# Sentiment Polarity and Emotion Detection from Tweets Using Distant Supervision and Deep Learning Models

Muhamet Kastrati[1], Marenglen Biba[1], Ali Shariq Imran[2], and Zenun Kastrati[3]

[1] Department of Computer Science, University of New York Tirana, Tirana, Albania
muhamet.kastrati@gmail.com, marenglenbiba@unyt.edu
[2] Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
ali.imran@ntnu.no
[3] Department of Informatics, Linnaeus University, 351 95 Växjö, Sweden
zenun.kastrati@lnu.se

**Abstract.** Automatic text-based sentiment analysis and emotion detection on social media platforms has gained tremendous popularity recently due to its widespread application reach, despite the unavailability of a massive amount of labeled datasets. With social media platforms in the limelight in recent years, it's easier for people to express their opinions and reach a larger target audience via Twitter and Facebook. Large tweet postings provide researchers with much data to train deep learning models for analysis and predictions for various applications. However, deep learning-based supervised learning is data-hungry and relies heavily on abundant labeled data, which remains a challenge. To address this issue, we have created a large-scale labeled emotion dataset of 1.83 million tweets by harnessing emotion-indicative emojis available in tweets. We conducted a set of experiments on our distant-supervised labeled dataset using conventional machine learning and deep learning models for estimating sentiment polarity and multi-class emotion detection. Our experimental results revealed that deep neural networks such as BiLSTM and CNN-BiLSTM outperform other models in both sentiment polarity and multi-class emotion classification tasks achieving an F1 score of 62.21% and 39.46%, respectively, an average performance improvement of nearly 2-3 percentage points on the baseline results.

**Keywords:** Sentiment polarity · Emotion detection · Distant supervision · Emoji · Deep learning · Twitter · Classification.

## 1 Introduction

Nowadays, microblogging and social networks are highly influential in a wide range of settings, from daily communication, sharing ideas, opinions, emotions and reactions with others, shopping behavior, political issues, and reacting to crises, just to name a few [1]. Over the past few years, researchers have shown a growing interest in text-based emotion detection on online social networks, notably Twitter and Facebook. The huge amount of text generated by Twitter users is a rich source to obtain people's emotions, which are an integral part of human life and have a strong influence on people's behaviors and actions [2].

Emotion detection from text is a sub-field of sentiment analysis that aims to extract and analyse emotions that can be explicit or implicit in the sentence [3]. While sentiment analysis is concerned with classifying sentiments as positive, negative, or neutral, emotion detection on the other hand deals with extracting fine-grained emotions such as anger, disgust, fear, joy, sadness, and surprise.

There are various learning approaches used to detect emotions in text, including the lexicon-based approach [4], the rule-based approach [5], the machine learning-based approach [6–9], and the deep learning-based approach [10–14].

Machine learning and deep learning models are widely employed to build sentiment analysis and emotion recognition systems [14–16]. More recently, deep neural networks such as CNN and RNN (including its variants LSTM and GRU) have gained popularity due to the state-of-the-art performance obtained on various natural language processing (NLP) tasks [17]. Supervised learning is the most widely used approach in machine learning, including deep and shallow learning [18]. However, training supervised learning models require a large amount of human-labeled data, which is not the case for several real-world applications, and text emotion detection is no exception [7].

To address this issue, we have collected a large-scale emotion dataset of tweets from Twitter. Inspired by the research study conducted in [19], emotion-indicative emojis are used for automatic labeling of the dataset. Then, several supervised conventional machine learning algorithms and deep learning models are tested on the newly collected dataset to establish the baseline results and examine an approach on sentiment polarity and emotion detection that better suits the dataset in order to improve the performance of the classifier models.

The core contributions of this work are:

- Collecting and curating a real-world large-scale dataset of tweets that are automatically labelled with categorical emotions based on Ekman's model [20] employing distant supervision using emotion indicative emojis.
- New knowledge with regard to performance comparison of supervised conventional machine learning algorithms and deep neural networks for sentiment polarity classification and emotion detection on our created dataset.

