Classification Approaches to Identify Informative Tweets

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

Abstract

Social media platforms have become prime forums for reporting news with users sharing what they saw, heard or read on social media. News from social media is potentially useful for various stakeholders including aid organizations, news agencies, and individuals. However, social media also contains a vast amount of non-news content. For users to be able to draw on benefits from news reported on social media it is necessary to reliably identify news content and differentiate it from non-news. In this paper, we tackle the challenge of classifying a social post as news or not. To this end, we provide a new manually annotated dataset containing 2,992 tweets from 5 different topical categories. Unlike earlier datasets, it includes postings posted by personal users who do not promote a business or a product and are not affiliated with any organization. We also investigate various baseline systems and evaluate their performance on the newly generated dataset. Our results show that the best classifiers are the SVM and BERT models.

1 Introduction

In the last decade, social media have become the platform par excellence for all kinds of online information exchange, such as content creation, consumption, and sharing; commenting on and engaging with contents posted by others. During unwanted situations like natural calamities, accidents, etc., users provide informative postings on social media websites to report about the incidents, to share an update about them and inform others about what they saw, heard or read. In this case, the users play the role of journalists and report the news to the public. However, there is also a vast amount of data that does not contain news-like information such as personal information, chats among friends, etc. Analyzing social media posts for whether they are news or not would allow e.g. aid providers during natural calamities to determine relevant information and plan appropriate actions. Furthermore, journalists could use such analysis to determine newsworthy information or even gain updates about events they have been reporting.

This paper contributes to the task of classifying social media posts, specifically Twitter messages, as news or non-news by providing data and a set of benchmark results for the task. The main contribution of the paper includes dataset[1] containing 2992 tweets manually labeled as *news* or *not*. To the best of our knowledge, related datasets are either event specific (Freitas and Ji, 2016) or queried with news-related keywords or hashtags like the name of news agencies (Liu et al., 2017). Unlike these datasets, our data consists of news reported by individual users and not just specific to tweets posted by news agencies. The dataset is developed to include tweets coming from first-hand reporters and witnesses of an event, which would be useful in the aforementioned scenarios. Although these first-hand reports can be very important in a given situation, the tweets coming from individuals are not identified as news by hashtags and are therefore more difficult to classify as news or not, in particular as individual tweets are more likely than organizational ones to report personal information. Furthermore, our dataset contains a variety of topics, unlike previously reported data which is focused on an event. We also investigate the behaviour of the dataset, find patterns and regulari-

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

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 7–15,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_002

ties using text visualisations.

For news classification, we adopt a supervised machine learning paradigm and report the performance of seven classifiers, which can be used as baselines in future work. We report the results of SVM (Chang and Lin, 2011), Logistic Regression (Fan et al., 2008), Random Forest (Breiman, 2001), Decision Tree (Breiman et al., 1984) and Xgboost classifier (Chen and Guestrin, 2016). In addition to shallow learning approaches, we train a Multi-Layer-Perceptron (MLP) model (Hinton, 1989) and also we use the pre-trained BERT-base model (Devlin et al., 2018). In the end, we claim the usabilty of our dataset by performing cross-domain experiments.

In this paper, we first discuss the related work (Section 2). In Section 3 we describe the dataset which we plan to make publicly available. In Section 4, we describe our experiments and present the results of the baseline systems used. We conclude and outline our future directions in Section 5.

