Proceedings of CSR-NLP I @LREC 2022, pages 40–45
Marseille, 25 June 2022

Detecting Violation of Human Rights via Social Media

Yash Pilankar, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, Pramod Pathak
School of Computing
National College of Ireland
Mayor Street Lower, IFSC, Dublin 1, Ireland
x19216858@student.ncirl.ie, firstname.lastname@ncirl.ie

Abstract

Social media is not just meant for entertainment, it provides platforms for sharing information, news, facts and events. In the digital age, activists and numerous users are seen to be vocal regarding human rights and their violations in social media. However, their voices do not often reach to the targeted audience and concerned human rights organization. In this work, we aimed at detecting factual posts in social media about violation of human rights in any part of the world. The end product of this research can be seen as an useful assest for different peacekeeping organizations who could exploit it to monitor real-time circumstances about any incident in relation to violation of human rights. We chose one of the popular micro-blogging websites, Twitter, for our investigation. We used supervised learning algorithms in order to build human rights violation identification (HRVI) models which are able to identify Tweets in relation to incidents of human right violation. For this, we had to manually create a data set, which is one of the contributions of this research. We found that our classification models that were trained on this gold-standard dataset performed excellently in classifying factual Tweets about human rights violation, achieving an accuracy of upto 93% on hold-out test set.

Keywords: Human Rights Violation, Fact Checking, Machine Learning

1. Introduction

Over the last two decades, we have seen astounding growth and development across a wide range of disciplines including science and technology, and the adaptation of modern ideologies has significantly accelerated the growth of every nation. However, there are still many locations around the world, where people do not even enjoy basic human rights and freedoms. In current affairs, there have been countless situations befalling around the world, where human rights are continuously being violated, and unfortunately these incidents go unnoticed. Activists across the world aim to bring such issues to light as soon as they become aware of such incidents. Likewise, media reporters and activists are on field risking their lives to cover such incidents and bring them in front of the world. We refer the readers to one recent heartbreaking incident that was reported by Laskar and Sunny (2021) in Hindustan Times. It is regarding Danish Siddiqui,[1] India's one of most renowned and Pulitzer prize-winning photojournalists, who was reportedly killed while reporting an instance where human rights were being violated. Russia's invasion of Ukraine is the world's centre of attention today, and the escalation in violations of human rights law, including deaths of civilians resulting from unlawful attacks are being reported everyday.

Over the past few years, we have seen many crowd-sourced technological solutions, one of which is Ushahidi.[2] It is a map-based tracker that is used to monitor event-based situation and tags the location over the map. Similarly, Syria Tracker[3] is a tool that is used for reporting incidents about human rights abuse and tag the location of the incidents in map so that its neighborhoods become aware about the situation.

There have been a staggering growth and usage of micro-blogging platforms since the beginning of this century. In fact, social media has become one of the mediums, where people raise voices for human rights and tend to share factual and truthful events occurring nearby for justice. Over the past decade, NLP researchers both from academia or industry investigated sentiment analysis by analyzing data from a variety of micro-blogging websites. However, significant portion of these works aimed at identifying characteristics or opinions of user-generated content such as users' emotions, intentions, mood, behaviors and sentiments (Neethu and Rajasree, 2013; Waseem and Hovy, 2016; Haque et al., 2019; Singh et al., 2020a; Singh et al., 2020b). Recently, Alhelbawy et al. (2020) developed a HRVI platform for Arabic to monitor human rights violation in several countries in Central Asia. For this, they built Naïve Bayes and Support Vector Machine (SVM) classifiers on tweet data. As in Alhelbawy et al. (2020), we focused in detecting human rights violation in Tweets. Unlike Arabic as in Alhelbawy et al. (2020), we considered English Tweets for our investigation so that incidents about human right violations worldwide can be traced.

