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

Anoop Rao Enaganthi
A20531044
Illinois Institute of Technology
aenaganthi@hawk.iit.edu

Nachiketh Nallamaddi
A20549679
Illinois Institute of Technology
nnallamaddi@hawk.iit.edu

Ajay Babu Popuri
A20547319
Illinois Institute of Technology
apopuri1@hawk.iit.edu

April 2, 2024

1 introduction

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.

2 Project Description

In the digital age, social media, especially Twitter, has become a significant source of data, reflecting diverse emotions and perspectives in real-time. This wealth of information, particularly evident during the COVID-19 pandemic, underscores the importance of effective sentiment analysis. Twitter's brief, informal nature, filled with slang and jargon, makes it a complex medium for accurately deciphering human emotions.

The pandemic has revealed social media's impact in the propagation of information and underlined the advanced sentiment analytics required. It is an important tool that traces the public's general sentiment and helps one navigate these murky waters, where high emotional outbursts are fueled by misinformation.

Sentiment analysis in social media presents a dual challenge technically, it demands processing vast text data using sophisticated algorithms; sociologically, it requires understanding human emotions and communication nuances. This intersection of technology and sociology is essential for effective analysis, especially during globally impactful events like the pandemic. This fusion of data analysis, technology, psychology, and sociology is vital in capturing the nuances of human sentiment in the digital realm, offering insights into the public mood as reflected in social media

Our 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

2.1 Description of the Data

In our project, three distinct datasets from Kaggle are utilized, each serving for a specific purpose:

Twitter Sentiment Dataset: This dataset contains a set of posts from Twitter, where each of them has been annotated for sentiment: positive, negative, or neutral. This dataset would contribute to understanding the public opinions and their emotions collected over Twitter, it will be a perfect training dataset for models in sentiment and natural language processing analysis. It has all kinds of topics and sentiments gathered from discussions on social media.

Women's E-commerce Dataset: This dataset is centered around women's e-commerce and includes detailed product descriptions, customer reviews, and ratings. It's an invaluable resource for understanding consumer behavior in the e-commerce space, specifically regarding women's products. This data, when analyzed, gives an insight into customer preference, level of satisfaction, and trending products that become highly critical inputs to market research and targeted marketing strategies in the e-commerce industry.

Sentiment Analysis Dataset on Hatred-Speech Detection from Twitter: This dataset is tailored for the detection of hate speech in Twitter posts. It contains tweets classified as containing hate speech or not containing hate speech. This is important in developing models that are able to identify, track, and potentially flag or filter hate speech from social media platforms. This greatly boosts the safety in communication while online and promotes healthy discourse within digital spaces.

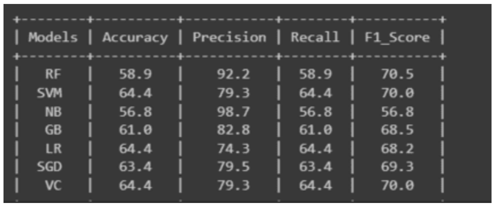
Each of these three datasets provided a unique insight from general Twitter sentiment analysis into specific consumer behavior in women's e-commerce and hate speech identification. With these, they offer a good framework for the analysis of various factors in digital communication and consumer trend that are pivotal in the understanding and addressing of challenges currently affecting the social media dynamics and online retails

2.2 Current Progress

In the **exploratory phase** of our analysis we have used the Women's E-commerce dataset from Kaggle and have applied various data visualization techniques, such as bar charts, histograms, and pie charts, with the intention to be able to understand the underlying patterns. A bar chart of the 'Happy' and 'Unhappy' groups from the emotions in the dataset, plotted using Python's pyplot module, showed that there were more subjects who felt 'Happy' than 'Unhappy'. This initial discovery underscores the importance of understanding emotional weight in consumer feedback, which can be critical for product and service adjustments in the e-commerce domain.

