Identification of Good and Bad News on Twitter

Piush Aggarwal
University of Duisburg-Essen
piush.aggarwal@stud.uni-due.de

Ahmet Aker
University of Duisburg-Essen
a.aker@is.inf.uni-due.de

Abstract

Social media plays a great role in news dissemination which includes good and bad news. However, studies show that news, in general, has a significant impact on our mental stature and that this influence is more in bad news. An ideal situation would be that we have a tool that can help to filter out the type of news we do not want to consume. In this paper, we provide the basis for such a tool. In our work, we focus on Twitter. We release a manually annotated dataset containing 6,853 tweets from 5 different topical categories. Each tweet is annotated with good and bad labels. We also investigate various machine learning systems and features and evaluate their performance on the newly generated dataset. We also perform a comparative analysis with sentiments showing that sentiment alone is not enough to distinguish between good and bad news.

1 Introduction

Social media sites like Twitter, Facebook, Reddit, etc. have become a major source of information seeking. They provide chances to users to shout to the world in search of vanity, attention or just shameless self-promotion. There is a lot of personal discussions but at the same time, there is a base of useful knowledgeable content which is worthy enough to consider for the public interest. For example in Twitter, tweets may report about news related to recent events such as natural or man-made disasters, discoveries made, local or global election outcomes, health reports, financial updates, etc. In all cases, there are good and bad news scenarios.

Studies show that news, in general, has a significant impact on our mental stature (Johnston and Davey, 1997). However, it is also demonstrated that the influence of bad news is more significant than good news (Soroka, 2006; Baumeister et al., 2001) and that due to the natural negativity bias, as described by (Rozin and Royzman, 2001), humans may end up consuming more bad than good news. Since bad news travels faster than good news (Kamins et al., 1997; Hansen et al., 2011) the consumption may increase. This is a real threat to the society as according to medical doctors and, psychologists exposure to bad news may have severe and long-lasting negative effects for our well being and lead to stress, anxiety, and depression (Johnston and Davey, 1997). (Milgrom, 1981; BRAUN et al., 1995; Conrad et al., 2002; Soroka, 2006) describe crucial role of good and bad news on financial markets. For instance, bad news about unemployment is likely to affect stock markets and in turn, the overall economy (Boyd et al., 2005). Differentiating between good and bad news may help readers to combat this issue and a system that filters news based on the content may enable them to control the amount of bad news they are consuming.

The aim of this paper is to provide the basis to develop such a filtering system to help readers in their selection process. We focus on Twitter and aim to develop such a filtering system for tweets. On this respect the contributions of this work are:

- We introduce a new task, namely the distinction between good and bad news on Twitter.
- We provide the community with a new gold standard dataset containing 6,893 tweets. Each tweet is labeled either as good or bad. To the best of our knowledge, this is the first dataset containing tweets with good and bad

Proceedings of Recent Advances in Natural Language Processing, pages 9–17,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_002

labels. The dataset is publicly accessible and can be used for further research[1].

- Provide guidelines to annotate good/bad news on Twitter.
- We implement several features approaches and report their performances.
- The dataset covers diverse domains. We also show out-of-domain experiments and report system performances when they are trained on in-domain and tested on out-of-domain data.

In the following, we first discuss related work. In Section 3 we discuss the guidelines that we use to annotate tweets and gather our dataset. Section 4 provides description about the data itself. In Section 5 we describe several baseline systems performing the good and bad news classification as well as features used to guide the systems. Finally, we conclude the paper in Section 7.

2 Related Work

In terms of classifying tweets into the good and bad classes no prior work exists. The closest studies to our work, are those performing sentiment classification in Twitter (Nakov et al., 2016; Rosenthal et al., 2017). Kouloumpis et al. (2011) use n-gram, lexicon, part of speech and micro-blogging features for detecting sentiment in tweets. Similar features are used by Go (2009). More recently researchers also investigated deep learning strategies to tackle the tweet level sentiment problem (Severyn and Moschitti, 2015; Ren et al., 2016). Twitter is multi-lingual and in Mozetič et al. (2016) the idea of multi-lingual sentiment classification is investigated. The task, as well as approaches proposed for determining tweet level sentiment, are nicely summarized in the survey paper of Kharde et al. (2016). However, Balahur et al. (2010) reports that there is no link between good and bad news with positive and negative sentiment respectively.