2 Related Work

A widely accepted analysis of news values are defined by Galtung and Ruge's twelve news factors (Harcup and ONeill, 2017). According to this research, generally, a news story should be selected if it is published in context of potential figures, celebrity or organisation, fulfilling public need and interest, related to curiosity and amazement, a propaganda, positive-negative events, focusing on a huge crowd or relevant to the audience. In the last few years, there have been several studies published on the application of computational methods in order to identify news from tweets. Sankaranarayanan et al. (2009) built a news processing system, called *TwitterStand* using an unsupervised approach to classify tweets collected from pre-determined users who frequently post news about events. Sriram et al. (2010) use lexical and structural features based multi-class classification on manually annotated tweets having different categories (including *news*). Castillo et al. (2011) investigate tweet newsworthiness classification using features representing the message, user, topic and the propagation of messages. Others use features based on social influence, information propagation, syntactic and combinations of local linguistic features as well as user history and user opinion to select informative tweets (Inouye and Kalita, 2011; Yang et al., 2011; Li et al., 2012; Ren et al., 2013; Chua and Asur, 2013). Freitas and Ji (2016) use content based features like slang usage, sentiment terms, etc. to identify newsworthy tweets. Liu et al. (2017) use unsupervised approaches like clustering to identify news related topics among twitter postings. We differ from related work in various aspects. First, our dataset consists of tweets not specific to news agencies. Messages posted by news agencies can be easily tracked using e.g. the news agencies' hashtags. However, news posts reported by normal users will not have such hashtags and are difficult to determine. Next, such normal user-generated contents are of more value since they are the first source of information and tracking and knowing about them can e.g. in natural disaster situations be life-saving. Furthermore, our dataset is not specific to a particular topic but contains tweets from 5 different categories that are topically not related. Finally, we investigate various supervised techniques on this dataset to provide the community with various baselines.

Label	**Tweet**
News	*Indian cities and towns became less clean after Prime Minister Narendra Modi's Swaach Bharat mission*
News	*Unsafe abortion could induce some health related implications such as health risks to the girl or woman including #HUV/AIDS risks and #STDs*
Not News	*@chamberlainusoh If #Ebola has no known cure, what's then the need of going to the hospital*
Not News	*Honestly: ambient intelligence is a concept in the Internet of Things. But really do we want soo much controll handed over to devices?*

Table 1: Examples of news and not-news tweets

3 Dataset

Our dataset contains tweets labelled as news or not. Tweets are collected from five different categories and get the labels using crowd-source experiments. For annotation instructions, we sum-

Category	Topics	Collected	Annotated
Health	Ebola	90,430	287
	HIV	31,566	275
Natural Disaster	Hurricane Harvey	1,458,000	304
	Hurricane Irma	4,698,000	302
Terrorist Attack	Macerata oohmm	492,159	297
	Stockholm Attack	344,396	307
Geography and Env.	AGU17	29,997	310
	Swachh Bharat	19,868	283
Science and Edu.	IOT	6,326,806	319
	Nintendo	104,695	308

Table 2: Categories, their topics and distributions for the dataset generation

marised Galtung and Ruge's (Harcup and ONeill, 2017) twelve news factors and consider a text statement as news story if it holds informative elements or noticeable events. Similarly, tweets with no informative content are considered as not newsworthy. Table 1 illustrate examples of *news* and *not news*. With this, we believe to have a simple and sophisticated annotation task.

Data Collection Our data contain tweets from 5 categories with which we aim to have wider topic coverage. Furthermore, for each category, two different sub-topics are chosen to make the dataset more diverse. The first and second columns of Table 2 represent categories and their corresponding topics.

To collect the data, we used the following strategies. For the health category, for *Ebola* tweets, we used tweet-ids provided by Tamine et al. (2016) and for *HIV*, we used different hashtags shown in Table 3. For the natural disaster category, we collected *Hurricane Harvey* and *Irma* tweets from Littman (2017). From Tweet Catalog portal[2], we collected tweets related to *Macerata* and *Stockholm attack*. We use *AGU17* tweets from Pikas (2018) and for *Swachh Bharat Abhiyan* (Clean India Campaign), we looked for tweets containing hashtags shown in the second row of Table 3. For *IOT*, we used tweets from Bian et al. (2016) and for *Nintendo*, we used one of the kaggle datasets[3] which consists of tweets that were collected during the Nintendo E3 2018 Conference. The third column of Table 2 represents number of tweets collected for the aforementioned topics.

[2]https://www.docnow.io/catalog/

[3]http://tiny.cc/iookbz

Topic	Hash-Tags
Hiv	#AIDS, #aids, #hiv, #HIV, #PLHIV, #StopHIV, #EndAIDS, #HIVTreatmentWorks
SB	#MyCleanIndia, #SwachhBharat, #SwachhBharatSwasthBharat, #Killpollution, #SwachhBharatSwasthBharat

Table 3: Hashtags for tweets collection (here SB refers to *Swachh Bharat)*

Data Annotation From the collected tweets, we first filtered out all the tweets which are not in English language. Then we removed re-tweets and finally removed duplicates based on lower-cased first four words of tweets keeping only the first one, then we randomly pick 500 tweets from each topic.

