Sentiment Analysis in Social Media Data Using Machine Learning

Roli Shrivastava[1], Shatendra Kumar Dubey[2]

[1,2]Department of Information Technology, NRI Institute of Information Science and Technology, Bhopal (M.P), India
[1]roli.rlrl@gmail.com, [2]shatendradubey@gmail.com

* Corresponding Author: Roli Shrivastava

Abstract: *Sentiment analysis is an emerging trend nowadays to understand people's sentiments in multiple situations in their quotidian life. Social media data would be utilized for the entire process i.e the analysis and Classification processes and it consists of text data and emoticons, emoji's, etc. Many experiments were conducted in the antecedent studies utilizing Binary and Ternary Classification whereas Multi-class Classification gives more precise and precise Classification. In Multi-class Classification, the data would be divided into multiple sub-classes predicated on the polarities. Machine Learning and Deep Learning Techniques would be utilized for the Classification process. Utilizing Social media, sentiment levels can be monitored or analysed. This paper shows a review of the sentiment analysis on Social media data for apprehensiveness or dejection detection utilizing various machine learning techniques. Multi Class Classification with machine Learning Algorithm shows higher precision value during the sentiment analysis.*

Keywords: *Sentiment analysis, social network analysis, Multiclass classification, Emoticons & Emoji's, Machine learning.*

1. Introduction

In the last few decades, the number of users on social media has grown exponentially. Facebook has 1.28 billion active users and Twitter has 241 million active users [1]. These platforms serve as the hub for people to express their opinions. Thus, a massive amount of data is provided by Social Media platforms, which indicate the behaviour and sentiment of the users. Sentiment Analysis (SA) of this data can be beneficial in many domains. SA has been a focus of many researchers owing to its vast applications in the industry [2]. It brings forth an automated way to mine the opinions and sentiments of the target audience. Organizations and companies can gain financial benefits by analysing the sentiments of consumers. The government can understand the public's perceptions and sentiments regarding their policies [3]. While the common supervised ML models namely Decision Tree, Support Vector Machine (SVM), and Naïve Bayes ML algorithms have been implemented widely for different tasks in multiple domains, such as data privacy, and healthcare security assessment [4,5]. The most naïve approach to SA is to classify the text into binary classes i.e., positive and negative. Binary ML classifiers yield high accuracy [6]. However, multi-class classification proves to be more beneficial as it provides an in-depth insight into the sentiments. Thus, the decision-makers are better equipped to make accurate decisions. Aspect-based SA is another tool of paramount importance that gives the awareness of the users' sentiment regarding a specific aspect of a particular entity. The problem arises when baseline ML algorithms fail to give an optimal accuracy for multi-class classification problems. Naïve Bayes and SVM are the most commonly used algorithms for SA [7,8]. These algorithms cease to perform optimally in a social media text. Degradation in performance arises because social media text frequently contains erroneous spellings, domain-centric slang, spam reviews, and negation in sentences. Due to these challenges, the SA of social media becomes a challenging task. Many researchers have put in efforts to enhance the performance of the multi-class classifier. To overcome this challenge, various architectures have been proposed. However, on account of the complexity of the social media text, the data embodies high-level features and is not linearly separable. Thus, the baseline ML models are incapable of completely capturing and learning these features. However, more complex deep learning (DL) models such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and Bidirectional Encoder Representation of Transformer (BERT) can perform more optimally [9,10]. Besides the model selection and designing an optimal architecture, pre-processing of the textual data plays a key role in augmenting the overall performance [11]. Thus, effective preprocessing of the social media text has been a focus of many researchers. The baseline pre-processing of textual data involves a number of crucial steps: 1) stop words elimination, 2) tokenization 3) word normalization, and 4) vectorization [8]. Which language model i.e., unigram, bigram, or trigram fits best the particular NLP task at hand and comes under the umbrella of pre-processing techniques.

