INTERNATIONAL JOURNAL OF ELECTRONICS AND COMPUTER APPLICATIONS

ORIGINAL ARTICLE

Article access online

OPEN ACCESS

Received: 08.09.2024

Accepted: 10.11.2024

Published: 12.12.2024

Citation: Zore S, Bhosale A, Chavan P. (2024). Sentiment Analysis. International Journal of Electronics and Computer Applications. 1(2): 50-54. https://doi.org/10.54839/ijeaca.v1i2.8

* **Corresponding author**.

soniyazore93@gmail.com

Funding: None

Competing Interests: None

ISSN
Print: XXXX-XXXX
Electronic: 3048-8257

Sentiment Analysis

Soniya Zore[1]*, Amol Bhosale[2], Pratibha Chavan[2]

1 PG student, Department of Electronics and Telecommunication, Trinity College of Engineering & Research, Kondhawa Budruk, Pune, 411048, Maharashtra, India
2 Assistant Professor, Department of Electronics and Telecommunication, Trinity College of Engineering & Research, Kondhawa Budruk, Pune, 411048, Maharashtra, India

Abstract

Social media platforms like Twitter have become a rich source of real-time data and public sentiment. Analysing sentiment on Twitter is essential for various applications, from brand monitoring to political analysis. This work focuses on Twitter sentiment analysis, employing natural language processing (NLP) techniques to categorize tweets as positive, negative, or neutral. We collect a large dataset of tweets, pre-process the text, and train machine learning models to predict sentiment. Our goal is to provide insights into public sentiment on various topics, trends, and events, which can be valuable for decision-makers in diverse domains.

Keywords: NLP; Twitter sentiment analysis; Public sentiment

Introduction

Disasters can strike suddenly, causing widespread devastation and requiring a swift response. Social media, particularly Twitter, has become a valuable platform for sharing information during emergencies. Analysing this real-time data can aid disaster relief efforts by:

- **Identifying affected areas**

Twitter offers a real-time window into disaster zones. Here's how to pinpoint affected areas:

Geotagged Tweets: Analyze tweets with location data (GPS) to directly identify impacted regions.

Location Mentions: Employ Natural Language Processing (NLP) to extract locations mentioned in tweets, even without explicit geotags.

Spatial Clustering: Cluster tweets based on location keywords. Areas with a high concentration of disaster-related messages likely represent affected zones.

Image and Video Analysis: Utilize image recognition to analyze pictures and videos shared on Twitter, potentially revealing damaged infrastructure or landscapes.

- **Understanding the nature of the disaster**

By analyzing tweet content (keywords, location) and sentiment, we can classify disaster type (flood, earthquake) and gauge its severity (urgency, desperation) to guide targeted response.

- **Assessing public sentiment and needs**

During disasters, understanding public sentiment and needs is crucial for effective response. Here's how Twitter data can help:

Sentiment Analysis: Utilize sentiment analysis tools to categorize tweets as positive, negative, or neutral. This reveals the emotional state of people in affected areas (e.g., fear, frustration, hope).

Hashtags: Track relevant hashtags. They can condense specific needs (e.g., #EvacuationNeeded, #BloodDonationDrive) and facilitate targeted aid distribution.

Location-Specific Analysis: Combine location data with sentiment analysis to understand the emotional landscape across affected areas. This helps prioritize resources and identify areas with the most urgent needs.

Call to Action Tweets: Identify tweets requesting specific actions (e.g., search and rescue, medical attention). These can be used to direct help towards those in immediate danger.

Sentiment analysis, a field of natural language processing (NLP), has emerged as a crucial tool for understanding public opinion, emotional responses, and attitudes expressed on social media platforms like Twitter. With the exponential growth of user-generated content on Twitter, ranging from personal anecdotes to global events, sentiment analysis offers valuable insights into the collective mood of the platform's diverse user base.

This burgeoning field utilizes advanced algorithms and machine learning techniques to categorize tweets as neutral, negative, or positive based on the underlying sentiment conveyed in the text. By analyzing linguistic cues, context, and emotive expressions, sentiment analysis enables researchers, businesses, and policymakers to gauge public sentiment on various topics, ranging from product reviews and brand perceptions to social issues and political discourse.

In this study, we delve into the realm of sentiment analysis on Twitter, exploring methodologies, challenges, and applications in deciphering the complex nuances of human emotions expressed through microblogging. By leveraging state-of-the-art techniques and harnessing the vast wealth of data available on Twitter, we aim to uncover trends, patterns, and insights that illuminate the intricate interplay between language, sentiment, and social dynamics in the digital age.