To annotate tweets whether they are *news* or *not* we used the crowd-sourcing platform Figure Eight[4]. We showed each annotator 5 tweets per page and paid 3 US Cents per tweet. To ensure quality, we used 125 test questions created by 5 different annotators[5]. 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%

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

[5]These are non-crowd annotators. All are post-graduate students and use Twitter to post information on a daily basis. We considered a tweet as test instance if at least 4 annotators agreed on the class label.

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 which are annotated by at least 3 annotators. In addition to the news and not news categories, we also allowed a third category, namely *not sure*. We filtered out tweets where annotators were unsure about their judgment. We use a total 5000 tweets to annotate. Of these, 2992 were classified as news or not news. The other 2008 tweets were discarded because the annotators were not sure about their decision. The topic-wise number of successful annotations are displayed in the fourth column of Table 2. Further, we randomly split the resulting dataset into train and test set. Table 4 shows the distribution of each set.

Label	**Train**	**Test**	**Total**
NEWS	756	253	1,009
NOT NEWS	1,731	252	1,983
All	2487	505	2,992

Table 4: Dataset distribution

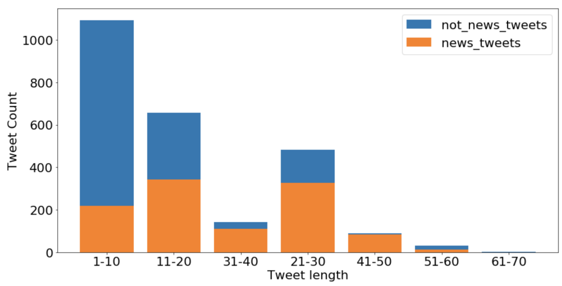

Figure 1: Length based distribution of tweets labelled with news and not news

Inter Annotator Agreement To evaluate the quality of the annotation, we compute Fleiss' kappa (Fleiss, 1971) scores between the annotators. For test questions, we record a kappa of 0.522, which indicates good agreement. For instances uploaded to the crowdsource platform, the majority class label for each tweet is collected and we compared it to the labels provided by the annotators. Such strategy is also followed by earlier studies (Zubiaga et al., 2016). In the end, an agreement of 0.443 is recorded, indicating a moderate agreement among the annotators.

Not News	**News**
new orleans	*tropical storm*
stay safe	*african migrants*
hope everyone	*italy attack*
blog post	*northern league*
please stay	*target immigrants*
safe everyone	*attack targets*
go time	*tropical depression*
new blog	*caribbean sea*

Table 5: List of most frequent bi-grams in the *news* and *not-news* corpora

Data Analysis To analyse the generated dataset, we perform several experiments (Mien, 2017) that visualise differences in the behaviour of news and not-news tweets. Also, the analysis helps in finding patterns and regularity among the data which certainly play a major role for deciding features and the further classification process. Before experimentation, we pre-processed the generated dataset by removing numbers, stop-words and tweet specific keywords like *hiv*, *macerata*, etc. from the tweet texts and lower-cased them. First, we analyse tweet length distribution for each tweet label. In Figure 1, each bar presents the tweet count for each label with respect to the word length interval. From the Figure, it can be concluded that *news* tweets are much less frequent than *not news* tweets if their length is less than 10 words, but as the length of the tweets get increases, *news* tweets become dominant over *not news* ones.

To learn about the number and kind of topics present in a body of text, two tweet corpora are created by concatenating the tweet posts for each label (*news* and *not news*) and most frequent bi-grams are extracted (see Table 5). From the Table, we can see, *not news* tweets generally focus on conversation related words whereas newsworthy tweets include instances associated with events, group references, etc.

