, IS Iraq, ISIS Aleppo, ISIS US, ISIS Rusia.
SS Note 7	7,362	7,465	Galaxy Note 7, #note7, #GalaxyNote7, thegalaxynote7, #SamsungGalaxyNote7.
Nobel prize	6,812	2,780	Nobel prize 2016, Nobel peace, Nobel chemistry, Nobel economy, Nobel physics.
SpaceX	4,982	2,417	Facebook SpaceX, SpaceX Explosion, Falcon 9 exploded, Falcon 9 explosion.

step. Tweets in each time step were assigned into clusters. The intuition of clustering tweets is that even the number of tweets per day is small, directly extracting a subset of these tweets may eliminate other important ones. By clustering, our goal is to keep representative tweets as many as possible. The time step can be arbitrary, e.g. per hour.
To foster the real-time aspect of a tweet summarization system, we adopted the Affinity Propagation (AP)[4] algorithm for clustering (Frey and Dueck, 2007) because traditional clustering methods such as k-means (MacQueen, 1967; Forgy, 1965) require a pre-defined number of clusters k. However, in real-time scenarios, identifying k is nontrivial (Busch et al., 2012; Nguyen et al., 2015). AP identifies a subset of data points as exemplars and forms clusters by assigning remaining data points into one of the exemplars. After clustering and eliminating days which contain a very small number of tweets (less than 10), we formed six sets corresponding to six events in 26 days and each day includes a set of clusters, which can be seen as subtopics.

Standard reference creation Once clusters have been formed, we have to create standard references for evaluation. We followed the two-stage method (Shou et al., 2013) in order to avoid tremendous human labor. In the extraction stage, since the number of tweets in each cluster is quite large, we applied three different extractive methods to create reference candidates. `Luhn` is a heuristic method for extraction (Luhn, 1958). `Lexrank` is a graph-based method, which builds a sentence similarity graph and selects important ones based on their eigenvector centrality (Erkan and Radev, 2004). `DSDR-non` bases on non-negative linear data reconstruction (He et al., 2012). The extracted tweets from the three methods form three candidate sets. In practice, suppose that n is the number of tweets in each cluster, we conditioned the number of extracted tweets is $n_{ext} = \frac{n}{2}$ if $n \leq 30$; otherwise $n_{ext} = \frac{n}{3}$.
In the selection step, we asked two annotators to select references from the candidate ones via a Web interface.[5] Each annotator reads whole candidate references in each cluster (after extracting) and estimates the importance of each candidate reference. A gold-standard reference is a tweet, which satisfies two conditions: (i) it is important in the viewpoint of each annotator regarding the event and (ii) it belongs to at least two over three candidate sets. Each cluster contains more than five and less than 25 tweets. Since the judgment of annotators is objective, therefore we kept the selected tweets from the two annotators as the references. As a result, the outputs of each extractive method have to compare to two references.[6] We show the upper bounds of ROUGE scores by using extracted tweets from the three methods in Table 2.

Table 2: Upper bound ROUGE scores.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
Luhn	0.556	0.502	0.250
LexRank	0.581	0.516	0.273
DSDR-non	0.419	0.310	0.145

Data observation Unlike other types of data, tweets are created to share human thoughts and emotions under a length constraint. Consequently, people tend to put emotional words and hashtags in their tweets. The emotional words show their interests and the hashtags include important information regarding an event. We performed an observation of sentiment words and hashtag utilization as follow. We projected each tweet to a dictionary[7] of sentiment words to assess whether this tweet contains the sentiment aspect. Meanwhile, we counted the number of tweets consisting of hashtags. Figure 2 shows the observation.

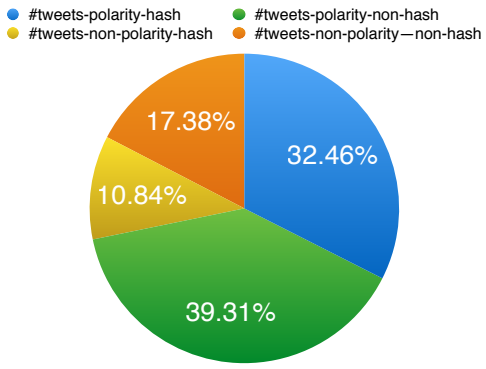


Figure 2: Hashtag and polarity observation on six datasets.

