TWEETPY

Tweet Classification Web App

Yagay Khatri, 2110110605

Priyansh Singhal, 2110110397

Tanisha Bashishth, 2110110542

IR Project

Supervisor: Dr. Sonia Khetarpaul

Department of Computer Science & Engineering

SHIV NADAR
INSTITUTION OF EMINENCE DEEMED TO BE
UNIVERSITY
DELHI NCR

TABLE OF CONTENTS

ABSTRACT

With the ever-growing volume of unstructured data on social media, classifying tweets into relevant categories has become an essential task in fields like information retrieval and natural language processing. This project focuses on building a tweet classification system that organizes tweets into various categories such as Politics, News, Events, Sports, Entertainment, and more. We began by manually looking and scraping a diverse set of English-language tweets to create a reliable ground truth for training our classification model.

To optimize the system's performance, we applied multiple machine learning algorithms, combining them through ensemble techniques to improve accuracy. The model's effectiveness was evaluated using standard metrics like precision, recall, and F-score, both overall and per category, to ensure its robustness in capturing the distinct characteristics of each class.

Additionally, a simple and intuitive web interface allows users to interact with the system, inputting tweets and receiving real-time classification results. The findings show that combining different algorithms leads to a significant improvement in classification accuracy, highlighting the potential of this approach for broader social media analysis tasks.

GitHub Link: https://github.com/yagaykhatri19/TweetPy
Render Link: https://tweetpy.onrender.com/

INTRODUCTION

The rise of social media platforms has radically changed the way information is shared and consumed, with Twitter being one of the most influential platforms for real-time updates, discussions, and public opinion. Tweets, however, are often short, frequent, and informal, posing significant challenges in processing and analyzing them effectively. Despite these challenges, tweets are a valuable source of real-time data, providing insights into news, events, public opinion, and much more.

Given the immense volume and unstructured nature of tweets, classifying them into predefined categories—such as Politics, News, Events, Sports, Entertainment, and others—becomes crucial for extracting meaningful information. Automated tweet classification allows researchers, businesses, and policymakers to quickly filter relevant data from the overwhelming noise, aiding decision-making, trend analysis, and even sentiment analysis.

For example, businesses can monitor public sentiment towards their products, while news agencies can distinguish between breaking news and casual posts. However, manually classifying this data is not only labour-intensive but also prone to inconsistency. Thus, developing a robust, automated classification system is critical for enhancing the efficiency and scalability of analyzing social media content.

The automation of tweet classification plays a vital role in managing large-scale data. For businesses, it provides a way to track customer feedback and emerging trends in real time. For journalists and media organizations, it helps in identifying the most relevant and urgent stories from a stream of millions of tweets. Furthermore, governments and policymakers can use such systems to gauge public opinion on various issues, aiding in faster and data-driven decision-making.

This project presents a tweet classification system with the following key contributions:

1. **Annotated Dataset**: We scraped and collected a diverse set of English-language tweets into predefined categories, creating a high-quality dataset that can serve as a foundation for training machine learning models.

2. **Multi-Algorithm Classification Model**: We combine various machine learning algorithms to build an ensemble model that maximizes classification accuracy. This approach outperforms single-model methods and ensures robustness in categorizing tweets across multiple classes.

3. **User Interface**: We design a simple, intuitive web interface that allows users to easily interact with the classification system. The interface provides real-time classification results, making the tool accessible for non-technical users and applicable in real-world scenarios.

In summary, this project provides a scalable, efficient, and practical solution for classifying tweets and can be extended to other forms of social media content. It bridges the gap between unstructured social media data and structured insights, making it a valuable tool for a wide range of applications in data analysis, sentiment analysis, and information retrieval.

LITERATURE REVIEW

The classification of tweets and text into meaningful categories has been widely studied, with researchers exploring various approaches to improve accuracy and usability. Below is an overview of the papers reviewed and the novelty of this project in comparison.