We also find some of the terms which are frequently available in both text corpora. We plot lexical dispersion which displays occurrence of terms with respect to word offset in the corpus[6]. Each word on the y-axis has a strip representing the entire text in terms of offset, and a mark on the strip indicates the occurrence of the word at that off-

[6]taking only the first 10,000 terms for each corpus

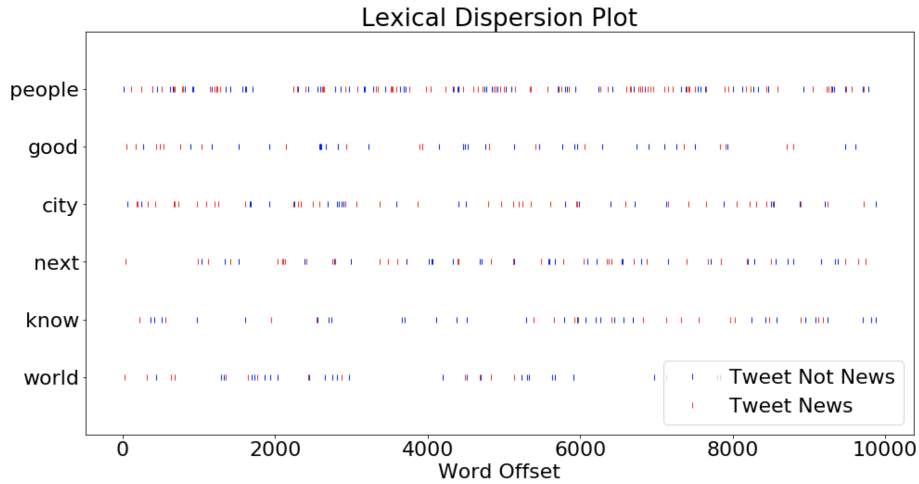

Figure 2: Lexical dispersion distribution of commonly used terms found in the Twitter corpus annotated with news and not news labels

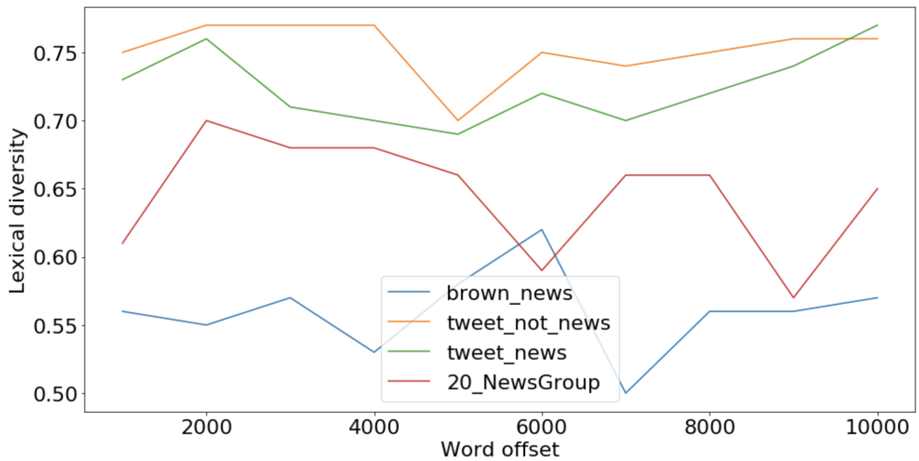

Figure 3: Lexical diversity distribution of different corpora dispersed on word offset interval

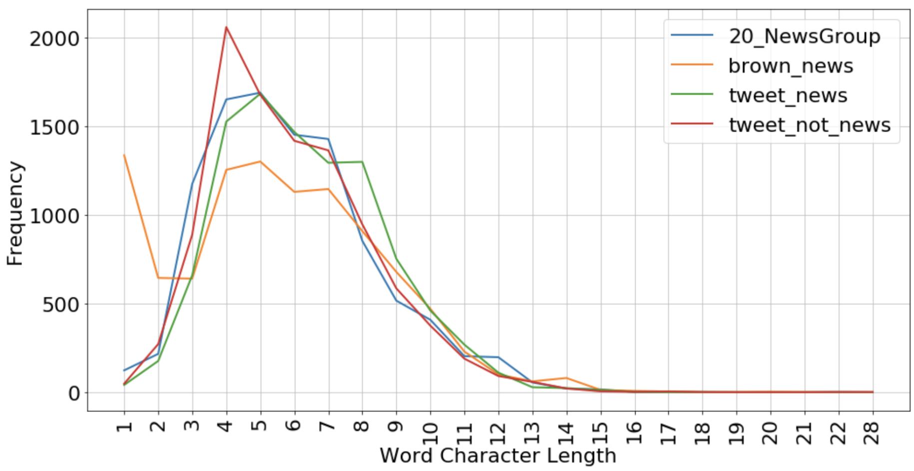

Figure 4: Word's character length distribution for different corpora

set. Analysis shows that there are very few news tweets which contain *good* and *know* terms. However, terms like *people* and *city* are equally dispersed among the tweets in both categories (news and not news).