2. Related Work

2.1 Based on Machine Learning Techniques

Chen et al.[1] proposed a deep sentiment representation model based on CNNs and long short-term memory recurrent neural network (LSTM). The model uses two layers of CNNs to capture the partial features of the text. The model can capture more accurate partial features, after which the features are fed to the LSTM, which can capture the contextual information. Finally, we combine the improved deep learning model with a one-versus-rest training mechanism and apply it to multi-class sentiment classification. Experimental results demonstrate that the model we designed for multi-class sentiment analysis achieves a accuracy of 78.42% on data set D1 is better than the existing SVMs (support vector machines), CNN, LSTM and CNN-LSTM.

Ruz et al.[2] adopt a Bayes factor approach for this purpose, yielding more realistic networks. The results show the effectiveness of using the Bayes factor measure as well as its competitive predictive results when compared to support vector machines and random forests, given a sufficient number of training examples. Also, the resulting networks allow to identify the relations amongst words, offering interesting qualitative information to historically and socially comprehend the main features of the event dynamics.

Arora et al.[3] tried to analyze health tweets for Depression, Anxiety from the mixed tweets by using Multinomial Naive Bayes and Support Vector Regression (SVR) Algorithm as a classifier.

Katchapakirin et al. [4] employs Natural Language Processing (NLP) techniques to develop a depression detection algorithm for the Thai language on Facebook where people use it as a tool for sharing opinions, feelings, and life events. Results from 35 Facebook users indicated that Facebook behaviours could predict depression level.

Kumar et al.[5] proposed h a novel model, AD prediction model, for anxious depression prediction in real-time tweets. The model is trained using three classifiers (multinomial naïve bayes, gradient boosting, and random forest) and majority voting using an ensemble voting classifier is done. Preliminary results are evaluated for tweets of sampled 100 users and the proposed model achieves a classification accuracy of 85.09%.

Cornn et al.[6] developed machine learning (logistic regression, support vector machines), a BERT-based model, and neural networks with and without word embeddings (CNN) for this classification task. It was found that the CNN model without word embeddings performed the best, with approximately 92.5% accuracy after 4 epochs, followed by the BERT-based model with 85.7% accuracy.

Ahmad et al.[7] aims at proposing a terrorism-related content analysis framework with the focus on classifying tweets into extremist and non-extremist classes. Based on user-generated social media posts on Twitter, we develop a tweet classification system using deep learning-based sentiment analysis techniques to classify the tweets as extremist or non-extremist. The experimental results are encouraging and provide a gateway for future researchers.

Giakwad et al.[8] proposed a new approach that uses lexicon database to assign each word in a text a value called as `Impact Factor'. The Impact Factor is nothing but how a single word is affecting the whole sentence in which it is used. Every word in a sentence has its own Impact Factor and it tries to influence the overall semantic of the sentence. Higher the value of Impact Factor of a word in the sentence, the more influential it is. The approach proposed in this paper makes use of lexicon based approach as well as machine based learning. It uses AFINN lexicon database to assign Impact Factor to words and Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Naive Bayesian (NB) machine learning algorithms for training and testing the model.

Jayakrishnan et al.[9] an SVM classifier is used for sentence level multi-class emotion detection in Malayalam. The proposed approach uses different syntactic features such as n-gram, POS related, negation related, level related features etc, for better classification. The classifier classifies the Malayalam sentences into different emotion classes like happy, sad, anger, fear or normal etc. with level information such as high, low etc. It also states whether the sentence is dialogue, question or not for better hearing experience from a speech synthesiser while reading the novel.

Solakidis et al.[10] developed an integrated framework that accepts user generated documents and then identifies the polarity of the text (neutral, negative or positive) and the sentiment expressed through it (joy, love, anger or sadness). We followed a semi-supervised approach which led to the development of two techniques for the automatic collection of training data without any human intervention. Our approach involves the detection and use of self-defining features that are available within the data. We take into account two emotionally rich features: a) emoticons and b) lists of emotionally intense keywords. These features are evaluated on data coming from a popular forum, using various classifiers and feature vectors. Our experimental results point to various conclusions about the effectiveness, advantages and limitations of applying such methods on Greek data. Using keywords we achieved 90% mean accuracy on identifying the subjectivity level and 93% on correctly identifying the polarity level, whereas using emoticons the mean accuracy for each of these levels was 74% and 77% respectively.

