Enhanced Sentiment and Emotion Detection in Tweets Using a Hybrid Machine Learning Approach

Anoop Rao Enaganthi
Illinois Institute of Technology
aenaganthi@hawk.iit.edu

Nachiketh Nallamaddi
Illinois Institute of Technology
nnallamaddi@hawk.iit.edu

Ajay Babu Popuri
Illinois Institute of Technology
apopuri1@hawk.iit.edu

March 2, 2024

Abstract

In an age when media sites like Twitter lead public discourse, an understanding of emotional nuances in what's being shared becomes key. This project details an advanced machine learning framework to categorize the tweets in refined emotions or emotional states, right from joy to surprise. The method, using a voting classifier with logistic regression and stochastic gradient descent, hence helps improve sentiment analysis. This indeed simplifies the process of pulling out emotional data from tweets in a much better way and hence further insights into societal sentiments. It's a big leap toward opinion mining.

1 Project Description

In today's digital context, social media, more so Twitter, has been described as a vast reservoir of big volumes of textual data. Some such data, reflecting the population's moods and perspectives, presents a special challenge for automated emotion detection and sentiment evaluation. Adding to this complexity is the fact that human emotion is complex in nature, and categorizing these emotions for the given task becomes extremely difficult due to the brief nature of tweets and the overwhelming use of jargon or informal language.

In addition, the COVID-19 outbreak has come to show the critical importance of social media in spreading information and hence increased awareness of the need to develop strong approaches towards sentiment analysis. One can only imagine the huge consequences that might stem from misinformation and the emotional components of a social media conversation. Sophisticated analytical methods are necessary for negotiating the complex arena of public sentiment.

Our research project has set out to address this difficulty through a framework that uses state-of-the-art NLP techniques and machine-learning algorithms. One big improvement in our methodology is the integration of a Voting Classifier that can gain the advantages from Logistic Regression and Stochastic Gradient Descent. The blended model is meant to aid in greater precision in categorizing the tweets in particular emotional groupings, hence yielding deeper comprehension of public sentiment

1.1 Existing Works

The sentiment analysis landscape was provided using Twitter data by several machine-learning and lexicon-based methods. The landscape is mostly oc-

cupied by methods that make use of algorithms like Support Vector Machine, Naive Bayes, Random Forest, and ensemble methods with an objective of binary sentiment classification. Besides, methods such as domain-specific lexicon generation are also used for emotion-based feature extraction to increase the accuracy of emotion recognition.

Lexicon-based methods have also been researched for sentiment extraction from Twitter data, with researchers developing domain-specific lexicons to increase recognition accuracy for emotions. These methods usually rely on lists of words that have been pre-annotated with the emotions they convey and on the existence of such words within the tweet. Despite being very simple and transparent, these lexicon-based approaches might suffer from poor coverage of the wide range of emotional expressions and context-dependent interpretations.

Our proposed work presents a new methodology using a voting classifier that combines Logistic Regression and Stochastic Gradient Descent. This approach is designed to minimize errors more effectively than individual classifiers. Our model is validated across different datasets, for the first time one of them being a unique 6-Emo's dataset from Kaggle, for even wider ranges of emotions to classify, which demonstrates its versatility and better performance for emotion recognition on Twitter data.

1.2 Preliminary plan

The preliminary steps are outlined as follows:

1. **Dataset Acquisition:**(Done)The initial step involves acquiring datasets from the Kaggle repository.

2. **Data Visualization and Preprocessing:**(In Progress)To understand the dataset characteristics.

3. **Feature Extraction:**(To be done)Post-preprocessing, the data is transformed into TF-IDF matrices for further analysis.

4. **Building the Models:**(To be done)Several Machine Learning models will be developed that includes Random Forest (RF),Support Vector Machine (SVM),Naive Bayes (NB),Gradient Boosting Machine (GBM),Logistic Regression (LR),Stochastic Gradient Descent (SGD) and Voting Classifier (VC) which integrates LR and SGD

5. **Evaluation Metrics:**(To be done)The models will be evaluated using metrics like F1 measure.

6. **Dataset Information:**(To be done)The project utilizes four types of datasets, where three are binary-class and one is a six-class dataset.

7. **Methodology Overview:**(To be done) A stepwise algorithmic structure will be employed for a clear understanding of the methodology.

References

[1] P. Routray, C. K. Swain, and S. P. Mishra, "A survey on sentiment analysis," *Int. J. Comput. Appl.*, vol. 76, no. 10, pp. 1–8, Aug. 2013.

[2] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNNLSTM)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020.

[3] C. Kariya and P. Khodke, "Twitter sentiment analysis," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 212–216.

[4] A. Alsaeedi and M. Zubair, "A study on sentiment analysis techniques of Twitter data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019.

[5] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, "GBSVM: Sentiment classification from unstructured reviews using ensemble classifier," *Appl. Sci.*, vol. 10, no. 8, p. 2788, Apr. 2020.

[6] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles", *Decis. Support Syst.*, vol. 66, pp. 170–179, Oct. 2014.