More specifically, in this work, we focused on identifying "factual" information from Tweets rather than

[1] https://en.wikipedia.org/wiki/Danish_Siddiqui

[2] https://www.ushahidi.com/

[3] https://syriatracker.crowdmap.com/

categorising opinionated Tweets. In order to do this, we crawled Tweets in relation to events and incidents about human right violations. Then, we made use these Tweets in order to create a gold-standard dataset which is used to build and evaluate our HRVI classifiers. Everyday thousands of social media posts about entertainment or so become viral; however, posts about human rights violation are not seen, cornered, and does not reach to the targeted audience. Our work aims at aiding organizations whose intention is to keep peace and harmony within the nations and society by tracking situation and incidents in relation to human right violations. We employed a number of machine learning (ML) algorithms in order to build our HRVI classification models. Our expectation was that our HRVI systems would be able to identify specific factual Tweets. One of the main contributions of this work is the creation of the gold-standard data for the HRVI task, which, we believe, could serve as an invaluable asset as far as this line of NLP research is concerned. To the best of our knowledge, there is no readily available dataset that one can freely use for HRVI via social media platforms. We also believe that our work would not only advance NLP research but also have positive societal and political impacts.

2. Related Works

For sentiment analysis gathering relevant dataset has always been a challenge but can be collected following a set of standard methods, e.g. crawling, scrapping and REST API. Jiang et al. (2017) used scrapping technique for getting microblogs from Sina.[4] Twitter is one of the most widely-used social media platforms in the world and one can use its API to easily to fetch and crawl millions of Tweets. Waseem and Hovy (2016) created a corpus mainly on hatespeech by collecting over 130K Tweets using Twitter API. Likewise, Davidson et al. (2017) collected a set of Tweets (25K) in order to create a corpus for hatespeech. Further, in order to produce a gold-standard data for their task, Waseem and Hovy (2016) prepared a set of rules which were used in their annotation task. They ended up with a dataset containing 16K Tweets. In case of Davidson et al. (2017), they performed the annotation task with the help of CrowdFlower[5] users.

Unlike the strategy described above, Zahoor and Rohilla (2020) took a different approach for annotation as they utilized TextBlob, a NLP library, for getting sentiment (i.e. positive, negative and neutral) of posts. Neethu and Rajasree (2013) proposed a simple method for creating lexical feature vector for collecting Tweets from Twitter, and their annotation task was performed manually. Hamdan et al. (2015) used feature extraction methods such as polarity score over ten different lexicon along with a slang dictionary of Twitter for handling social media post containing slang. Lim et al. (2020) used features from pre-trained language model (Embedding from Language Models (ELMo)), and Term Frequency–Inverse Document Frequency (TF-IDF). Interestingly, they observed that use of parts-of-speech (PoS) feature does not help in classification of micro-blog texts.

Event-based sentiment classification was one of the important turnarounds regarding 2016 Presidential Election in United States. Somula et al. (2020) performed an experiment taking Tweets posted during that time into account for the prediction of the election winner. The event-based sentiment classification has also been been adopted in different campaigns. For example, Fitri et al. (2019) performed predictive analysis on anti-LGBTQ campaign in Indonesia. The similar strategy was also taken into account for monitoring human rights abuse in Iraq (Alhelbawy et al., 2020). Alhelbawy et al. (2020) developed a map-based platform called Ceasefire[6] that reports location where any human rights were violated. It also offers a feature that fetches Tweets from Twitter about any human rights abuse and tags the locations mentioned in the Tweets. As the portal was specifically developed for peace in Iraq, it was limited to Arabic only. Alhelbawy et al. (2020) used vector space learning technique for text, i.e. word2vec (Mikolov et al., 2013), and TF-IDF method for weighted scheme. They tested a number of ML techniques and algorithms (e.g. Linear SVM, Gaussian SVM and Naïve Bayes) in their task, and achieved the highest accuracy when they applied the combination of CNN and LSTM.

As for sentiment analysis, there have been a plethora of works that studied this area of NLP considering both high-resource and low-resource languages. We refer the interested readers some of the notable works (Lim et al., 2020; Kanakaraddi et al., 2020; Qin et al., 2020) who investigated sentiment analysis using more advanced ML techniques such as bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2018a).

3. Experimental Setups

3.1. Collecting Tweets

To the best of our knowledge, there is no publicly available dataset for HRVI task. For this, we created a gold-standard data set for this task. We used Twitter API, Tweepy,[7] in order to collect Tweets. We are interested in collecting those Tweets which would be relevant for the task. This is in fact a challenging and time-consuming task. One of the ways is to collect Tweets from those Twitter accounts which are reliable and post

[4] https://en.wikipedia.org/wiki/Sina_Weibo

[5] https://en.wikipedia.org/wiki/Figure_Eight_Inc.