The rest of the paper is organized as follows: Section 2 presents related work on emotion analysis and approaches used for dataset creation. Section 3 presents the research method followed by an overview of the experimental settings provided in Section 4. Section 5 presents the results and analysis, while conclusions and directions for future work are given in Section 6.

## 2 Related Work

During the past decade, several studies have been conducted with regard to the sentiment analysis tasks in Twitter posts. Most of these studies can generally be grouped into two main research directions based on their core contributions: i) data curation/labeling techniques for sentiment analysis tasks, ii) polarity/emotion classification. The first group entails studies concerning data collection and (semi)automatic labeling techniques. For instance, the research work conducted in [21], introduced for the first time distant supervision labels (emoticons) for classifying the sentiment polarity of tweets. The study presents one of the most widely used Twitter sentiment datasets for sentiment analysis tasks known as Sentiment140. Another similar study which uses distant supervision strategy for automatic labeling is presented in [22]. In particular, hashtags and text emoticons for sentiment annotation are applied in both studies to generate labels. A similar study that applies emojis as distantly supervised labels to detect Plutchik's emotions is conducted in [23].

There is another strand of research which focuses on creating datasets for emotion detection task. For example, the research study in [6] presents Twitter Emotion Corpus annotated using distant supervision with emotion-specific hashtags for emotion annotation. An extended dataset called Tweet Emotion Intensity dataset is presented later in [8] where the authors created the first dataset of tweets annotated for anger, fear, joy, and sadness intensities using best–worst scaling technique. The researchers in [24] present the first emoji sentiment lexicon, known as the Emoji Sentiment Ranking as well as a sentiment map that consists of 751 most frequently used emojis. The sentiment of the emojis is computed from the sentiment of the tweets in which they occur. A similar work is conducted in [19] where a large-scale dataset of tweets in Urdu language for sentiment and emotion analysis is presented. The dataset is automatically annotated with distant supervision using emojis. A list of 751 most frequently used emojis are applied for annotation.

The second group of research works focuses on polarity and emotion classification using conventional machine learning algorithms and deep neural networks. Such a study is conducted in [12], where the authors proposed a classification approach for emotion detection from text using deep neural networks including Bi-LSTM, and CNN, with self-attention and three pre-trained word-embeddings for words encoding. Another similar example where LSTM models are used for estimating the sentiment polarity and emotions from Covid-19 related tweets is proposed in [14] and in [25]. The later study also introduced a new approach employing emoticons as a unique and novel way to validate deep learning models on tweets extracted from Twitter. Another study focusing on emotion recognition using both emoticon and text with LSTM is conducted in [13].

## 3 Design and Research Methodology

This study uses a quantitative research approach composed of five major phases. The first phase entails the collection of emoji tweets on Twitter, belonging to the time period from 01 January until 31 December 2021. To be able to collect enough tweets to meet our needs, we selected 41 emojis indicative of the emotion used in [19] and collected tweets that contained at least one of the selected emojis, and only those tweets that were tagged by Twitter as English (retweets excluded). In the second phase of the study, a text pre-processing is performed to remove extra attributes related to tweets (author id, date of creation, language, source, etc.), duplicate tweets, extract emojis from tweets, remove hashtags/mentions, URLs, emails, phone number, non-ASCII characters and tweets with length less or equal to five characters. Additionally, all tweets were converted to lowercase. In the third phase, automatic labeling of collected tweets is carried out through distant supervision using emotion-indicative emojis. Consequently, all emoji tweets are properly classified into one of Ekman's six basic categorical emotions, including anger, disgust, fear, joy, sadness, sadness, or surprise. In the fourth phase, a representation model to prepare and transform the tweets to an appropriate numerical format to be fed into the emotion classifiers is performed. A bag of word representation model with its implementation, term frequency inverse document frequency ($tf - idf$) is employed.

The final phase of the study involves the sentiment analyser (binary classification) and the emotion analyser for the multi-class classification of tweets along the six basic categorical emotions, namely anger, disgust, fear, joy, sadness, or surprise. The analyser involves several classifiers including conventional machine learning algorithms and deep neural networks for emotion detection. A high-level pipeline of the proposed sentiment and emotion analyser depicting all the five phases elaborated above is illustrated in Figure 1.