Lexical diversity (ld) distribution of the generated dataset is also analysed, which can be defined for the given text t as:

$$ld = \frac{\textit{count of unique words in } t}{\textit{count of total words in } t} \tag{1}$$

For analysing the lexical diversity (Johansson, 2008), first 10,000 terms for each tweet corpus are taken and divide them into chunks of size 1000 words. For each chunk, ld is calculated (Equation 1) and plotted it with respect to word offset intervals as shown in Figure 3.

We also plot the same distribution for two well-known news corpora. The first corpus (also called 20-NewsGroup[7]) comprises around 18000 newsgroups posts on 20 topics. For the other corpus (Brown Corpus[8]), we focus only on news genre which include news from 44 different categories. From Figure 3, it can be interpreted that lexical diversity for news-related corpora (brown_news, tweet_news and 20_NewsGroup) is low compared to not news tweet corpora.

We also analyse the distribution of word length in terms of the number of characters and compare it among different corpora as discussed above. We took a subset of each corpus (first 10,000 terms) and plot the frequency of each word length for each corpus (see Figure 4). The figure illustrate that in not news tweets, most words have a length of (size) 4 whereas in news corpora most words hold 5 characters.

Finally, we tried to figure out the n-gram distribution pattern among different corpora. We plot n-gram distribution for each corpus (see Figure 5) where n is 1 to 5. In the Figure, the x-axis has different values of n-grams and the y-axis has the number of times the n-gram has occurred[9]. The figure shows that news instances of tweets capture more bi-grams than not-news ones.

[7] http://qwone.com/~jason/20Newsgroups/

[8] http://tiny.cc/bytkbz

[9] here only those n-grams are chosen which are occurred more than 1 time.

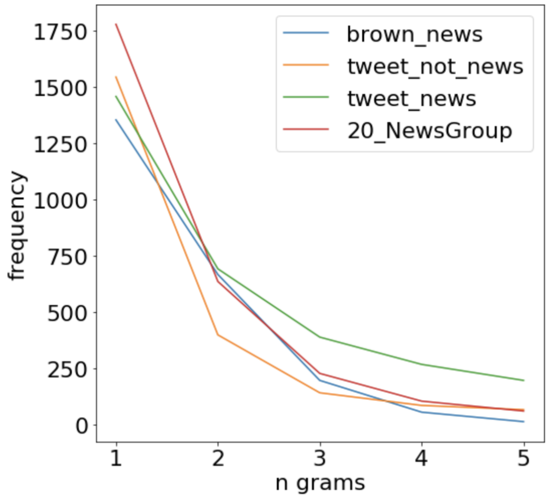

Figure 5: N-gram frequency distribution for corpora

4 Experiments and Results

As our task is to identify whether a particular tweet is news or not, we treat it as a binary classification task. We train our baseline classifiers on the training set and evaluate the resulting models on the test set where label distribution is in proportion with that of training set (see Table 4 for the training and testing split).

Preprocessing and Feature Extraction Tweets are lower-cased and use Ark Tokenizer (Gimpel et al., 2011) for segmentation. After these preprocessing steps, we represent each posting by a dense embedding, created by the mean of the individual words embeddings. We use the pre-trained embeddings provided by (Mikolov et al., 2018), which are trained on the common crawl corpus. In addition to posting embeddings, we also extract syntactic features in the form of TF-IDF vectors (Salton and McGill, 1986) for up to 3 grams having vocabulary size as vector dimensions.