[6] http://iraq.ceasefire.org/

[7] https://docs.tweepy.org/en/stable/api.html

Tweets on subjective matters such as human rights violation. We looked at the profiles of various NGOs and peace-maker organizations, e.g. Human Rights Watch, Amnesty, United Nations, and Refugees, and came up with a list of relevant Twitter accounts. We also looked at personal Twitter profiles of many activists, e.g. Malala Yousafzai, Nadia Murad, who have been vocal over human rights and their violation. In sum, we collected those Twitter account names that are related to the context of human rights and violations of human rights, which are required for the creation of our dataset. Finally, we used Tweepy with the list of user accounts and collected Tweets. Our second approach is based on search query functionality available in Tweepy. We turned on language filtering functionality and set it to English. This does not consider Tweets of non-English languages. We also turned on filtering for RetTweets in order to avoid them. This helped in removing redundant Tweets. Lastly, we supplied a list of search keywords such as "child abuse", "ban on education", "attack on civilians". Using the two above approaches we collected a list of 15,590 Tweets which are considered for the annotation task (cf. Section 3.2).

3.2. Annotation Process

This section describes our data annotation process. In Section 2, we talked about different data annotation methods for the sentiment classification tasks. Our task is identification of human rights violation through social media posts. In short, it is a binary classification task where given a Tweet it checks whether there is any incident about human rights violation. We labelled a Tweet with "1" when we see that it contains information, event, fact or incident about human rights violation. The concerned Tweet may also contain location where the incident occurred. The additional clues that we considered for tagging were: (i) there may be a victim such as any community, person, group of people, and (ii) information about the assailant who violated the rights. The dataset that we created is different from the existing sentiment analysis tasks, where sentiments such as feelings and opinions of the user are checked based on content of the post alone. In our case, it is more focused on facts that is encoded in the Tweet. In sum, we labeled each of the collected clean Tweets with either one of the two categories: '1' (indicating the violation of human rights), '0' (normal post that does not indicate any violation of human rights). Note that since data annotation is an expensive and time-consuming task, we had only single annotator for this task, who is a native speaker of English and has excellent knowledge in Tweets or micro blogs.

On completion of annotation task, we ended up with a list of 10,077 annotated Tweets. Figure 1 shows the distribution of these instances in each class ('1' and '0'). As can be seen from Figure 1, this is a highly imbalanced dataset. We see that the number of instances of the minority class ('1': indicating human rights violation) is 1,057, and the same of the majority class is 9,020 ('0': indicating no human rights violation). We split the data set into two parts: train and test sets. The train and test sets contain 7,557 and 2,520 instances, respectively.

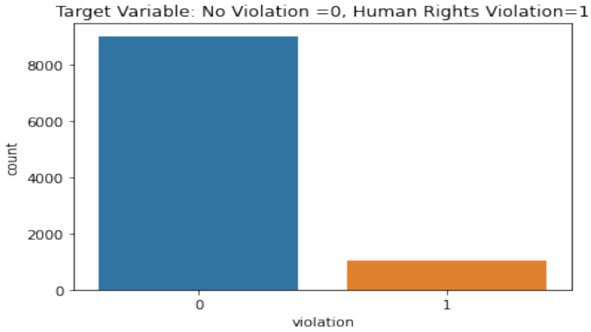

Figure 1: Class Distribution.

3.3. Quality of Annotation

At the end of the annotation task, each tweet is associated with one of the two tags: '1' or '0'. Since we have one annotator, one value is associated with each of the 10,077 Tweets. In order to measure how good annotation process was, a set of 200 Tweets were randomly sampled from the data set such that they are equally distributed across the both classes, and annotated by another annotator. The second annotator who only annotated this small set of Tweets (200) were instructed with the annotation guidelines that were given to the first annotator. On completion of this annotation task, we computed inter-annotator agreement using Fleiss' Kappa (Fleiss and Cohen, 1973) at Tweet level. For each tweet, we count an agreement whenever two annotators agree with the annotation result. We found the Kappa score to be high (i.e. 0.90) for the annotation task. This indicates that our tweet labeling task is to be excellent in quality.