For example, the rapidly evolving political landscape in India, the role of social media platforms like Twitter in shaping public opinion and influencing electoral outcomes has become increasingly significant. With the advent of the 2019 Indian General Election, Twitter emerged as a key battleground for political parties to engage with voters, disseminate their messages, and counter opposition narratives.

Our study focuses on analyzing tweets by utilizing advanced sentiment analysis techniques to discern prevailing sentiments among users. Leveraging state-of-the-art machine learning algorithms, and few deep learning techniques like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, we aim to provide insights into the mood of the user and predict outcomes.

By analyzing Twitter data through these methods, we can gain valuable insights into public sentiment and needs, allowing for a more efficient and responsive disaster response.

This project explores how Long Short-Term Memory (LSTM) networks, a deep learning technique, can be used to classify tweets related to disasters.

Literature Review

Several researchers have proposed strategies for classifying tweets based on their informative contents. One approach involves leveraging deep learning techniques, which have demonstrated significant success in both image and text processing domains. For instance, in the realm of image processing, the instances where articles contain images and titles that are not directly related.

To tackle this issue, paper[1] introduced a convolutional neural network (CNN) model. This model utilizes both image and text features for similarity prediction. By extracting features from both modalities, the model offers insights into the image-text relationship, potentially enhancing search quality.

The highest spoken language in the world is English. Many researchers took nice efforts for sentiment analysis from English text. A. Tumasjan, et.al. did Twitter based prediction found only one time tweet by 50% users and rest 50% made 90% of the tweets, In 2009[2]. They worked on data set having approx. 100K tweets which is too unclear to predict for the big population of a country. B. Joyce and J. Deng collected tweets from presidential election of 2016 using streaming API filtered with specific words like Hillary Clinton and Donald Trump[3]. The keywords they included were democratic and republican and the full name of the top political candidates like Donald Trump. Around 79 million tweets in a period of two months and 10,000 unique users were collected initially. Then 1.9 million tweets containing emojis were filtered. Redundancy removal brought that size to 783K tweets. Each tweet's emotional orientation was detected by Emojis Sentiment Ranking; by converting to Unicode. Then applied regex on Unicode and found score from the customize list of 522 unique emoji characters. A sentiment classifier was built using a multinomial Naïve Bayes classifier. To test the classifier, they used hand-annotated tweets as test data and passed it to the classifier. They achieved 74.9 % accuracy on this test data. Emoji based judgement is not enough for accurate prediction, as emoji may not present in each tweet, or is used as a sarcastic tone. The word ordering is significant in sentiment analysis. For example, one sentence is "I have to read this book" and another sentence is "I have this book to read". Both examples have different meaning. This simple example is one of the examples where we can say word

ordering matters while doing sentiment analysis.

Additionally, the evolving tactics employed by spammers necessitate robust solutions that can circumvent traditional security mechanisms. Paper[4] proposed a method that employs low-level n-gram features to thwart tokenization, thereby addressing this challenge. Researchers have explored various approaches for analysing language-dependent devices, utilizing publicly available databases to assess the performance of multiple systems. They incorporate n-gram analysis from tweet tones into development methodologies. Notably, they have demonstrated the capability of technology to swiftly detect spam tweets, a critical aspect in real-time Twitter scenarios. Twitter's significance as a data source has propelled the popularity of tweet sentiment analysis.

In this context, paper[5] introduces a method for the partial analysis of standard tweets by adopting a classification strategy with notable successes. Given the prevalence of emotionally charged phrases in tweets, effectively addressing specific conceptual phrases could significantly enhance cognitive analysis techniques' efficacy. The study establishes a methodology focused exclusively on analysing tweets containing positive emotion phrases, showcasing promising results across both tasks.

The challenges inherent in social media data analysis, such as noise, brevity, and the need to categorize incoming messages, underscore the importance of access to human-specific information. To this end, paper[6] employs machine learning classifiers trained on adjective usage, alongside releasing a word2vec software trained on a vast corpus of disaster-related tweets. Additionally, the author addresses language variations in tweets by providing common lexical resources for lexical variants.

In the realm of tweet informative contents and event detection, paper[7] proposes a Convolutional Neural Network (CNN)-based method. This approach involves training a CNN model on recent earthquake-related tweets, with a focus on discerning informative tweets and real-time event detection. The CNN model plays a pivotal role in predicting Twitter keywords associated with seismic events, aiding in post-event earthquake detection with a high level of accuracy and pre-announcement confirmation from official disaster sources. Notably, existing programs in the literature primarily rely on small datasets or necessitate logical variable names for operation.