Baseline Classifiers To classify *news* and *not news* we train the following classifiers: SVM (Chang and Lin, 2011) with regularization parameter (C) as 10 and rbf as kernel, Logistic Regression (Fan et al., 2008) with 0.1 as inverse regularization strength, Random Forest (Breiman, 2001) with 15 as maximum depth and 500 trees. We use Decision Tree (Breiman et al., 1984) with 2 minimum sample leaves and 3 as minimum sample split and Xgboost classifier (Chen and Guestrin,

Approach	Embeddings Precision	Recall	F_1	TF-IDF(1-3 gram) Precision	Recall	F_1
SVM	**.854**	**.851**	**.851**	.808	.808	.808
BERT	.835	.841	.838	-	-	-
Random Forest	.839	.838	.837	**.821**	**.820**	**.820**
Logit Reg.	.828	.827	.827	.812	.812	.812
Xgboost	.823	.822	.822	.802	.794	.792
MLP Classifier	.801	.772	.767	.809	.794	.791
Decision Tree	.733	.733	.733	.755	.754	.754
Majority Vote - all NOT	.331	.500	.399	.331	.500	.399

Table 6: Classifiers evaluation results

2016). In addition to shallow learning approaches, we train a model called Multi-Layer-Perceptron (MLP) (Hinton, 1989) with Sigmoid activation function (Cybenko, 1989), 0.001 as l2 penalty (Ng, 2004), adaptive as learning rate (Schaul and LeCun, 2013) and 0.1 as tolerance. Apart from the mentioned hyper-parameters, we use default-parameters provided by scikit-learn (Pedregosa et al., 2011). Finally, we use the pre-trained BERT-base model (Devlin et al., 2018) to create a vector representation of a posting. We fine-tune the model on the training dataset using a sequence length of 64 and batches of 32 and training epochs of 2.

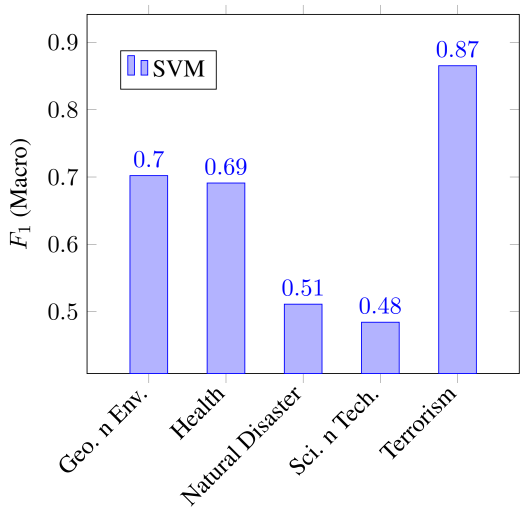

Figure 6: Cross domain performance of SVM for each tweet category

Evaluation and Results We evaluate the performance of the classifiers using the test set (Table 4). We report Precision, Recall, and Macro F_1 (Powers and Ailab, 2011) for all the classifiers. We use the majority class (all-NOT) as the additional baseline. Table 6 shows the performance scores. The results show that the SVM classifier with the posting vector-representation achieves the best F-Score, followed by BERT. Using content based semantic features like word embeddings we were able to achieve better performance than using syntactic based features like TF-IDF vectors.

Dataset Usability Using cross domain experiments, we investigate the practical usability of our dataset where we train our best model on in-domains and test on out-of-domain data. For this purpose, we split the dataset into a training set consisting of all examples that belong to 4 categories and the left out category instances are used to create a held-out test set. We train a SVM classifier with fasttext embeddings on the training set. Figure 6 illustrates the results of the model tested on different domains. The models achieve an average macro F_1 score of 65% which is much higher than the majority class baseline. We also see low F_1 scores in the cases of *Science n Technology* and *Natural Disaster* domains. For Science n Technology, one possible reason is availability of only 2% of true news labels. In case of Natural Disaster, we found 56% news true labels. Therefore, to find the root cause, we perform an experiment where we add a small proportion of out-of-domain data to the training set. We transfer 12% of the instances of Natural Disaster from test-set to train-set. The model achieve an F_1 score of 69% which is a substantial increase from its previous value. The analysis show the practical usability of the dataset. In some cases, model may under-fit, such cases can be handled by adding small amounts of out-of-domain data.