In Figure 2, the number of tweets containing polarity is considerable (around 72%). It shows that polarity analysis may potentially affect the extraction step. The number of tweets which owns the polarity aspect and includes hashtags is large (32.46%) whereas only 17.38% of tweets do not contain polarity and hashtags. Note that the number of tweets containing polarity may change if we use a classifier instead of using a dictionary.

[4]http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

[5]https://s242-097.jaist.ac.jp/doc-sum-annotator/annotate

[6]This is similar to DUC, which includes four references from four annotators for each topic.

[7]http://mpqa.cs.pitt.edu/

2.2. Summarization with Hybrid TF-IDF

Tweet scoring Term frequency – inverse document frequency (TF-IDF) (Luhn, 1958) is a well-known method for information retrieval and text summarization.

$$TF\_IDF = tf_{i,j} * \log_2 \frac{N}{df_j} \quad (1)$$

where: $tf_{i,j}$ is the frequency of term T_j in the document D_i, N is the total number of documents, and df_j is the number of documents containing the term T_j. After scoring, sentences containing terms with high weights are extracted as a summary.

Equation (1) composes TF and IDF with the logarithm to balance the effect of the IDF component. In the context of document summarization, TF-IDF has been shown advantages to select important sentences (Luhn, 1958). For tweet extraction, since tweets are informal documents; therefore, Eq. (1) exists two issues. Firstly, if we consider all tweets in a cluster as a document, Eq. (1) can compute the TF across all tweets. However, the IDF is limited due to only one document. On the other hand, we can assume that each tweet as a document to tackle the limitation of IDF. However, the TF is problematic because each tweet consists of a handful of words, hence, it receives a small TF value.

From the limitations of the traditional TF-IDF for tweet extraction, we, therefore, adopted a hybrid TF-IDF method (Inouye and Kalita, 2011). It differs the traditional one by regarding all tweets as a single document when computing TF and each tweet as a separate document when calculating IDF. Eqs. (2) - (6) present the hybrid TF-IDF model.

$$h_{TFIDF}(t) = \frac{\sum_{i=0}^{\#WordsInTweet} W(w_i)}{nf(t)} \quad (2)$$

$$W(w_i) = tf(w_i) * \log_2(idf(w_i)) \quad (3)$$

$$tf(w_i) = \frac{\#OccurencesOfWordInAllTweets}{\#WordsInAllTweets} \quad (4)$$

$$idf(w_i) = \frac{\#Tweets}{\#TweetsInWhichWordOccurs} \quad (5)$$

$$nf(t) = \#WordsInTweet \quad (6)$$

where: w_i is a term i^{th} in the tweet t, $W()$ returns the weight of a term, $tf()$ returns the TF score, $idf()$ is the IDF score, $nf()$ is a normalization factor for the tweet t because the traditional TF-IDF model usually biases to select longer tweets. We used stemming (Porter, 2011) in `NLTK` (Bird et al., 2009) for non-stopwords because stop words contribute insignificantly in the scoring step.

Tweet selection After scoring tweets in each subtopic by using Eq. (2), top *m* ranked tweets having the highest scores were selected as a summary for each cluster.

3. Experimental Setup

Settings In the preprocessing step, URLs were removed to reduce the noise. Standard Cosine similarity (threshold = 0.85) was also used to remove duplicate tweets (those have a very similar content). After being removed, clusters having more than 10 tweets were considered for summarization. The output of summarization methods was fix by $m = 15$ if $n > 15$; otherwise $m = \frac{n}{2}$; where n is the number of tweets in each cluster after preprocessing and removing duplicate ones.

Baselines We compared the hybrid method to basic models, which have been widely used for extractive summarization. **KL (Kullback-Leibler) Divergence** measures the difference of unigram probability distributions learned from seen documents (original documents) and unseen documents (summaries) based on KL-Divergence (Sripada and Jagarlamudi, 2009). **LSA** uses latent semantic analysis with the usage of SVD to rank tweets (Gong and Liu, 2001). **Sumbasic** bases on the impact of frequency on various aspects of summarization (Nenkova and Lucy, 2005). **TextRank** utilizes a graph-based ranking algorithm (Mihalcea and Tarau, 2004) for phrases and sentence extraction. **Retweet** represents the importance of a tweet based on retweet (Busch et al., 2012). **DSDR-linear** bases on data reconstruction with linear combination (He et al., 2012).

Evaluation method The evaluation was conducted on each cluster, by matching extracted tweets with the references. We employed `ROUGE-1.5.5` (Lin and Hovy, 2003) by using `pyrouge`[8] with ROUGE-1, 2, and SU F-score to balance precision and recall.