While machine learning models outperform lexicon and rule-based algorithms in sentiment classification tasks, in paper[8] LSTM model's accuracy doesn't meet expectations. However, there are avenues for enhancing its performance. For instance, the use of GloVe for word embedding lacks emoticon vector representations, hindering the model's ability to interpret emoticons in text. Employing alternative embedding with emoticon vectorization could improve accuracy. In retrospect, opting for different embedding and integrating attention mechanisms to prioritize emoticons and punctuations would be strategies to boost the model's effectiveness.

Pradip Bhere et. al.,[9] introduced a tweet classifier framework designed to process raw tweets and classify them as informative or non-informative. The model has demonstrated superior performance compared to an existing system that employed a combination of CNN and ANN. By leveraging both the Continuous Bag of Words (CBOW) and Skip-Gram models of Word2Vec in conjunction with Convolutional Neural Networks (CNNs), achieved an evaluation accuracy of 84%.

In conclusion, the literature surveyed underscores the multifaceted challenges and advancements in the classification and analysis of tweets across various domains. Deep learning techniques, particularly convolutional neural networks (CNNs), have shown promise in extracting meaningful insights from both textual and visual content, thereby enhancing search quality and predictive capabilities. Additionally, the relentless evolution of spamming tactics necessitates robust solutions, with low-level n-gram features emerging as a viable approach to thwart tokenization and improve detection accuracy.

Furthermore, sentiment analysis on Twitter has emerged as a vital tool for gauging public opinion and understanding emotional responses to diverse topics. Leveraging machine learning classifiers and word embedding trained on vast tweet datasets, researchers have made strides in accurately categorizing tweets based on their sentiment. However, challenges persist, particularly in interpreting emoticons and punctuations, highlighting the need for innovative approaches such as attention mechanisms to enhance model performance.

In summary, while considerable progress has been made in tweet classification and sentiment analysis, there remains ample room for further research and innovation. Continued exploration of advanced techniques and methodologies will be crucial in unlocking the full potential of tweet data for diverse applications ranging from disaster detection to election forecasting and beyond.

To overcome the above problem, in our work we developed a model that will try to analyse the sentiment based on word number. Deep learning approaches like Long-short Term Memory [LSTM] is one of the Recurrent Neural Network approaches where model is trained with the word's ordering whereas in the already existing work, they only considered emojis while doing the analysis. Due to the above-mentioned reasons, LSTM seems to be the perfect fit for doing sentiment analysis. So, we will be using LSTM model for our analysis.

Methodology

A. Data Acquisition

A publicly available Twitter disaster dataset will be used. This dataset should be labelled, indicating whether a tweet is related to a disaster or not. But following are some more methods available.

Obtaining Twitter data for disaster analysis requires strategic planning. Here are two main approaches:

1. Twitter API

- **Application:** Apply for a developer account with Twitter to access the Twitter API. This grants programmatic access to tweet data.
- **Keyword Filtering:** Define relevant keywords associated with disasters (e.g., ”flood,” ”earthquake,” ”hurricane”).
- **Location Targeting:** Specify locations of interest or filter based on geotagged tweets to pinpoint affected areas.
- **Historical vs. Real-Time:** Choose between gathering historical data for past events or collecting real-time data during ongoing disasters.

2. Streaming Tools

- **Tweepy (Python):** Utilize libraries like Tweepy (for Python) to connect to the Twitter Streaming API and collect tweets in real-time.
- **Focus on Speed:** Ideal for capturing the fast-moving nature of disasters and gathering the latest updates.
- **Data Volume:** Be prepared to handle a potentially high volume of tweets during a disaster.

3. Additional Considerations

- **Rate Limits:** Twitter API enforces rate limits on data collection requests. Plan your queries accordingly to avoid exceeding limits.
- **Data Filtering:** Develop a filtering process to eliminate irrelevant tweets (e.g., advertisements, spam) and focus on disaster-related information.
- **Ethical Considerations:** Respect user privacy and avoid collecting or using personally identifiable information (PII) without proper consent.

By effectively utilizing the Twitter API or streaming tools, you can acquire valuable data for disaster analysis. Remember to tailor your approach based on the specific disaster and your research goals.

The data set is obtained from Kaggle website, which is a popular site and an online community for data scientist and machine learning (ML) practitioner, which provides a wide variety of data science and ML problems. The problem am tackling is “Natural Language Processing with Disaster Tweets”, which is a competition posted by the website, it contains two datasets that consist of 7560 tweets, a training dataset (train.csv, 4983 rows), and testing dataset (test.csv, 2567 rows), the difference between the two datasets is that the training dataset contains a target attribute while the testing dataset does not, and that it because it is a competition where they want to test competitors. Consequently, for this project I will be selecting the training dataset only as the main dataset, and divide it between a training and testing subsets, in order to be able to evaluate the model after all.