Deshpande et al.[11] aims to apply natural language processing on Twitter feeds for conducting emotion analysis focusing on depression. Individual tweets are classified as neutral or negative, based on a curated word-list to detect depression tendencies. In the process of class prediction, support vector machine and Naive-Bayes classifier have been used. The results have been presented using the primary classification metrics including F1-score, accuracy and confusion matrix.

Chatterjee et al.[12] The World Health Organization (WHO) suggests that people who are depressed are regularly not correctly diagnosed and others s who are misdiagnosed are prescribed antidepressants. Thus, there is a strong need to automatically assess the risk of depression. Social media platforms increasingly come closer to become a true digitization of the human social experience. In many cases people would in fact prefer to express themselves online than offline.

Jose et al.[13] propose a novel approach for accurate sentiment classification of twitter messages using lexical resources SentiWordNet and WordNet along with Word Sense Disambiguation. Thus we applied the SentiWordNet lexical resource and Word Sense Disambiguation for finding political sentiment from real time tweets. Our method also uses a negation handling as a pre-processing step in order to achieve high accuracy.

Table 1. Comparison of various machine learning techniques

Ref	Classifier	Accuracy (%)	Dataset
[1]	SVM	74.18	Twitter
[2]	Naive Bayes	74.2	Twitter
[3]	Multinomial Naive Bayes	78.0	Twitter
[4]	SVM	68.57	Facebook
[5]	Random Forest	81.04	Twitter
[6]	Logistic Regression	84.8	Reddit
[7]	KNN	72.0	Twitter
[8]	SVM	82.0	Twitter
[9]	SVM	90.0	Various sources
[10]	Multinomial Naive Bayes	92.2	Twitter
[11]	Multinomial Naive Bayes	83.0	Twitter
[12]	Naive Bayes	76.6	Facebook
[13]	Naive Bayes	71.46	Twitter

2.2 Based on Multi-class Sentiment Analysis

Tanna et al.[14] aims to automate this task of analyzing the reactions and the posts and generate a report based on the outcome. A custom social media platform is proposed that would give the ability to the users to perform activities such as post, like, comment, and share. Also, the proposed platform can be used to share content on other social media platforms as well. As a result, the users would have a single platform with the ability to do a lot more than any existing social media platform facilitates.

Arora et al.[15] tried to analyze health tweets for Depression, Anxiety from the mixed tweets by using Multinomial Naive Bayes and Support Vector Regression (SVR) Algorithm as a classifier.

Asad et al.[16] propose a data-analytic based model to detect depression of any human being. In this research, machine learning is used to process the scrapped data collected from SNS users. Natural Language Processing (NLP), classified using Support Vector Machine (SVM) and Naïve Bayes algorithm to detect depression potentially in a more convenient and efficient way.

Lyu et al.[17] apply sentiment analysis in text-mining technology to analyze the textual data of social media Weibo, explore people's attitudes toward child abuse incidents and the reasons behind them, and discuss how their emotions will become the potential driving force for improving China's child protection policies. This analysis may serve as a potential impetus for the government to develop child protection policies and build a robust child welfare system.

Tariq et al.[18] propose a methodology to classify the patients associated with chronic mental illness diseases (i.e. Anxiety, Depression, Bipolar, and ADHD (Attention Deficit Hyperactivity Disorder) based on the data extracted from the Reddit, a wellknown network community platform.The experimental results indicate the effectiveness of Co-training based classification rather than the state of the art classifiers by a margin of 3% on average in par with every state of art technique.

Jabreel et al.[19] development of a novel deep learning-based system that addresses the multiple emotion classification problem in Twitter. We propose a novel method to transform it to a binary classification problem and exploit a deep learning approach to solve the transformed problem. Our system outperforms the state-of-the-art systems, achieving an accuracy score of 0.59 on the challenging SemEval2018 Task 1:E-cmulti-label emotion classification.