Thus, unlike related work, we do tweet level good vs. bad news classification. We also show that similar to Balahur et al. (2010), there is no evidence that positive sentiment implies good news and negative sentiment bad news.

[1] https://github.com/aggarwalpiush/goodBadNewsTweet

Figure 1: Good news tweet

Figure 2: Bad news tweet

3 Good vs Bad News

News can be good for one section of society but bad for other section. For example, win or loss related news are always subjective. In such cases, agreement towards news types (good or bad) is quite low. On the other hand, news related to natural disaster, geographical changes, humanity, women empowerment, etc. show very high agreement. Therefore, while defining news types, topicality plays an important role.

We consider news as good news if it relates to low subjective topics and includes positive overtones such as recoveries, breakthroughs, cures, wins, and celebrations (Harcup and ONeill, 2017) and also beneficial for an individual, a group or society. An example of good news is shown in Figure 1. In contrary to that, the bad news is defined as when it relates to the low subjective topic and include negative overtones such as death, injury, defeat, loss and is not beneficial for an individual, a group or society. An example of bad news is shown in Figure 2. Based on these definitions/guidelines we have gathered our dataset (see next Section) of tweets containing the good and bad labels.

4 Dataset

Data collection To collect tweets for annotation, we first choose low subjective ten topics which can be divided into five different categories. Then,

Category	Topics	Collected	Annotated
Health	Ebola	892	852
	Hiv	663	630
Natural Disaster	Hurricane Harvey	2,073	1,997
	Hurricane Irma	795	772
Terrorist Attack	Macerata oohmm	668	625
	Stockholm Attack	743	697
Geography and Env.	AGU17	652	592
	Swachh Bharat	21	21
Science and Edu.	IOT	627	602
	Nintendo	78	65
	Total	**7,212**	**6,853**

Table 1: Categories, their topics, and distributions for the dataset generation.

we retrieve the examples from Twitter using its API[2]. Next, we discard non-English tweets and re-tweets. We also remove duplicates based on lower-cased first four words of tweets keeping only the first one. Thereafter, we filter only those tweets which can be regarded as news by using an in house SVM classifier (Aggarwal, 2019). This classifier is trained on tweets annotated with the labels news and not news. We use this classifier to remove *not news* tweets from the annotation task[3]. We select only tweets where the classifier prediction probability is greater than or equal to 80%. In Table 1, we provide information about the topics and categories as well as statistics about the collected tweets that will be used for annotation (column *collected*).

Data Annotation For data annotation, we use the figure-eight crowdsourcing service[4]. Before uploading our collected examples, we carried out a round of trial annotation of 300 randomly selected instances from our tweet collection corpus. The aim of the trial annotation was

- to ensure the newsworthiness quality of our collected examples.
- to create test questions to ensure the quality of the annotators, for the rest of the data, which was carried out using crowdsourcing.
- to test our guidelines described in Section 3.

We ask three annotators[5] to classify the selected examples into good and bad news. We also allowed a third category *cannot say*. We computed Fleiss' kappa Fleiss (1971) on the trial dataset for the three annotators. The value is 0.605 which indicates rather a high agreement. We used 247 instances agreed by all the three annotators as test questions for the crowdsourcing platform.

During the crowd annotation, we showed each annotator 5 tweets per page and paid 3 US Cents per tweet. For maintaining quality standards, in addition to the test questions, we applied a restriction so that annotation could be performed only by people from English speaking countries. We also made sure that each annotation was performed maximum by 7 annotators and that an annotator agreement of min. 70% was met. Note if the agreement of 70% was met with fewer annotators then the system would not force an annotation to be done by 7 annotators but would finish earlier. The system requires 7 annotators if the minimum agreement requirement is not met. We only choose instances that are annotated by atleast 3 annotators. In addition to the good and bad news categories we also ask annotators to mandatory provide their confidence score (range between 0-100%) for the label they have annotated[6]. We discarded all the tweets where we did not have at least 3 annotators with each having min. 50% confidence value. We also discarded tweets that are annotated by less than three annotators. We

[2] https://www.tweepy.org

[3] Since we want humans to annotate tweets as good and bad news we apply this approach to filter tweets that are not news at all and so avoid our annotators spending valuable time on annotating tweets that are not our target.