1. **Ashwin V,** ***"Twitter Tweet Classifier"***
   This paper explores the use of Naive Bayes (NB) for classifying tweets into predefined categories. While NB is computationally efficient, its performance is limited when handling complex and multi-faceted tweets. Our approach extends this work by combining NB with Logistic Regression (LR) in an ensemble model to improve classification accuracy and robustness.

2. **Rabia Batool et al.,** ***"Precise tweet classification and sentiment analysis"***
   The authors focus on precise tweet classification integrated with sentiment analysis, using advanced NLP techniques. Although their work highlights the importance of preprocessing and domain-specific features, it does not focus on ensemble models or user accessibility. Our project addresses this gap by combining algorithms for improved precision and offering a user-friendly interface.

3. **Alper Kursat Uysal and Serkan Gunal,** ***"The impact of preprocessing on text classification"***
   This paper emphasizes the significance of preprocessing techniques, such as stop-word removal, stemming, and lemmatization, on the performance of text classification models. We adopted these preprocessing techniques to enhance the quality of input data and ensure effective feature extraction in our classification pipeline.

4. **Johnson Kolluri et al.,** ***"Text Classification Using Machine Learning and Deep Learning Models"***
   The study compares traditional machine learning algorithms like SVM and NB with deep learning approaches. While deep learning showed superior results, the computational overhead was significant. Inspired by this, we chose lightweight algorithms like NB and LR, balancing accuracy and efficiency for real-time applications.

5. **IR Textbook and In-Class Lectures**
   The principles and techniques of text preprocessing, feature extraction, and evaluation metrics were adapted from standard information retrieval concepts discussed in textbooks and class lectures. These foundational concepts were tailored to address the specific challenges of tweet classification.

Novelty of Our Approach

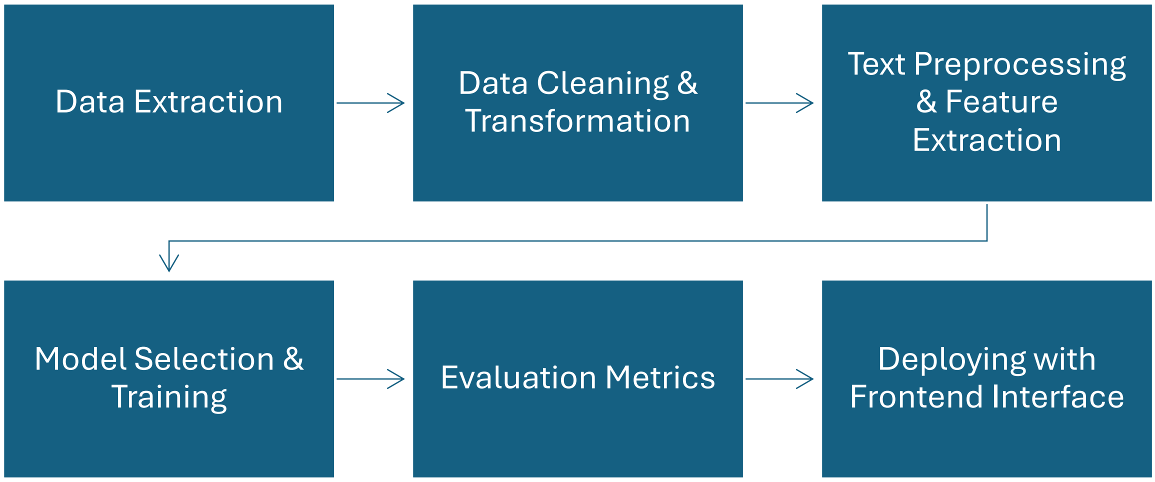
Our project stands out by comparing multiple machine learning algorithms and combining the complementary strengths of Naive Bayes and Logistic Regression in an ensemble model to optimize accuracy and robustness. Additionally, we introduced a **user-friendly web interface**, allowing users to classify tweets conveniently in real time, bridging the gap between theoretical research and practical application. This integration of improved accuracy, computational efficiency, and usability makes our approach both innovative and applicable to real-world scenarios.