4. Results and Discussion

4.1. Experimental Results

ROUGE scores with gold-standard references We report the summarization performance of the hybrid model on our dataset with the average of ROUGE scores in 26 days compared to the baselines.

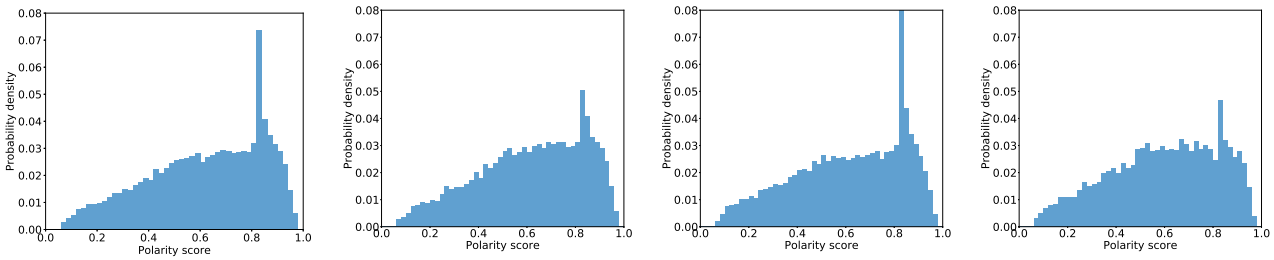
Table 3: The average ROUGE scores over six datasets. Text means the hybrid model significantly outperforms with $p \leq 0.05$. **Bold** is the best, *italic* is second bet.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
KL	0.394	0.263	0.146
LSA	0.462	0.368	0.175
Sumbasic	0.444	0.298	0.174
TextRank	**0.495**	**0.418**	**0.213**
Retweet	0.384	0.264	0.129
DSDR-lin	0.460	0.351	0.183
h-TFIDF	*0.482*	*0.384*	*0.199*

The ROUGE scores indicate that the hybrid model obtains very competitive results, where it significantly outperforms almost methods (using the pair $t-$test[9]), except for TextRank. It confirms the efficiency of the model in summarizing short texts (Inouye and Kalita, 2011). However, a large margin between the ROUGE scores of the hybrid model and the upper bounds (Table 2) suggest that its performance can be improved. TextRank is the best model for all metrics

[8] parameters: -c 95 -2 -1 -U -r 1000 -n 4 -w 1.2 -a -m

[9] https://docs.scipy.org/doc/scipy-0.19.0/reference/generated/scipy.stats.ttest_ind.html



(a) All data (b) H-TFIDF (c) KL (d) DSDR-non

Figure 3: Polarity distribution of original and extracted tweets from three models.

because it is similar to LexRank, which achieves the highest upper bound of ROUGE scores in Table 2. Small margins between the hybrid method and TextRank indicate that we can still increase the performance of the hybrid model. Methods based on the content quality analysis (except for KL and retweet) are competitive, confirming that content quality is a critical factor for tweet selection (Inouye and Kalita, 2011; Duan et al., 2012; Shou et al., 2013; Nguyen et al., 2015).

ROUGE scores with hashtags We also evaluated all the methods by using hashtags. The intuition is that tweets usually include hashtags, which show important information regarding user's interests. To do that, we extracted all hashtags of each cluster to form its artificial references.

Table 4: The average ROUGE scores over six datasets.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
KL	0.122	0.033	0.015
LSA	0.111	*0.034*	*0.017*
Sumbasic	**0.137**	*0.034*	*0.018*
TextRank	0.100	0.030	0.011
Retweet	0.101	0.024	0.007
DSDR-lin	0.117	0.033	0.011
DSDR-non	*0.123*	**0.036**	0.010
Luhn	0.105	0.032	0.011
LexRank	0.118	0.033	0.013
h-TFIDF	0.113	0.031	0.010

ROUGE scores from Table 4 indicate that the hybrid model is still better than some methods, but results are slightly worse than those in Table 3. The margin among these models is small because hashtags are short and single words.