[4] https://www.figure-eight.com/

[5] All are post-graduate students who are fluent in English and use Twitter to post information on a daily basis.

[6] We found this strategy better than providing the option *cannot say* and later allowed us to discard annotations where the confidence score was less than 50%.

use a total 7,212 tweets to annotate. After all filterings, we remained with 6,853 instances which were classified as good and bad news. Topic-wise number of successful annotations are displayed in the fourth column of Table 1.

Inter Annotator Agreement To calculate agreement between the annotators of the crowdsourcing annotation results, we select the top three confident annotator labels for each sample. Based on this, we record an agreement of 0.614 as Fleiss' Kappa (Fleiss, 1971) score indicating a good agreement among the annotators. We also claim stability in our annotation task because of the score similarity with that of trail annotation.

5 Method

We experiment with several machine learning approaches and features. Before using the tweets in decision making, we also apply a simple preprocessing on them. In the following, we briefly outline these.

5.1 Preprocessing

We use the ArkTokenizer (Gimpel et al., 2011) to tokenize the tweets. In addition to tokenization, we do lowercasing and remove digits if available in text.

5.2 Features

We extract nine features for each tweet and divide them into *Structural*, *TF-IDF* and *Embeddings* features.

5.2.1 Structural features

Emoticons: We extract all the emoticons from the training data and use them as a binary feature, i.e. does a tweet contain a particular emoticon or not.

Interjections: We use existing list of interjections[7] and use them similar to *Emoticons* as binary feature.

Lexicons: We use existing positive and negative lexicons[8] and use them as a binary feature.

Sentiment: We use the textblob[9] tool to compute sentiment score over each tweet. The score varies between -1 (negative) to 1 (positive).

POS-Tag: This feature includes 36 different pos-tags (uni-gram) and are used as binary features.

Significant terms: Using tf-idf values we also extract the top 300 terms (uni-gram and bi-gram, 300 in each case) from the training data and use them as binary features. Note, we extract for good and bad news separate uni-grams and bi-grams.

Tweet Characteristics: This feature contains tweet specific *characterstics* such as the number of favorite counts, tweet replies count and number of re-tweets.

5.2.2 TF-IDF

In this case, we simply use the training data to create a vocabulary of terms and use this vocabulary to extract features from each tweet. We use tf-idf representation for each vocabulary term.

5.2.3 Embeddings

Finally, we also use fasttext based embedding (Mikolov et al., 2018) vectors which are trained on common crawl having 600 billion tokens.

5.3 Classifiers

We investigate 8 classifiers for our task including Multi-Layer Perceptron (MLPC), Support Vector Machine with linear (LSVC) and rbf (SVC) kernel, K Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and Decision Tree (DT). In addition, we also fine-tune BERT-base model (Devlin et al., 2018). Each classifier, except the BERT, has been trained and tested on each possible combination of the three feature types.

6 Results

Overall results We performed a stratified 5-fold cross-validation. We evaluate each resulting model on a held-back development dataset containing 264 good news postings and 764 bad news ones. The 5-fold cross validation has been performed on the training data containing 4,332 bad news and 1,493 good news instances. For

[7] https://www.vidarholen.net/contents/interjections/

[8] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[9] https://textblob.readthedocs.io/en/dev/

Feature set	SVC	XGB	LSVC	KNN	RF	MLPC	DT	LR
Structural	.78	.78	.77	.77	.77	.63	.74	.78
Embeddings	.88	.86	.87	.86	.85	.85	.72	.87
TF-IDF	.86	.85	.86	.83	.84	.84	.83	.87
Structural + Embeddings	.86	.85	.87	.79	.86	.86	.78	.87
Structural + TF-IDF	.87	.87	.87	.80	.87	.86	.81	.87
Embeddings + TF-IDF	**.89**	.87	**.89**	.87	.88	.87	.81	**.89**
ALL	.88	.88	**.89**	.82	.87	.86	.82	.88