Bouzazi et al.[20] propose a novel approach that, in addition to the aforementioned tasks of binary and ternary classification, goes deeper in the classification of texts collected from Twitter and classifies these texts into multiple sentiment classes. While in this work, we limit our scope to 7 different sentiment classes, the proposed approach is scalable and can be run to classify texts into more classes. Nevertheless, the approach proves to be very accurate in binary classification and ternary classification: in the former case, we reach an accuracy of 81.3% for the same dataset used after removing neutral tweets, and in the latter case, we reached an accuracy of classification of 70.1%.

Imran et al.[21] analyze reaction of citizens from different cultures to the novel Coronavirus and people's sentiment about subsequent actions taken by different countries. Deep long short-term memory (LSTM) models used for estimating the sentiment polarity and emotions from extracted tweets have been trained to achieve state-of-the-art accuracy on the sentiment140 dataset. The use of emoticons showed a unique and novel way of validating the supervised deep learning models on tweets extracted from Twitter.

Yao et al.[22] proposes a domain-specific sentiment analysis approach specifically for tweets posted during hurricanes (DSSA-H). DSSA-H can retrieve hurricane-relevant tweets with a trained supervised-learning classifier, Random Forest (RF), and classify the sentiment of hurricane-relevant tweets based on a domain-adversarial neural network (DANN). DSSA-H contributes to an intelligent and adaptive disaster information system for the data-rich human and the built environment system.

Li et al.[23] introduce HEMOS (Humor-EMOji-Slang-based) system for fine-grained sentiment classification for the Chinese language using deep learning approach. We investigate the importance of recognizing the influence of humor, pictograms and slang on the task of affective processing of the social media. Our experimental results show that the

proposed method can significantly improve the state-of-the-art methods in predicting sentiment polarity on Weibo, the largest Chinese social network.

Harb et al.[24] describe an analysis framework to investigate the emotional reaction on Twitter to mass traumatic events and use it to derive conclusions about eight mass shooting events. The framework encompasses the crawling of pre/post-events tweets to compare the emotional reactions, the classification of the sentiment in terms of Ekman's basic emotions, and the use of data extracted from Twitter users' profiles to understand these reactions in the light of users' demographics (age and gender), proximity to the event, and the number of victims. This analysis framework could be used to derive conclusions on the emotional reactions to all sorts of mass traumatic events.

3. Sentiment Analysis

The classification of the block of text is whether positive, negative, or neutral, is called Sentiment Analysis. The main aim will be to analyze people's interest in a way that it can help the businesses expand. It depicts not only on polarity (positive, negative neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms. It is the contextual mining of words that indicates the social sentiment of a brand. It also helps to determine the business whether the product in which they are manufacturing is going to make a demand in the market or not. Figure 1 shows the necessary sentiment analysis steps.

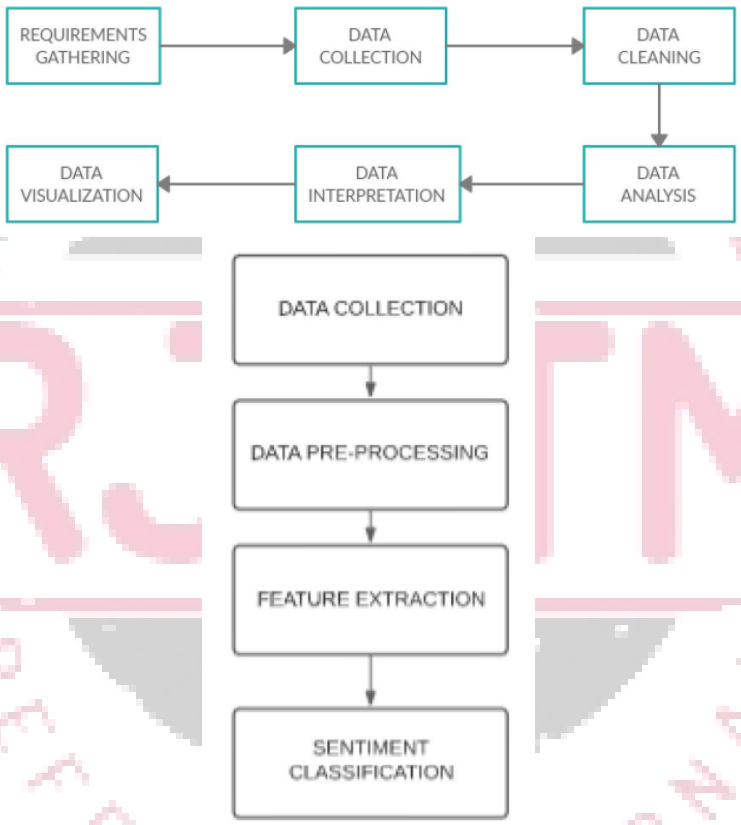

Fig. 1 Sentiment analysis steps

Two techniques that involves in Sentiment analysis are:

– Rule-based sentiment analysis It uses rules and a collection of words labelled by polarity to identify the text opinion. Sentiment value typically need to be combined with extra powers to understand sentences containing sarcasm, negations, or dependent clauses.

– Machine learning-based sentiment analysis Involves training a Machine Learning model to understand the polarity based on the word order using a sentiment labelled training set.

4. Social Network Analysis

Social network analysis means the process of identifying and realizing the relationships and data few between people, groups, organizations, computers and other connected information entities. The network nodes are the groups and people, whereas the links show relationships between the respective nodes. It performs analysis like visual and mathematical analysis of human relationships. Researchers measures the activity of network for a node that involves the concept of degrees—the total number of connections a node has directly. A centralized network is identified by one or a few nodes known as central nodes. Central nodes are damaged or removed if the network quickly fragments into unconnected sub-networks. It can become a single point of failure. A centralized system around a well-connected hub can fail if that hub is removed or disabled. Hubs are nodes with a high degree centrality.

5. Conclusion

Sentiment analysis/Opinion Mining understands the feelings, replications as well as judgements amassed or extracted from texts or other data utilized in data analysis or mining, web mining, and convivial media analytics because sentiments are to judge human comportment. They can be categorized into positive, negative, or neutral. It discovers opinions, then convey the posture and categorize them division-sapient. The data amassed in the process, apperceiving their sentiments, culling features, relegating sentiments and conclusively calculating the sentiment polarity. In Multi class Classification, the data was classified into many subclasses predicated on the sentiment polarity where we can expect a precise or precise classification. Social media data additionally consists of emoticons and emoji's, where they withal have sentiment score values as they additionally can be utilized for the sentiment analysis process or classification. The foremost paramount step is to ascertain that not to lose the emoticons during the pre-processing data stage since it contains the sentiment value. It can withal be utilized for sentiment analysis. To extract the features from the pre-processed data, feature extraction techniques are utilized. Utilizing classifcation procedure, the data was polarized into sentiment classes predicated on the sentiment values which was done by sundry machine learning and deep learning algorithms.

References

[1] B. Chen, Q. Huang, Y. Chen, L. Cheng, and R. Chen, "Deep Neural Networks for Multi-class Sentiment Classification," Proc. - 20th Int. Conf. High Perform. Comput. Commun. 16th Int. Conf. Smart City 4th Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2018, pp. 854–859, Jun. 2018, doi: 10.1109/HPCC/SMARTCITY/DSS.2018.00142.

[2] P. Arora and P. Arora, "Mining Twitter Data for Depression Detection," 2019 Int. Conf. Signal Process. Commun. ICSC 2019, pp. 186–189, Mar. 2019, doi: 10.1109/ICSC45622.2019.8938353.

[3] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," Futur. Gener. Comput. Syst., vol. 106, pp. 92–104, 2020, doi: https://doi.org/10.1016/j.future.2020.01.005.

[4] K. Katchapakirin, K. Wongpatikaseree, P. Yomaboot, and Y. Kaewpitakkun, "Facebook Social Media for Depression Detection in the Thai Community," 2018 15th Int. Jt. Conf. Comput. Sci. Softw. Eng., 2018, doi: 10.29007/TSCC.

[5] A. Kumar, A. Sharma, and A. Arora, "Anxious Depression Prediction in Real-time Social Data A R T I C L E I N F O", Accessed: Aug. 04, 2023. [Online]. Available: http://ssrn.com/link/ICAESMT-2019.html=xyz

[6] K. Cornn, "Identifying Depression on Social Media".