3.4. Overview on our HRVI Systems

Figure 2 illustrates the working architecture of our HRVI system. Each of the components of the HRVI model is clearly shown in Figure 2, and they are placed under three different layers: data layer, logical layer and client layer. The data layer includes tasks such as data collection, cleaning and annotation. Spacy,[8] an open-source software library for advanced natural language processing, is used for data cleaning. It also took into account abbreviations, slangs, #tags, links, user tags, and provided us a clean data. We performed tonenisation, stop word removal and encoding (word embedding) based on the requirements of our learning algorithms. TF-IDF weighting is used for classical machine learning algorithms, i.e. random forest (RF), support vector machine (SVM)). RF is an extension of Bagging technique, which includes subspace sampling

[8] https://spacy.io/

strategy. Hyperparameters for the RF classifier were tuned using GridSearchCV,[9] a hyperparameter search technique using cross validation. As for SVM, we used the default set of hyperparameters of Scikit-learn[10] for our experiments.

Vaswani et al. (2017) introduced Transformer as an efficient alternative to recurrent or convolutional neural networks. Based on the Transformer architecture, Devlin et al. (2018b) proposed a powerful NN architecture – BERT – for a variety of NLP tasks including text classification such as sentiment analysis. BERT is a multi-layer bidirectional Transformer encoder architecture which provides context-aware representations from an unlabeled text by jointly conditioning from both the left and right contexts within a sentence. It can also be used as a pre-trained model with one additional output layer to fine-tune downstream NLP tasks, such as sentiment analysis, and natural language inferencing. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using the labeled data from the downstream tasks. There were two steps in BERT training: *pre-training* and *fine-tuning*. During pre-training, the model is trained on unlabeled data. As for fine-tuning, it is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using the labeled data from the downstream tasks (e.g. sentiment analysis). This strategy has been successfully applied to different fact checking tasks in social media (e.g. Williams et al. (2020)). In this work, we also investigated human rights violation identification in social media sphere using BERT.

4. Results and Discussion

In order to evaluate our HRVI models, we used metrics that are widely used for evaluating classifiers, i.e. accuracy, recall, precision and F1. In Table 1, we show the performance of our classifiers in terms of these metrics. As can be seen from Table 1, RF and SVM performed excellently in the task. BERT produces a moderate performance (an F1 of 0.80 and 75% accuracy on the test data).

	Acc	Precision	Recall	F1
RF	0.93	0.93	0.93	0.92
SVM	0.93	0.93	0.93	0.92
BERT	0.75	0.91	0.75	0.80

Table 1: Performance of our HRVI models.

We also show their performance using confusion matrix, which provides more insights on how they perform on each class. In Figure 3, we show confusion matrix for the RF classifier. As can be seen from Figure 3, RF is able to classify most of the test set instances correctly. However, it misclassified 161 instances of the positive class, i.e. they were incorrectly classified as normal Tweets ('0') (i.e. false negative (FN)). In sum, it performed poorly on the positive class (i.e. true-positive rate (TPR): 103 / 264 = 39.01%), and we are mainly interested in that class.

We show confusion matrix for SVM in Figure 4. We obtained a slightly improved classification performance. In other words, we obtained a slightly higher TPR (108/264 : 40.04%) this time. Again, its performance is below par on the class we are interested in. We show confusion matrix of classification results obtained with BERT in Figure 5. We see from Figure 5 that performance of BERT is much worse than that of RF and SVM.

Our dataset is a mixture of different types of posts including personal information, opinions, events, information about articles and publications on human rights. Moreover, this is a class-imbalanced data set (10% of Tweets were based on factual Tweets about human right violation). We manually looked at the Tweets of both classes. We observed that a number of Tweets of majority class seems to be factual Tweets at first glance. However, they were either instructive texts or expressions about opinion on human rights. As an example, we show a Tweet that belong to the majority class (class '0'): "*Students have the right to protest. Violence against peacefully protesting students —or anywhere else—can't be justified under any circumstances. As protests spread to campuses, we urge authorities to respect the right to dissent by peaceful protesters*". It is an opinion and not a factual post. It has a negative polarity. However, it does not express the fact that any harm was caused or any human right was violated. Such type of Tweets of training data could be one the reasons for poor TPR. We conjecture that another reason for poor TPR is the nature of our gold-standard dataset, which is class-imbalanced. Investigating this area (i.e. dealing with class-imbalanced data) is part of our future research plans.