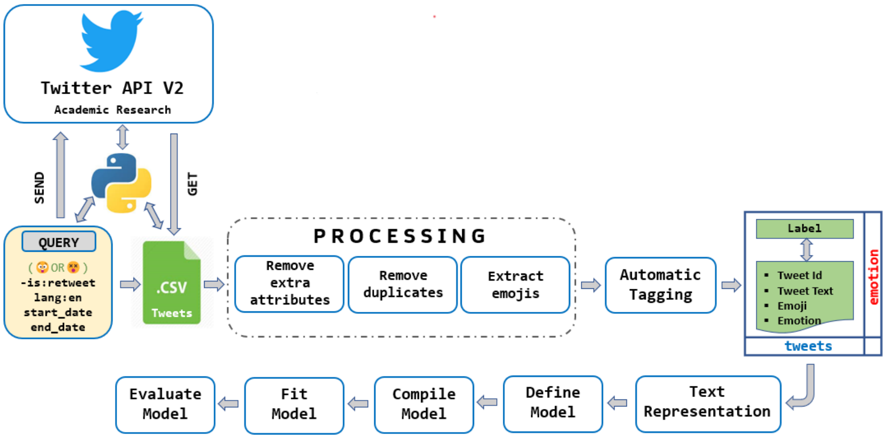

**Fig. 1.** High-level pipeline of the proposed solution.

## 4 Experimental Settings

This section briefly describes the dataset (emoji tweets) as well as the classifier models used to perform the sentiment and emotion classification task.

### 4.1 Dataset

The dataset consists of 1,832,279 tweets posted between January 1 and December 31, 2021, with the same distribution of tweets every day. The whole data collection process was conducted through Twitter API v2 for academic research product track using Python 3. The dataset is balanced for sentiment (51% for positive and 49% for negative), but is imbalanced for emotion, and its statistics are given in Table 1.

**Table 1.** Dataset statistics

| Sentiment Polarity | # of instances | % | Emotion | # of instances | % |
|---|---:|---:|:---:|---:|---:|
| Positive | 934,435 | 51 | Joy | 547,047 | 30 |
| | | | Surprise | 387,388 | 21 |
| Negative | 897,844 | 49 | Sadness | 298,742 | 16 |
| | | | Disgust | 207,838 | 11 |
| | | | Anger | 207,514 | 11 |
| | | | Fear | 183,750 | 10 |
| **Total** | **1,832,279** | **100** | **Total** | **1,832,279** | **100** |

### 4.2 Conventional Machine Learning Models

The conventional machine learning models employed in this study for sentiment and emotion classification include Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and AdaBoost, as they are known for their good performance [26] and efficiency even for handling millions of tweets [27]. All the algorithms are trained in scikit-learn library in Jupyter Notebook in Anaconda, with default values for all parameters for all classifiers.

### 4.3 Deep Neural Networks

We selected DNN, BiLSTM, CNN (Conv-1D), GRU and CNN-BiLSTM combined, as these models are well known for their state-of-the-art performance in almost all NLP tasks, including sentiment and emotion analysis [12,17,28,29]. All these models are trained and tested in google colab using Keras Python library for deep learning using the TensorFlow backend. Table 2 presents various deep neural networks along with their model configurations as well as their accuracy obtained on the test set (on 10% test data) for each of the models.

## 5 Results and Analysis

We conducted a set of experiments to investigate the performance of both conventional machine learning and deep learning models on the classification of