During the data **preprocessing** stage, our focus was to refine the raw dataset into a usable format. We began by checking for missing values and deciding on strategies for their imputation, however, the dataset contained no null values, allowing us to proceed without the need for data imputation. Our preprocessing steps included removing extraneous characters such as emojis and symbols, deleting numbers, converting text to lowercase, and stripping away common stop words and suffixes through stemming. These actions are essential in reducing noise and standardizing the data, thereby enhancing the performance of our subsequent analyses.

The next step was **feature extraction**, which helped in configuring cleaned data into a structured form so that it would be fit to go into a machine learning model. We would reduce the text data into a 2D feature matrix by summing the importance of the words in comparison to the frequency of all the words that appear in the corpus using a TF-IDF matrix. Not all these features turned out to be influential for classification. Feature extraction becomes a very important step for model learning, where relevant predictors are centered. Based on this, various machine learning models were trained, including Support Vector Machines, Random Forest, Naive Bayes, Decision Trees, Gradient Boosting Machines, Logistic Regression, Stochastic Gradient Descent, and a



```
+--------+-----------+------------+--------+----------+
| Models | Accuracy  | Precision  | Recall | F1_Score |
+--------+-----------+------------+--------+----------+
|   RF   |   58.9    |    92.2    |  58.9  |   70.5   |
|  SVM   |   64.4    |    79.3    |  64.4  |   70.0   |
|   NB   |   56.8    |    98.7    |  56.8  |   56.8   |
|   GB   |   61.0    |    82.8    |  61.0  |   68.5   |
|   LR   |   64.4    |    74.3    |  64.4  |   68.2   |
|  SGD   |   63.4    |    79.5    |  63.4  |   69.3   |
|   VC   |   64.4    |    79.3    |  64.4  |   70.0   |
```

Figure 1: Pretty Table comparing all the algorithms

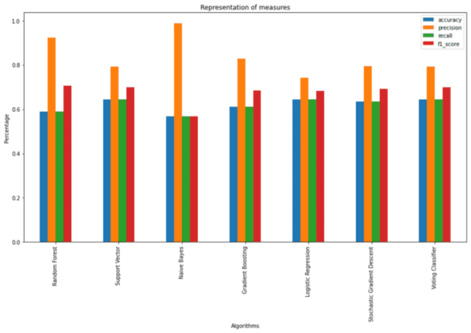


Figure 2: Bar graph Visualizing all the algorithms

Voting classifier for performance comparison in sentiment categorization.

Challenges faced include managing the imbalance in emotional categories and ensuring that the feature selection process effectively captures the nuances of sentiment without overfitting. This careful balance is crucial for the robust performance of our models, as seen in the performance metrics chart and bar graph visualizations for Women's E-commerce Dataset, highlighting the ongoing iterative process of model optimization.

2.3 Pending Work

Pending tasks for the project encompass crucial steps in data enhancement and model fine-tuning, particularly for Twitter Sentiment Dataset, and Sentiment Analysis dataset on Hatred-Speech Detection from Twitter. Currently, the prominent issue at hand is the imbalanced nature of these datasets. Imbalance can skew the performance of predictive modeling, as the models tend to be biased towards the majority class, leading to unreliable prediction metrics for the minority class.

To address this, we plan to implement an Over Sampling technique. This technique involves augmenting the underrepresented class in the dataset by duplicating examples or generating synthetic samples, which can help in achieving a more balanced class distribution. This approach is expected to enhance the robustness of our machine learning models, allowing them to learn more effectively from an equitable representation of all classes.

After balancing the datasets, witter Sentiment Dataset will be merged with Sentiment Analysis Dataset on Hatred-Speech Detection from Twitter, giving an integrated dataset for sentiment and hate-speech detection. The newly combined dataset will then be taken through some pre-processing to make sure it is apt for training a model. This will, in turn, include additional text normalization, noise reduction, and an extensive feature extraction process so that enough linguistic variations suitable for sentiment and hate-speech classification are caught within the features.

Lastly, with the enriched dataset, we will proceed to retrain our suite of machine learning models. This step will be followed by a detailed analysis of the model performances through cross-validation and an exploration of advanced ensemble techniques to boost the predictive accuracy. The aim is to finalize a model that not only performs well across all metrics but also generalizes effectively on unseen data, which is vital for real-world applications.

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.