B. Data Pre-processing

The raw tweet data needs cleaning and preparation before feeding it into the LSTM model. This may involve:

- **Removing irrelevant information:** Removing hashtags, URLs, user mentions, and special characters.
- **Text normalization:** Lowercasing text, correcting typos, and stemming/lemmatization (converting words to their base form).
- **Stop word removal:** Eliminating common words like ”the”, ”a”, ”an” that don’t contribute to the meaning.

Result and Discussion

The report will present the achieved accuracy and other evaluation metrics of the LSTM model. We will discuss the effectiveness of the model in classifying disaster tweets and identify any limitations.

Figures 1, 2 and 3 display the performance metrics of three different classification algorithms: Naive Bayes, Random Forest, and Support Vector Machine (SVM).

Naive Bayes Accuracy: 0.79
Naive Bayes Classification Report:

	precision	recall	f1-score	support
0	0.79	0.85	0.82	874
1	0.78	0.70	0.74	649
accuracy			0.79	1523
macro avg	0.79	0.78	0.78	1523
weighted avg	0.79	0.79	0.79	1523

Fig 1. Naive Bayes Report

Comparison

Naive Bayes has the highest accuracy among the three algorithms, viz. SVM and Random Forest

Naive Bayes also has the highest F1-score for both classes, indicating a good balance between precision and recall.

Random Forest Accuracy: 0.76
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.78	0.83	0.80	874
1	0.74	0.68	0.71	649
accuracy			0.76	1523
macro avg	0.76	0.75	0.76	1523
weighted avg	0.76	0.76	0.76	1523

Fig 2. Random Forest Report

SVM Accuracy: 0.77
SVM Classification Report:

	precision	recall	f1-score	support
0	0.79	0.82	0.80	874
1	0.74	0.71	0.72	649
accuracy			0.77	1523
macro avg	0.77	0.76	0.76	1523
weighted avg	0.77	0.77	0.77	1523

Fig 3. Support Vector Classifier (SVC) Report

Random Forest has the highest recall for class 0, while Naive Bayes has the highest recall for class 1.

SVM has comparable performance to Naive Bayes but slightly lower accuracy and F1-scores.

Conclusion

This work demonstrates the potential of using LSTM deep learning for disaster analysis on Twitter data. The ability to automatically classify disaster tweets can be a valuable tool for emergency response teams and humanitarian organizations.

References

1) Kundu S, Srijith PK, Desarkar MS. Classification of Short-Texts Generated During Disasters: A Deep Neural Network Based Approach. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE. 2018. Available from: https://doi.org/10.1109/ASONAM.2018.8508695.
2) Tumasjan A, Sprenger T, Sandner P, Welpe I. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*. 2010;4(1):178–185. Available from: https://doi.org/10.1609/icwsm.v4i1.14009.
3) Joyce B, Deng J. Sentiment analysis of tweets for the 2016 US presidential election. In: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE. 2017;p. 1–4. Available from: https://doi.org/10.1109/URTC.2017.8284176.
4) Ashour M, Salama C, El-Kharashi MW. Detecting Spam Tweets using Character N-gram Features. In: 2018 13th International Conference on Computer Engineering and Systems (ICCES). IEEE. 2019. Available from: https://doi.org/10.1109/ICCES.2018.8639297.
5) Phan HT, Nguyen NT, Van Cuong T, Hwang D. A Method for Detecting and Analyzing the Sentiment of Tweets Containing Fuzzy Sentiment Phrases. In: 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). IEEE. 2019. Available from: https://doi.org/10.1109/INISTA.2019.8778360.
6) Imran M, Mitra P, Castillo C. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016. 2016;p. 1638–1643. Available from: https://pure.psu.edu/en/publications/twitter-as-a-lifeline-human-annotated-twitter-corpora-for-nlp-of-.
7) Van Quan N, Yang HJ, Kim K, Oh AR. Real-Time Earthquake Detection Using Convolutional Neural Network and Social Data. In: 2017 IEEE Third International Conference on Multimedia Big Data (BigMM). IEEE. 2017. Available from: https://doi.org/10.1109/BigMM.2017.58.
8) Mollah MP. An LSTM model for Twitter Sentiment Analysis. 2022. Available from: https://doi.org/10.48550/arXiv.2212.01791.
9) Bhere P, Upadhyay A, Chaudhari K, Ghorpade T. Classifying Informatory Tweets during Disaster Using Deep Learning. *ITM Web of Conferences*. 2020;32:1–5. Available from: https://doi.org/10.1051/itmconf/20203203025.