**Table 2.** Configuration of the model and accuracy of the deep learning models tested.

| Classifier | Model Configuration / Parameters | Sentiment Polarity | Emotion Detection |
|---|---|---|---|
| DNN | Embedding Layer with 100 Dimension,<br>GlobalMaxPooling1D,<br>Layers with 128, 64, 32 with ReLU,<br>(Dense 2 with Sigmoid)/Dense 6 with Softmax. | 61.61% | 38.40% |
| CNN (1D) | Embedding Layer with 100 Dimension,<br>Layers with 64, 32 with ReLU,<br>GlobalMaxPooling1D,<br>Dense 32 with ReLU,<br>(Dense 2 with Sigmoid)/Dense 6 with Softmax. | 60.29% | 38.20% |
| BiLSTM | Embedding Layer with 100 Dimension,<br>BiLSTM Layers with 32, 32 with ReLU,<br>GlobalMaxPooling1D, Dense 10 with ReLU,<br>(Dense 2 with Sigmoid)/Dense 6 with Softmax. | 62.06% | 39.69% |
| GRU | Embedding Layer with 100 Dimension,<br>GRU Layers with 32, 32 with ReLU,<br>GlobalMaxPooling1D, Dense 10 with ReLU,<br>(Dense 2 with Sigmoid)/Dense 6 with Softmax. | 62.11% | 39.38% |
| CNN-BiLSTM | Embedding Layer 100, SpatialDropout1D(0.3),<br>Conv1D with 32 with ReLU,<br>BiLSTM with 32 with ReLU,<br>Flatten layer, Dense 64 with ReLU,<br>(Dense 2 with Sigmoid)/Dense 6 with Softmax. | 62.20% | 39.27% |

sentiment polarity and emotions task. The following parameter settings are used to conduct experiments. Dataset is divided in two sets: training and test sets, with 10% samples used for testing the model. Model training was set to 50 epochs and the *'EarlyStopping'* criteria with its arguments: *monitor*="*val_loss*" and *patience* = 3, is used to stop classifiers. The batch size of 2048 gave us the best result.

The findings, illustrated in Figure 2 and Figure 3, show that the best performance with regard to F1 score is achieved by deep learning models on both sentiment polarity and emotion classification tasks.

Class-wise performance with respect to F1 score for the task of sentiment polarity classification is shown in Figure 2. For the sake of space, we present results obtained from only two best performing models, including one from conventional machine learning (Logistic Regression) and one from deep learning (CNN-BiLSTM). The results show that CNN-BiLSTM generally outperforms the Logistic Regression model in sentiment polarity classification achieving an F1 score of 62.21%. It is interesting to note that BiLSTM slightly performs better than Logistic Regression, achieving an F1 score of 59.85% for the negative class and 64.28% for the positive class. This slight improvement is accounted to the network architecture and it might be higher if more complex architectures would have been used to train the BiLSTM.

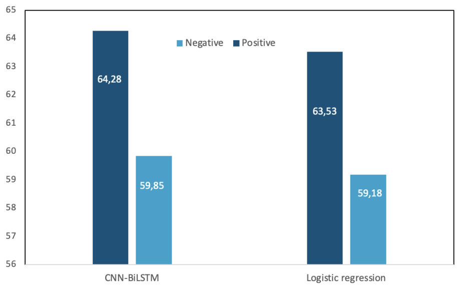

**Fig. 2.** Performance of Logistic Regression and CNN-BiLSTM for sentiment polarity

Next, we examined the class-wise performance of classifiers on the task of emotion classification. The obtained results from two best performing models, one from conventional machine learning (SVM) and one from deep learning (BiLSTM) are illustrated in Figure 3. The result show that BiLSTM generally outperforms the SVM (LinearSVC) model in multi-class emotion classification achieving an F1 score of 39.46%. It is worth pointing out that a better performance is achieved by BiLSTM at all classes of emotions.