5 Conclusion

In this paper, we release a new dataset containing 2992 tweets annotated as news or not. This dataset will be publicly available for the research community. To the best of our knowledge, this is the first dataset that consists of Twitter postings with 5 diversified categories consisting of postings from first-hand reporters and witnesses of an event, which would be useful in emergency situations such as natural disasters to gain knowledge about the happenings. We experimented with seven different supervised machine learning techniques and showed that best performances can be achieved using the SVM and BERT models. These techniques serve as baselines.

In the future, we would like to put more focus on data augmentation and further categorization of newsworthy tweets as good or bad news.

Acknowledgments

I thank *Dr. Ahmet Aker* (University of Duisburg-Essen) and the anonymous reviewers for their most useful feedback. 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[10] and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media".

References

Jiang Bian, Kenji Yoshigoe, Amanda Hicks, Jiawei Yuan, Zhe He, Mengjun Xie, Yi Guo, Mattia Prosperi, Ramzi Salloum, and Franois Modave. 2016. Mining Twitter to Assess the Public Perception of the Internet of Things. *PLOS ONE* 11(7):1–14. https://doi.org/10.1371/journal.pone.0158450.

Leo Breiman. 2001. Random Forests. *Machine Learning* 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 1984. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA. http://cds.cern.ch/record/2253780.

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

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '11, pages 675–684. https://doi.org/10.1145/1963405.1963500.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2(3):27:1–27:27. https://doi.org/10.1145/1961189.1961199.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '16, pages 785–794. https://doi.org/10.1145/2939672.2939785.

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

G. Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2(4):303–314. https://doi.org/10.1007/BF02551274.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* .

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* 9:1871–1874. http://dl.acm.org/citation.cfm?id=1390681.1442794.

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

Jesse Freitas and Heng Ji. 2016. Identifying News from Tweets. In *NLP+CSS@EMNLP*.

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, pages 42–47. http://aclweb.org/anthology/P11-2008.

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

Geoffrey E. Hinton. 1989. Connectionist learning procedures. *Artificial Intelligence* 40(1):185 – 234. https://doi.org/https://doi.org/10.1016/0004-3702(89)90049-0.

David Inouye and Jugal K. Kalita. 2011. Comparing Twitter Summarization Algorithms for Multiple Post Summaries. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. pages 298–306. https://doi.org/10.1109/PASSAT/SocialCom.2011.31.

V Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics* 53:61–79.

Hao Li, Yu Chen, Heng Ji, Smaranda Muresan, and Dequan Zheng. 2012. Combining Social Cognitive Theories with Linguistic Features for Multigenre Sentiment Analysis. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*. Faculty of Computer Science, Universitas Indonesia, pages 127–136. http://aclweb.org/anthology/Y12-1013.

Justin Littman. 2017. Hurricanes Harvey and Irma Tweet ids. https://doi.org/10.7910/DVN/QRKIBW.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. pages 1483–1493. https://doi.org/10.1109/BigData.2017.8258082.

Mien. 2017. Visual text analytics with python - mien. https://medium.com/@therealmien/visual-text-analytics-with-python-6111ae8b16df.

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)*.

Andrew Y. Ng. 2004. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '04, pages 78–. https://doi.org/10.1145/1015330.1015435.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Christina Pikas. 2018. American Geophysical Union Annual Meeting Tweets 2017 https://doi.org/10.6084/m9.figshare.5756514.v1.

David M. W. Powers and Ailab. 2011. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation.

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*. ACM, New York, NY, USA, SIGIR '13, pages 513–522. https://doi.org/10.1145/2484028.2484052.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, USA, GIS '09, pages 42–51. https://doi.org/10.1145/1653771.1653781.

Tom Schaul and Yann LeCun. 2013. Adaptive learning rates and parallelization for stochastic, sparse, non-smooth gradients. *CoRR* abs/1301.3764. http://arxiv.org/abs/1301.3764.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short Text Classification in Twitter to Improve Information Filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '10, pages 841–842. https://doi.org/10.1145/1835449.1835643.

Lynda Tamine, Laure Soulier, Lamjed Ben Jabeur, Frederic Amblard, Chihab Hanachi, Gilles Hubert, and Camille Roth. 2016. Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, HT '16, pages 159–168. https://doi.org/10.1145/2914586.2914589.

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*. ACM, New York, NY, USA, SIGIR '11, pages 255–264. https://doi.org/10.1145/2009916.2009954.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11(3):e0150989.