OBJECTIVE

The primary goal of this project is to design, implement, and evaluate an automated tweet classification system capable of categorizing tweets into predefined categories. The system is intended to streamline the analysis of social media data, with a focus on English-language tweets rather than specifically Croatian Twitter data. The specific objectives of this project are:

1. **Create a Manually Annotated Dataset**: To ensure a reliable training and validation set, we manually scrape a diverse set of English-language tweets into categories such as Politics, News, Sports, Entertainment, and more. This annotated dataset will serve as the foundation for training the classification model and will enable the system to learn the distinct characteristics of each category.
2. **Develop a Multi-Algorithm Classification System**: The classification system will combine multiple machine learning algorithms to optimize accuracy and performance. By integrating different models, the system will be capable of learning from the annotated dataset and generalizing to new, unseen tweets, improving its ability to handle the wide variety of tweet content and topics.
3. **Optimize Classification Performance Using Evaluation Metrics**: The system's performance will be evaluated using standard classification metrics, including precision, recall, and F-score. These metrics will be calculated both overall and per category, enabling a comprehensive assessment of the system's effectiveness in distinguishing between different tweet categories and its overall robustness.
4. **Build a User-Friendly Web Interface**: A simple, interactive website will be developed where users can input tweets and receive real-time classification results. This user-friendly interface is designed to make the classification system accessible to non-technical users, allowing the tool to be used not only in academic settings but also for practical applications by businesses, media organizations, and the general public.
5. **Provide a Scalable Solution for Social Media Data Analysis**: The system will be designed to be scalable and adaptable for broader applications in social media analytics. This project aims to demonstrate how a multi-algorithm approach can efficiently process large volumes of unstructured text data from social media platforms, turning it into structured insights that can be used for real-time applications like trend analysis, sentiment analysis, and automated content categorization.

PROPOSED MODEL



Explanation of Project Stages:

- Data Extraction: Gathering tweets from various online sources to build a rich and diverse dataset for training and evaluation.
- Data Cleaning & Transformation: Removing unnecessary noise like URLs, special characters, and duplicates to ensure the dataset is clean and structured.
- Text Preprocessing & Feature Extraction: Preparing text by tokenizing, removing stop words, and lemmatizing while converting it into numerical features like TF-IDF for model compatibility.
- Model Selection & Training: Choosing suitable machine learning algorithms, training them on the dataset, and fine-tuning to achieve optimal performance.
- Evaluation Metrics: Measuring the effectiveness of the model using metrics like precision, recall, and F-score to ensure it categorizes tweets accurately.
- Deploying with Frontend Interface: Integrating the trained model with a user-friendly web interface, enabling real-time classification for practical use.

METHODOLOGY

Data Preprocessing

To prepare raw tweet data for effective analysis and classification, we employed several Natural Language Processing (NLP) techniques:

1. **Lowercasing**: All text was converted to lowercase to standardize input and eliminate case sensitivity.
2. **Noise Removal**: We removed URLs, special characters using regular expressions.
3. **Tokenization**: Tweets were split into individual words (tokens) using the nltk tokenizer.
4. **Stopword Removal**: Commonly used words (e.g., "the," "is") that do not add semantic meaning were filtered out using the NLTK stopword corpus.
5. **Lemmatization**: Each word was reduced to its base or dictionary form using the WordNet Lemmatizer to ensure consistency and reduce dimensionality.

The processed text was then stored in a new column for further analysis, creating a clean and structured dataset ready for feature extraction.

Feature Extraction

We used **TF-IDF (Term Frequency-Inverse Document Frequency)** to convert textual data into numerical features, emphasizing the importance of words that are frequent in a particular tweet but rare across the dataset. This representation was used as input for our machine learning models.