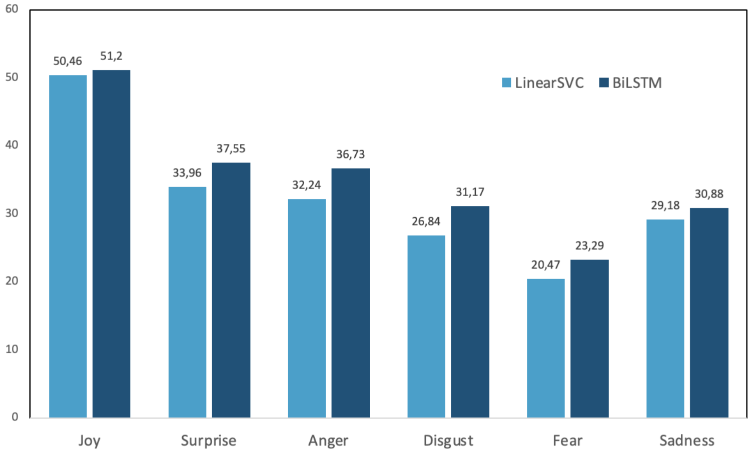

**Fig. 3.** Performance of LinearSVC and BiLSTM for emotion detection

**Sentiment assessment**. The next round of experiments is conducted to investigate the performance of various classifiers on the task of sentiment polarity classification. The results summarized in Table 3 show that a better performance is achieved by deep learning classifiers. In particular, the combined CNN-BiLSTM architecture slightly outperforms the other deep learning models achieving an F1 score of 62.21%.

**Table 3.** Performance of ML and DL models for sentiment polarity assessment

| Classifier | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| Naïve Bayes | 61.34% | 61.33% | 61.33% | 61.33% |
| Logistic Regression | 61.49% | 61.48% | 61.48% | 61.48% |
| SVM | 61.42% | 61.42% | 61.42% | 61.42% |
| Decision Tree | 64.11% | 51.63% | 57.20% | 51.63% |
| AdaBoost | 57.10% | 55.42% | 56.25% | 55.42% |
| DNN | 61.60% | 61.61% | 61.60% | 61.61% |
| CNN | 60.39% | 60.29% | 60.34% | 60.29% |
| BiLSTM | 62.09% | 62.06% | 62.07% | 62.06% |
| GRU | 62.13% | 62.11% | 62.12% | 62.11% |
| CNN-BiLSTM | 62.22% | 62.20% | 62.21% | 62.20% |

**Emotion Recognition.** Once the sentiment polarity has been assessed, in the second step, we identify emotions in tweets. In order to extract tweet emotions, we run the same experiments conducted for sentiment polarity assessment, except for the number of classes which here is different, 6 classes. The performance of five conventional machine learning and five deep learning models was tested for the multi-class emotion classification task. Table 4 shows precision, recall, F1 score, and accuracy obtained from these classifiers in our dataset. The empirical findings reveal that deep learning models perform slightly better than conventional machine learning ones. More precisely, the BiLSTM architecture slightly outperforms the other deep learning models achieving an F1 score of 39.46%, compared to the best performing conventional machine learning algorithm (NB) which achieved an F1 score of 38.06% on the same task.

**Table 4.** Performance of conventional ML and DL models for emotion detection

| Classifier | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| Naïve Bayes | 38.84% | 37.32% | 38.06% | 37.32% |
| Logistic Regression | 37.81% | 37.93% | 37.87% | 37.93% |
| SVM | 37.16% | 37.97% | 37.56% | 37.97% |
| Decision Tree | 26.00% | 30.48% | 28.06% | 30.48% |
| AdaBoost | 34.52% | 32.07% | 33.25% | 32.07% |
| DNN | 37.43% | 38.40% | 37.91% | 38.40% |
| CNN | 37.26% | 38.20% | 37.72% | 38.20% |
| BiLSTM | 39.23% | 39.69% | 39.46% | 39.69% |
| GRU | 38.52% | 39.38% | 38.95% | 39.38% |
| CNN-BiLSTM | 38.85% | 39.27% | 39.06% | 39.27% |

## 6 Conclusion and Future Work

This article presented and evaluated the use of emotion-indicative emojis to automatically label a large corpus of tweets with basic categorical emotions they express using Ekman's model. Supervised conventional machine learning and deep learning models are used for both sentiment polarity and detection of emotions from users' tweets on the created dataset. The experimental results showed that the BiLSTM and the combined CNN-BiLSTM architecture outperform the other models with a slight difference in accuracy and F1 score. The findings demonstrate that there is a moderate correlation between emojis and emotion annotations in tweets. As future work, we will focus more on further increasing the size of the dataset as the deep neural networks benefit from the presence of a huge amount of samples. We will also focus on addressing the class imbalance in the dataset and experiment with filter options to further clean the dataset from problematic instances/tweets. Additionally, experimenting with larger deep learning architectures, pre-trained word embedding models, and attention mechanism, is interesting to be further investigated in the future.