BERT-base model with its pre-trained embedding features: .92

Table 2: F_1(macro) scores of different classifiers on different feature types evaluated on the test data. Multi-Layer Perceptron (MLPC), Support Vector Machine with linear (LSVC) and rbf (SVC) kernel, K Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and Decision Tree (DT).

each model, we use grid-search method to select the hyper-parameters with best model's efficiency. The results reported are those obtained on the test data and are summarized in Table 2. Overall we see that the performances of the classifiers are all highly satisfactory. Among the more traditional approaches, the best performance is obtained through SVC, LSVC, and LR. We see also that these approaches work best when embeddings along with tf-idf features are used, although LSVC achieves the same results when all features are used. However, the best performance is achieved with the BERT-base model leading to 92% F_1 score. We computed also significance test using paired t-test between BERT and more traditional machine learning approaches[10]. However, after Bonferroni correction ($p < 0.007$) we found no significant difference between BERT and the other systems.

Structural feature analysis We also evaluate the structural features of the task independently (Figure 3). For this, we use the SVC classifier as it is one of the best performing traditional methods. From the figure, we see that the significant term feature gives the best performance. The difference to the other features is greater than the 23% F_1 score. The differences are also significant after Bonferroni correction ($p < 0.008$). In Table 3 we list some frequent uni-grams from the significant good and bad term lists. From the table, we see that the terms are certainly good indicators for distinguishing between the two classes.

Bad news	Good news
fake	services
racism	cured
fox	resistant
attack	energy
migrants	support
fears	arrested

Table 3: Top uni-grams from the good and bad news significant term lists.

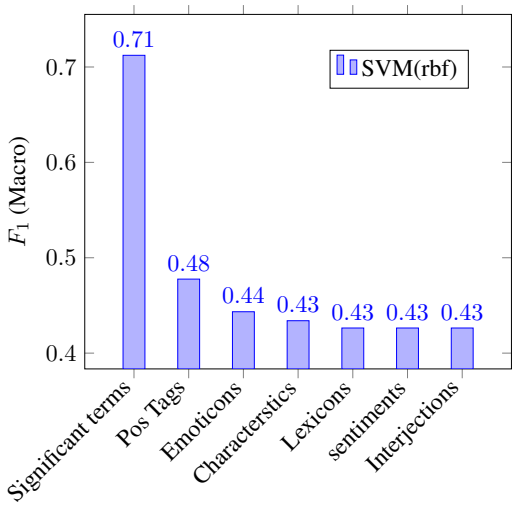

Figure 3: Structural features' performance using the SVM classifier evaluated on the test set.

[10]We always use the best result for every system.

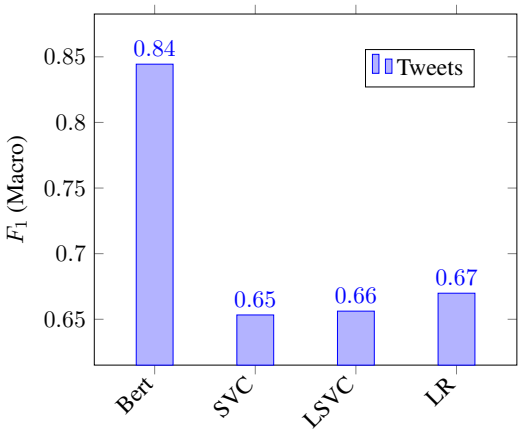

Figure 4: Out-of-domain performance of different systems.

Sentiment for good-vs-bad news We also tested whether sentiment score can predict good vs. bad news as Naveed et al. (2011) found a relationship between these two. For this, we use the textblob sentiment scorer and classify any tweet as good news when its sentiment score is greater than 0 otherwise bad. Using this strategy we could only achieve an F_1 score of 55%. This shows that tackling the good/bad news classification task using sentiment scores is not appropriate. This also confirms the findings of Balahur et al. (2010).

Out-of-domain experiments We also investigate how stable the models are when they are trained on in-domains and tested on out-of-domain data. For this purpose, we split our dataset into a training set consisting of all examples except instances belonging to the health category. We use four of the best-performing systems (BERT, SVC, LSVC, and LR) to train on this training set. The resulting models are tested on the held-out health data. Results are shown in Figure 4. From Figure 4 we see that BERT is stable and achieve an F_1 score of 84%. The performance of the other system drop by a great margin to the max. 67% F_1 score. From this, we can conclude that BERT is a better system to use for good-vs-bad Twitter news classification.