[7] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Human-centric Comput. Inf. Sci., vol. 9, no. 1, pp. 1–23, Dec. 2019, doi: 10.1186/S13673-019-0185-6/TABLES/12.

[8] G. Gaikwad and D. J. Joshi, "Multiclass Mood classification on twitter using lexicon dictionary and machine learning algorithms," Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 1, Jul. 2016, doi: 10.1109/INVENTIVE.2016.7823247.

[9] G. S. Solakidis, K. N. Vavliakis, and P. A. Mitkas, "Multilingual Sentiment Analysis Using Emoticons and Keywords," 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., vol. 2, pp. 102–109, Oct. 2014, doi: 10.1109/WI-IAT.2014.86.

[10] R. Jayakrishnan, G. N. Gopal, and M. S. Santhikrishna, "Multi-Class Emotion Detection and Annotation in Malayalam Novels," 2018 Int. Conf. Comput. Commun. Informatics, ICCCI 2018, Aug. 2018, doi: 10.1109/ICCCI.2018.8441492.

[11] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 858–862. doi: 10.1109/ISS1.2017.8389299.

[12] R. Chatterjee, R. K. Gupta, and B. Gupta, "Depression Detection from Social Media Posts Using Multinomial Naive Theorem," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012095.

[13] R. Jose and V. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," 2016, pp. 64–67. doi: 10.1109/SAPIENCE.2016.7684133.

[14] S. Tariq et al., "A Novel Co-Training-Based Approach for the Classification of Mental Illnesses Using Social Media Posts," IEEE Access, vol. 7, pp. 166165–166172, 2019, doi: 10.1109/ACCESS.2019.2953087.

[15] D. Tanna, M. Dudhane, A. Sardar, K. Deshpande, and N. Deshmukh, "Sentiment Analysis on Social Media for Emotion Classification," 2020 4th Int. Conf. Intell. Comput. Control Syst., pp. 911–915, May 2020, doi: 10.1109/ICICCS48265.2020.9121057.

[16] P. Arora and P. Arora, "Mining Twitter Data for Depression Detection," 2019 Int. Conf. Signal Process. Commun., pp. 186–189, Mar. 2019, doi: 10.1109/ICSC45622.2019.8938353.

[17] N. Al Asad, M. A. Mahmud Pranto, S. Afreen, and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," 2019 IEEE Int. Conf. Signal Process. Information, Commun. Syst. SPICSCON 2019, pp. 13–17, Nov. 2019, doi: 10.1109/SPICSCON48833.2019.9065101.

[18] Y. W. Lyu, J. C. C. Chow, and J. J. Hwang, "Exploring public attitudes of child abuse in mainland China: A sentiment analysis of China's social media Weibo," Child. Youth Serv. Rev., vol. 116, Sep. 2020, doi: 10.1016/J.CHILDYOUTH.2020.105250.

[19] M. Jabreel and A. Moreno, "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets," Appl. Sci. 2019, Vol. 9, Page 1123, vol. 9, no. 6, p. 1123, Mar. 2019, doi: 10.3390/APP9061123.

[20] M. Bouazizi and T. Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter," IEEE Access, vol. 5, pp. 20617–20639, Aug. 2017, doi: 10.1109/ACCESS.2017.2740982.

[21] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets," Ieee Access, vol. 8, p. 181074, 2020, doi: 10.1109/ACCESS.2020.3027350.

[22] F. Yao and Y. Wang, "Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach," Comput. Environ. Urban Syst., vol. 83, p. 101522, Sep. 2020, doi: 10.1016/J.COMPENVURBSYS.2020.101522.

[23] D. Li, R. Rzepka, M. Ptaszynski, and K. Araki, "HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media," Inf. Process. Manag., vol. 57, no. 6, Nov. 2020, doi: 10.1016/J.IPM.2020.102290.

[24] J. G. D. Harb, R. Ebeling, and K. Becker, "A framework to analyze the emotional reactions to mass violent events on Twitter and influential factors," Inf. Process. Manag., vol. 57, no. 6, Nov. 2020, doi: 10.1016/J.IPM.2020.102372.