5. Conclusions and Future Work

In this work, we investigated detection of violation of human rights via social media. We chose one of the most popular micro-blogging websites, Twitter, for our investigation. We used supervised learning algorithms in order to build human rights violation identification (HRVI) models such as Random Forest, Support Vector Machine, and state-of-the-art classification algorithm BERT. For this, we manually created a gold-standard dataset, which is in fact one of the main contributions of this work. The performance of our classifiers seem to be excellent in this task if we consider both classes. However, their performance are below par on positive class (i.e. on identifying Tweets in relation to incidents of human right violation). We identified a number of potential causes for this disparity. In order to counter this anomaly, in future, we plan to examine the fol-

[9] `https://scikit-learn.org/`

[10] `https://scikit-learn.org/stable/`

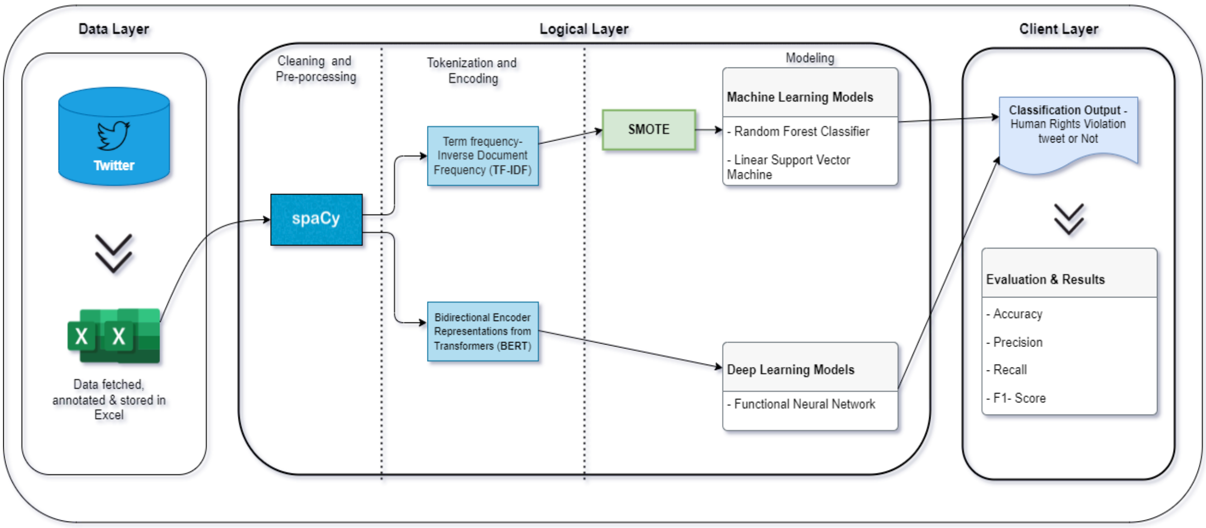
Data Layer
Logical Layer
Client Layer
Twitter
Data fetched, annotated & stored in Excel
Cleaning and Pre-porcessing
spaCy
Tokenization and Encoding
Term frequency-Inverse Document Frequency (TF-IDF)
Bidirectional Encoder Representations from Transformers (BERT)
SMOTE
Modeling
Machine Learning Models
- Random Forest Classifier
- Linear Support Vector Machine
Deep Learning Models
- Functional Neural Network
Classification Output - Human Rights Violation tweet or Not
Evaluation & Results
- Accuracy
- Precision
- Recall
- F1- Score