Model Selection and Training

After experimenting with multiple algorithms such as Decision Trees (DT) and Support Vector Machines (SVM), we finalized an **ensemble model** combining:

- **Naive Bayes (NB)**: A probabilistic classifier known for its simplicity and effectiveness in text classification tasks.
- **Logistic Regression (LR)**: A linear model that excels in binary and multi-class classification, offering robust performance and scalability.

Other algorithms like DT and SVM were excluded due to longer training times and comparable performance metrics (precision and recall) to the selected ensemble.

The ensemble approach leveraged the strengths of both NB and LR, balancing simplicity and performance.

Evaluation Metrics

The model was evaluated using standard classification metrics to assess its performance across all categories:

- **Precision**: The proportion of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**: The proportion of correctly predicted positive observations to all actual positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F-score**: The harmonic mean of precision and recall, balancing both metrics.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics were calculated both overall and for each category to ensure a comprehensive evaluation of the system's performance.

Integration with User Interface

To make the classification system accessible, we developed a **frontend interface** using Flask. This web-based application allows users to input a tweet and view its predicted category in real time. The integration ensures that the machine learning model is not just a research tool but a practical application for real-world use.

EXPERIMENTATION AND RESULTS

Logistic Regression Classification Report:

	precision	recall	f1-score	support
ARTS	0.35	0.38	0.36	300
ARTS & CULTURE	0.36	0.31	0.33	268
BLACK VOICES	0.43	0.38	0.40	300
BUSINESS	0.38	0.40	0.39	300
COLLEGE	0.61	0.48	0.54	229
COMEDY	0.58	0.45	0.51	300
CRIME	0.45	0.58	0.51	300
CULTURE & ARTS	0.56	0.30	0.39	215
DIVORCE	0.77	0.71	0.74	300
EDUCATION	0.59	0.52	0.55	203
ENTERTAINMENT	0.34	0.40	0.36	300
ENVIRONMENT	0.48	0.40	0.44	289
FIFTY	0.32	0.30	0.31	280
FOOD & DRINK	0.46	0.54	0.49	300
GOOD NEWS	0.35	0.30	0.32	280
GREEN	0.35	0.32	0.33	300
HEALTHY LIVING	0.24	0.26	0.25	300
HOME & LIVING	0.58	0.66	0.61	300
IMPACT	0.27	0.29	0.28	300
LATINO VOICES	0.65	0.39	0.49	226
MEDIA	0.51	0.53	0.52	300
MONEY	0.52	0.60	0.55	300
PARENTING	0.34	0.40	0.37	300
PARENTS	0.35	0.35	0.35	300
POLITICS	0.40	0.49	0.44	300
QUEER VOICES	0.79	0.63	0.70	300
RELIGION	0.58	0.48	0.53	300
SCIENCE	0.49	0.51	0.50	300
SPORTS	0.51	0.59	0.55	300
STYLE	0.38	0.38	0.38	300
STYLE & BEAUTY	0.56	0.58	0.57	300
TASTE	0.44	0.42	0.43	300
TECH	0.55	0.59	0.57	300
THE WORLDPOST	0.42	0.43	0.42	300
TRAVEL	0.47	0.56	0.51	300
U.S. NEWS	0.31	0.27	0.29	275
WEDDINGS	0.77	0.75	0.76	300
WEIRD NEWS	0.32	0.34	0.33	300
WELLNESS	0.29	0.37	0.33	300
WOMEN	0.35	0.39	0.37	300
WORLD NEWS	0.36	0.28	0.31	300
WORLDPOST	0.44	0.46	0.45	300