4.2. Polarity Observation

We argue that tweets usually include users' opinions defined as the polarity aspect (Turney, 2002; Pang et al., 2002; Liu, June 2015). Figure 2 also supports our argument. To reveal this aspect, we trained a polarity classifier to predict whether an input tweet contains polarity (sentiment/non-sentiment). We adapted Semeval datasets[10] because of the unfortunate lack of this kind of dataset for our task. From Semeval 2013 to 2016, we obtained 22,591 neutral tweets, 19,903 positive, and 7,840 negative tweets. We randomly selected 15,680 neutral ones as non-sentiment tweets, and 7,840 positive and 7,840 negative tweets to form sentiment tweets. We employed convolutional neural networks (CNN) (Kim, 2014; Kalchbrenner et al., 2014; Zhang and Wallace, 2015) for training as the setting of (Kim, 2014). The number of each region size is 300 and the dimension of a penultimate NN layer (with Dropout rate $p = 0.5$) is 100. We finally applied the trained model to our data. After predicting, a score of each tweet was converted to polarity intensity in [0, 1], where tweets with high scores (close to 1) are non-sentiment whereas those close to 0 are sentiment. Figure 3a shows that many tweets distribute in [0.4, 1]. The number of non-sentiment tweets (in [0.5, 1]) is larger than that of sentiment ones (in [0.5, 1]).

We also observed extracted tweets from three methods: hybrid TF-IDF, KL, and DSDR-non to investigate polarity. The distributions in Figures 3b, 3c and 3d are quite similar to Figure 3a, where tweets selected by the three models mainly range in [0.5, 1]. For example, the density of extracted tweets from DSDR-non, one of the competitive models, mainly distributes in [0.5, 1], showing that salient tweets tend to be non-sentiment. The same patterns appear in the result of hybrid TF-IDF and KL. The distribution in Figure 3 suggests a deeper analysis in combining polarity and content quality in the summarization process.

5. Conclusion

In this paper, we present a new dataset for tweet summarization. It includes six events collected from October 10 to November 9, 2019. The major property of our dataset is human involvement in creating gold-standard references, which provide reliability to evaluate extractive methods. Tweets also are assigned in sub-topics in consecutive days, which facilitate continuous tweet stream summarization. Experimental results conclude that the hybrid TF-IDF model obtains very competitive ROUGE scores.

We encourage to validate other advanced methods on our dataset. The preliminary analysis of polarity reveals the fact that tweets usually include users' opinions. It motivates a possible direction to exploit the polarity of tweets to improve the scoring step.

6. Acknowledgments

This work was supported by JSPS KAKENHI Grant number 15K16048, Japan; and Center for Research and Applications in Science and Technology, Hung Yen University of Technology and Education, under the grant number UTEHY.T026.P1718.04.

[10] http://alt.qcri.org/semeval2016/task4/

7. Bibliographical References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., and Lin, J. J. (2012). Earlybird: Real-time search at twitter. In *Proceedings of 28th International Conference on Data Engineering, pp. 1360-1369. IEEE*.

Duan, Y., Chen, Z., Wei, F., Zhou, M., and Shum, H.-Y. (2012). Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics, pp. 763-780. Association for Computational Linguistics*.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22, pp. 457-479*.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics, 21: 768?769*.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science 315(5814), pp. 972-976*.

Gong, Y. and Liu, X. (2001). Generic text summarization using relevant measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 19-25. ACM*.

He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., and He, X. (2012). Document summarization based on data reconstruction. In *AAAI, pp. 620-626*.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *WWW (Companion Volume): 159-162*.

Inouye, D. I. and Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of Third Inernational Conference on Social Computing and Privacy, Security, Risk and Trust, pp. 298-306. IEEE*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 655-665. Association for Computational Linguistics*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 1746-1751. Association for Computational Linguistics*.

Lin, C.-Y. and Hovy, E. H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 71-78. Association for Computational Linguistics*.

Liu, B. (June 2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development, 2(2), pp. 159-165*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297*.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 404-411. Association for Computational Linguistics*.

Nenkova, A. and Lucy, V. (2005). The impact of frequency on summarization. Technical report, Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005 101.

Nguyen, M.-T., Kitamoto, A., and Nguyen, T.-T. (2015). Tsum4act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 64-75. Springer International Publishing*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics*.

Porter, M. F. (2011). Snowball: A language for stemming algorithms.

Shou, L., Wang, Z., Chen, K., and Chen, G. (2013). Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 533-542. ACM*.

Sripada, S. and Jagarlamudi, J. (2009). Summarization approaches based on document probability distributions. In *Proceedings of The 23rd Pacific Asia Conference on Language, Information and Computation, pp. 521-529*.

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424. Association for Computational Linguistics*.

Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.