Figure 2: Project Architecture

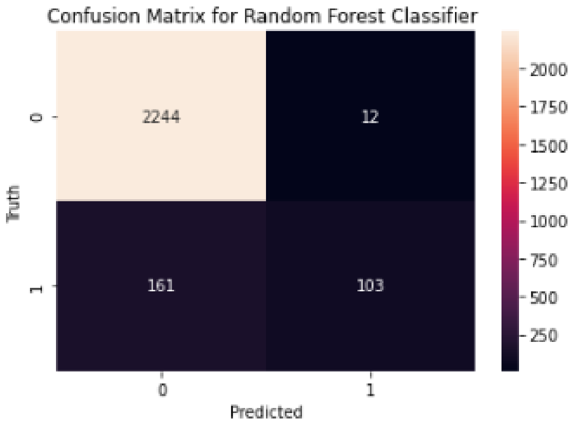

Figure 3: Confusion Matrix: RF

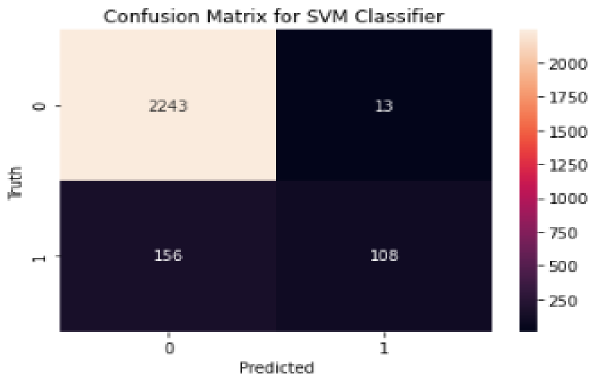

Figure 4: Confusion Matrix: SVM

lowing aspects of the task: (i) we want to increase the coverage for the positive class, (ii) exploring state-of-the-art strategies that deal with class-imbalanced text data, and (iii) play with different hyperparameters of the BERT model.

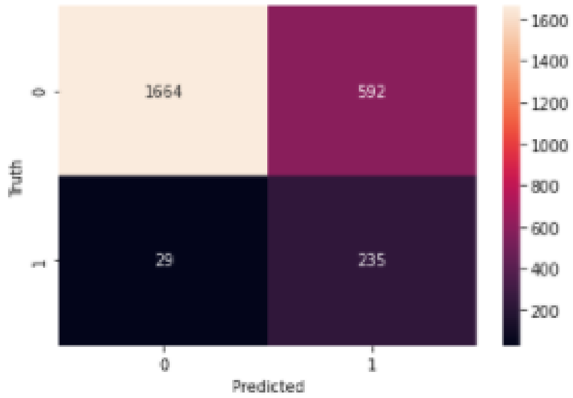

Figure 5: Confusion Matrix: BERT

Acknowledgement

A part of this research was conducted with the financial support of the Horizon 2020 project ITFLOWS under Grant Agreement No. 882986 and the ADAPT SFI Research Centre at Munster Technological University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

6. Bibliographical References

Alhelbawy, A., Lattimer, M., Kruschwitz, U., Fox, C., and Poesio, M. (2020). An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018b). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fitri, V. A., Andreswari, R., and Hasibuan, M. A. (2019). Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, 161:765–772. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Hamdan, H., Bellot, P., and Bechet, F. (2015). Lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Haque, R., Ramadurai, A., Hasanuzzaman, M., and Way, A. (2019). Mining purchase intent in twitter. In *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Jiang, D., Luo, X., Xuan, J., and Xu, Z. (2017). Sentiment computing for the news event based on the social media big data. *IEEE Access*, 5:2373–2382.

Kanakaraddi, S. G., Chikaraddi, A. K., Gull, K. C., and Hiremath, P. S. (2020). Comparison study of sentiment analysis of tweets using various machine learning algorithms. In *2020 International Conference on Inventive Computation Technologies (ICICT)*, pages 287–292.

Laskar, R. H. and Sunny, S. (2021). Indian journalist killed in line of duty by taliban. *The Hindustan Times*, Jul.

Lim, Y. Q., Lim, C. M., Gan, K. H., and Samsudin, N. H. (2020). Text sentiment analysis on twitter to identify positive or negative context in addressing inept regulations on social media platform. In *2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pages 96–101.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Neethu, M. S. and Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–5.

Qin, Q., Hu, W., and Liu, B. (2020). Feature projection for improved text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8161–8171, Online, July. Association for Computational Linguistics.

Singh, R. P., Haque, R., Hasanuzzaman, M., and Way, A. (2020a). Identifying complaints from product reviews: A case study on hindi. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2020)*, Dublin, Ireland.

Singh, R. P., Haque, R., Hasanuzzaman, M., and Way, A. (2020b). Identifying complaints from product reviews in low-resource scenarios via neural machine translation. In *Proceedings of ICON 2020: 17th International Conference on Natural Language Processing*, Patna, India.

Somula, R., Dinesh Kumar, K., Aravindharamanan, S., and Govinda, K. (2020). Twitter sentiment analysis based on us presidential election 2016. In Suresh Chandra Satapathy, et al., editors, *Smart Intelligent Computing and Applications*, pages 363–373, Singapore. Springer Singapore.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Williams, E., Rodrigues, P., and Novak, V. (2020). Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. *arXiv preprint arXiv:2009.02431*.

Zahoor, S. and Rohilla, R. (2020). Twitter sentiment analysis using lexical or rule based approach: A case study. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 537–542.