Detailed analysis on BERT Our overall but also the out-of-domain experiments show that BERT is outperforming the more traditional machine learning approaches. On the overall (1,028 testing instances) results, BERT fails only to classify 63 cases correctly. Using t-SNE distribution (van der Maaten and Hinton, 2008), we analyse BERT's 12th layer embedding vectors (having 300 dimensions) for random 100 test points (Figure 5). The analysis shows that BERT can classify semantics of good and bad news instances correctly even the instances are in proximity. From Figure 5, we see that mostly outliers are misclassified.

7 Conclusion and Future Work

In this paper, we presented a new dataset having 6,853 tweet post examples annotated with good and bad news labels. This dataset will be publicly available for the research community. We also presented a comparative analysis of supervised classification methods. We investigated nine different feature types and 8 different machine learning classifiers. The most robust result in our analysis was the contribution of the BERT-base model in in-domain but also in out-of-domain evaluations. Among structural features, significant terms significantly outperform the rest. We also showed that sentiment scores are not appropriate to classify good-vs-bad news.

In our future work, we plan to expand our investigation by including other features. We also plan to propose this model for the good-bad classification of news articles.

8 Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. ZE 915/7-1 "Data and Knowledge Processing (DKPro) - A middleware for language technology", by the Global Young Faculty[11] and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media".

References

Piush Aggarwal. 2019. Classification approaches to identify informative tweets. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*. Varna, Bulgaria, page To be published.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news.

Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is stronger than good. *Review of General Psychology* 5(4):323–370. https://doi.org/10.1037/1089-2680.5.4.323.