Naive Bayes Classification Report:

	precision	recall	f1-score	support
ARTS	0.33	0.45	0.38	300
ARTS & CULTURE	0.44	0.25	0.32	268
BLACK VOICES	0.41	0.31	0.36	300
BUSINESS	0.43	0.37	0.40	300
COLLEGE	0.66	0.45	0.53	229
COMEDY	0.50	0.47	0.49	300
CRIME	0.40	0.69	0.51	300
CULTURE & ARTS	0.61	0.19	0.29	215
DIVORCE	0.56	0.72	0.63	300
EDUCATION	0.69	0.45	0.54	203
ENTERTAINMENT	0.36	0.36	0.36	300
ENVIRONMENT	0.46	0.34	0.39	289
FIFTY	0.30	0.29	0.30	280
FOOD & DRINK	0.42	0.57	0.48	300
GOOD NEWS	0.42	0.34	0.38	280
GREEN	0.32	0.31	0.31	300
HEALTHY LIVING	0.24	0.22	0.23	300
HOME & LIVING	0.57	0.66	0.61	300
IMPACT	0.26	0.33	0.29	300
LATINO VOICES	0.82	0.27	0.41	226
MEDIA	0.52	0.52	0.52	300
MONEY	0.47	0.63	0.54	300
PARENTING	0.29	0.41	0.34	300
PARENTS	0.28	0.36	0.31	300
POLITICS	0.36	0.50	0.42	300
QUEER VOICES	0.67	0.60	0.64	300
RELIGION	0.60	0.45	0.51	300
SCIENCE	0.52	0.51	0.51	300
SPORTS	0.54	0.57	0.55	300
STYLE	0.41	0.36	0.38	300
STYLE & BEAUTY	0.52	0.58	0.55	300
TASTE	0.40	0.36	0.38	300
TECH	0.56	0.59	0.57	300
THE WORLDPOST	0.42	0.42	0.42	300
TRAVEL	0.48	0.57	0.52	300
U.S. NEWS	0.34	0.23	0.27	275
WEDDINGS	0.68	0.71	0.70	300
WEIRD NEWS	0.40	0.28	0.33	300
WELLNESS	0.28	0.40	0.33	300
WOMEN	0.34	0.29	0.32	300
WORLD NEWS	0.39	0.31	0.34	300
WORLDPOST	0.46	0.45	0.46	300

The above two images show us the precision, recall, and f1-score of Logistic Regression and Naïve Bayes algorithm.

Accuracy is a commonly used metric in text classification, but as learned in class it has limitations, especially with imbalanced datasets. For example, if 90% of tweets belong to the "News" category, a model predicting "News" for every tweet could achieve 90% accuracy without learning meaningful distinctions between categories. Additionally, accuracy does not provide insights into the types of errors made, such as falsely classifying a "Politics" tweet as "Entertainment" or missing "Events" entirely. These issues highlight why other metrics like precision, recall, and F1-score are often better for evaluating text classification models, as they offer a more detailed view of performance.

Ensemble Model Classification Report:

	precision	recall	f1-score	support
ARTS	0.38	0.44	0.41	300
ARTS & CULTURE	0.42	0.32	0.36	268
BLACK VOICES	0.38	0.37	0.38	300
BUSINESS	0.36	0.35	0.35	300
COLLEGE	0.60	0.49	0.54	229
COMEDY	0.54	0.46	0.50	300
CRIME	0.45	0.62	0.52	300
CULTURE & ARTS	0.57	0.32	0.41	215
DIVORCE	0.71	0.68	0.70	300
EDUCATION	0.55	0.50	0.52	203
ENTERTAINMENT	0.35	0.39	0.37	300
ENVIRONMENT	0.44	0.36	0.39	289
FIFTY	0.28	0.33	0.30	280
FOOD & DRINK	0.41	0.49	0.44	300
GOOD NEWS	0.36	0.38	0.37	280
GREEN	0.35	0.35	0.35	300
HEALTHY LIVING	0.29	0.30	0.30	300
HOME & LIVING	0.56	0.62	0.59	300
IMPACT	0.33	0.30	0.31	300
LATINO VOICES	0.76	0.39	0.52	226
MEDIA	0.55	0.55	0.55	300
MONEY	0.53	0.61	0.57	300
PARENTING	0.33	0.35	0.34	300
PARENTS	0.33	0.33	0.33	300
POLITICS	0.42	0.51	0.46	300
QUEER VOICES	0.75	0.59	0.66	300
RELIGION	0.63	0.56	0.59	300
SCIENCE	0.51	0.53	0.52	300
SPORTS	0.54	0.61	0.57	300
STYLE	0.45	0.47	0.46	300
STYLE & BEAUTY	0.54	0.54	0.54	300
TASTE	0.36	0.41	0.38	300
TECH	0.49	0.55	0.51	300
THE WORLDPOST	0.43	0.37	0.40	300
TRAVEL	0.50	0.60	0.55	300
U.S. NEWS	0.34	0.24	0.28	275
WEDDINGS	0.72	0.72	0.72	300
WEIRD NEWS	0.33	0.29	0.31	300
WELLNESS	0.31	0.42	0.36	300
WOMEN	0.36	0.36	0.36	300
WORLD NEWS	0.40	0.35	0.38	300
WORLDPOST	0.47	0.48	0.47	300

Our tweet classification system demonstrates robust performance across multiple categories, achieving notable accuracy and balanced precision, recall, and F1-scores for categories such as **Politics**, **News**, **Sports**, **Entertainment**, and others. By combining **Naive Bayes (NB)** with **Logistic Regression (LR)** in an ensemble model, we were able to capitalize on the complementary strengths of these algorithms. NB efficiently handles probabilistic text classification, while LR provides greater decision boundary flexibility, resulting in a more accurate and adaptable model overall.

Comparison with Existing Work

We compared our results with the findings from the paper *"Twitter Tweet Classifier"* (2016) by Ashwin V, which used **Naive Bayes** exclusively for tweet classification. Their approach yielded satisfactory results for certain categories but faced limitations in distinguishing categories with overlapping linguistic patterns or smaller representation in the dataset.

CONCLUSION AND LIMITATIONS

Conclusions

Our tweet classification system successfully combines **Naive Bayes (NB)** and **Logistic Regression (LR)** in an ensemble model, leveraging their strengths to achieve high **precision** and **recall** across multiple tweet categories. This approach allowed us to not only classify tweets accurately but also rank the **top three most likely categories** for each tweet, providing a nuanced understanding of ambiguous or multi-faceted tweets. The results demonstrate the model's capability to handle diverse datasets with varying linguistic patterns, showcasing its potential for applications in social media analytics and information retrieval.

Limitations

Text classification remains inherently challenging due to the complex nature of language. Tweets, in particular, often include slang, sarcasm, and emotional undertones that are difficult for machine learning models to interpret. Additionally, the short length of tweets limits contextual understanding, sometimes leading to misclassification.

Despite these challenges, the advancements in **natural language processing (NLP)** techniques and models—such as **transformers** and **contextual embeddings**—offer promising avenues for improvement. Incorporating these advanced techniques in the future could enhance our system's ability to capture emotional and contextual subtleties, improving its accuracy and adaptability to real-world applications.

While our model provides a strong foundation, its performance could be further enhanced with larger datasets, more diverse categories, and continued developments in AI-driven language understanding.

REFERENCES

1. Ashwin V, "Twitter Tweet Classifier"
2. Rabia Batool; Asad Masood Khattak; Jahanzeb Maqbool; Sungyoung Lee, "Precise tweet classification and sentiment analysis"
3. Alper Kursat Uysal; Serkan Gunal, "The impact of preprocessing on text classification"
4. Johnson Kolluri; Dr Shaik Razia; Soumya Ranjan Nayakier, “Text Classification Using Machine Learning and Deep Learning Models”
5. IR Textbook and In-class Lectures.