## References

1. Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20(3):531–558, 2018.
2. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE, 2012.
3. Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.
4. Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.
5. Maria Krommyda, Anastatios Rigos, Kostas Bouklas, and Angelos Amditis. Emotion detection in twitter posts: a rule-based algorithm for annotated data acquisition. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 257–262. IEEE, 2020.
6. Saif M Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.
7. Ian Wood and Sebastian Ruder. Emoji as emotion tags for tweets. In *Proc. of the Emotion and Sentiment Analysis Workshop, Portorož*, pages 76–79, 2016.
8. Saif M Mohammad and Felipe Bravo-Marquez. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*, 2017.
9. Anam Yousaf, Muhammad Umer, Saima Sadiq, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi. Emotion recognition by textual tweets classification using voting classifier (lr-sgd). *IEEE Access*, 9:6286–6295, 2020.
10. Muhammad Abdul-Mageed and Lyle Ungar. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proc. of the 55th annual meeting of the association for computational linguistics*, pages 718–728, 2017.
11. Niko Colnerič and Janez Demšar. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, 11(3):433–446, 2018.
12. Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. A comparison of word-embeddings in emotion detection from text using bilstm,

cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68, 2019.

13. Juyana Islam, Sadman Ahmed, MAH Akhand, and N Siddique. Improved emotion recognition from microblog focusing on both emoticon and text. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 778–782. IEEE, 2020.
14. Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *Ieee Access*, 8:181074–181090, 2020.
15. Maryam Edalati, Ali Shariq Imran, Zenun Kastrati, and Sher Muhammad Daudpota. The potential of machine learning algorithms for sentiment classification of students' feedback on mooc. In *Proceedings of SAI Intelligent Systems Conference*, pages 11–22. Springer, 2021.
16. Shpetim Sadriu, Krenare Pireva Nuci, Ali Shariq Imran, Imran Uddin, and Muhammad Sajjad. An automated approach for analysing students feedback using sentiment analysis techniques. In *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pages 228–239. Springer, 2022.
17. Muhamet Kastrati and Marenglen Biba. A state-of-the-art survey on deep learning methods and applications. *International Journal of Computer Science and Information Security (IJCSIS)*, 19(7), 2021.
18. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
19. Rakhi Batra, Zenun Kastrati, Ali Shariq Imran, Sher Muhammad Daudpota, and Abdul Ghafoor. A large-scale tweet dataset for urdu text sentiment analysis. 2021.
20. Paul Ekman. Facial expression and emotion. *American psy*, 48(4):384, 1993.
21. Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.
22. Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, 2010.
23. Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer, 2013.
24. Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.
25. Rakhi Batra, Ali Shariq Imran, Zenun Kastrati, Abdul Ghafoor, Sher Muhammad Daudpota, and Sarang Shaikh. Evaluating polarity trend amidst the coronavirus crisis in peoples' attitudes toward the vaccination drive. *Sustainability*, 13(10):5344, 2021.
26. Zenun Kastrati and Ali Shariq Imran. Performance analysis of machine learning classifiers on improved concept vector space models. *Future Generation Computer Systems*, 96:552–562, 2019.
27. HS Hemanth Kumar, YP Gowramma, SH Manjula, D Anil, and N Smitha. Comparison of various ml and dl models for emotion recognition using twitter. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 1332–1337. IEEE, 2021.
28. Zenun Kastrati, Lule Ahmedi, Arianit Kurti, Fatbardh Kadriu, Doruntina Murtezaj, and Fatbardh Gashi. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10):1–19, 2021.
29. Bilal Ahmed Chandio, Ali Shariq Imran, Maheen Bakhtyar, Sher Muhammad Daudpota, and Junaid Baber. Attention-based RU-BiLSTM sentiment analysis model for roman urdu. *Applied Sciences*, 12(7):3641, 2022.