[11] https://www.global-young-faculty.de/

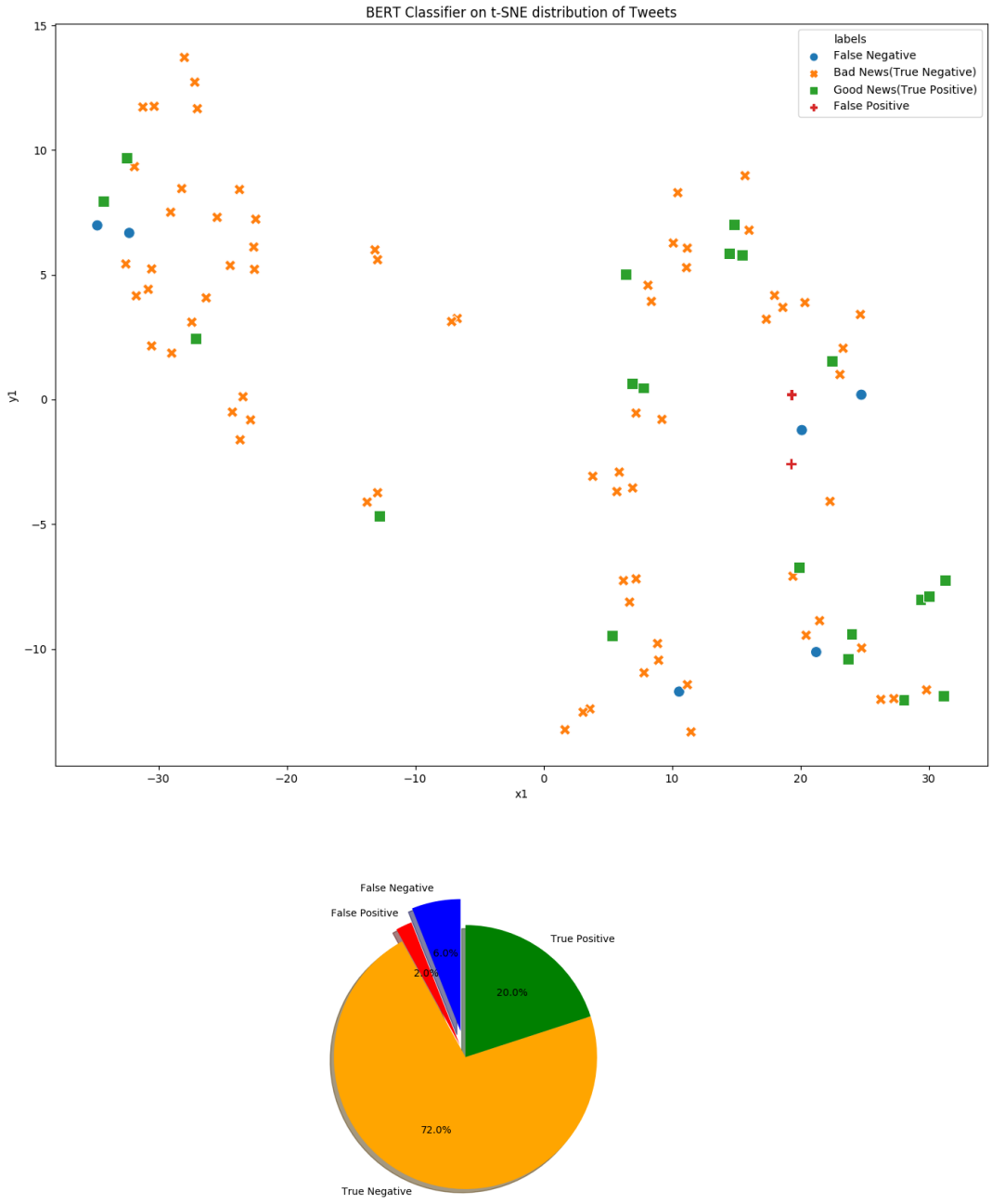

Figure 5: t-SNE distribution of random 100 test points with Bert's performance. The pie chart displays the percentage of BERT's misclassifications on these points.

John H. Boyd, Jian Hu, and Ravi Jagannathan. 2005. The stock market's reaction to unemployment news: Why bad news is usually good for stocks. *Journal of Finance* 60(2):649–672.

PHILLIP A. BRAUN, DANIEL B. NELSON, and ALAIN M. SUNIER. 1995. Good news, bad news, volatility, and betas. *The Journal of Finance* 50(5):1575–1603. https://doi.org/10.1111/j.1540-6261.1995.tb05189.x.

Jennifer Conrad, Bradford Cornell, and Wayne R. Landsman. 2002. When is bad news really bad news? *The Journal of Finance* 57(6):2507–2532. https://doi.org/10.1111/1540-6261.00504.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805. http://arxiv.org/abs/1810.04805.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382. https://doi.org/10.1037/h0031619.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 42–47. https://www.aclweb.org/anthology/P11-2008.

Alec Go. 2009. Sentiment classification using distant supervision.

Lars Kai Hansen, Adam Arvidsson, Finn Aarup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news - affect and virality in twitter. In *Communications in Computer and Information Science*, Springer Berlin Heidelberg, pages 34–43. https://doi.org/10.1007/978-3-642-22309-9_5.

Tony Harcup and Deirdre ONeill. 2017. What is news? *Journalism Studies* 18(12):1470–1488. https://doi.org/10.1080/1461670X.2016.1150193.

Wendy M. Johnston and Graham C. L. Davey. 1997. The psychological impact of negative tv news bulletins: The catastrophizing of personal worries. *British Journal of Psychology* 88(1):85–91. https://doi.org/10.1111/j.2044-8295.1997.tb02622.x.

Michael A. Kamins, Valerie S. Folkes, and Lars Perner. 1997. Consumer responses to rumors: Good news, bad news. *Journal of Consumer Psychology* 6(2):165–187. https://doi.org/10.1207/s15327663jcp0602_03.

Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971* .

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Paul R. Milgrom. 1981. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics* 12(2):380–391. http://www.jstor.org/stable/3003562.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one* 11(5):e0155036.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*. pages 1–18.

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, New York, NY, USA, WebSci '11, pages 8:1–8:7. https://doi.org/10.1145/2527031.2527052.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. pages 502–518.

Paul Rozin and Edward B. Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review* 5(4):296–320. https://doi.org/10.1207/S15327957PSPR0504_2.

Aliaksei Severyn and Alessandro Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. pages 464–469.

Stuart N. Soroka. 2006. Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics* 68(2):372–385.

Stuart N. Soroka. 2006. Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics* 68(2):372–385. https://doi.org/10.1111/j.1468-2508.2006.00413.x.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne.