Bahria University

Lahore Campus

Department of Computer Sciences

Natural Language Processing

ASSIGNMENT # 03 (Final)

DUE DATE: 3, January 2024

Instructor Name: Tayyab Mir **Course Code:**

Program: BSCS **Max marks: 10**

Instructions:

- The assignment should be submitted before the deadline.
- Late submission is not allowed.
- Plagiarism will be considered as serious academic offense and may result in F grade.

Name: **Mohsin Hussain Mirza** **Enrollment:** **03-134202-031** **Class:** **BSCS-7A**

Name: **Abdullah Hassan** **Enrollment:** **03-134202-112** **Class:** **BSCS-7A**

Contents

Challenges in Sentiment Analysis on Twitter: Navigating Noisy, Short, and Multilingual Texts

Problem Definition and Motivation

1. **Noisy and Informal Language:** Tweets often contain informal language, abbreviations, slang, and misspellings. Analyzing sentiment accurately becomes challenging due to this unstructured and diverse linguistic style.
2. **Short Texts:** Tweets are limited to 280 characters, leading to contextually incomplete expressions. Extracting sentiment from such concise texts requires understanding the context and implied emotions, often making it difficult to interpret.
3. **Contextual Ambiguity:** Sentiment can drastically change based on the context, sarcasm, irony, or even emoji usage. Identifying the true sentiment behind such nuanced expressions is complex.
4. **Subjectivity and Multilingual Content:** Twitter is used globally, resulting in a wide range of languages and cultural contexts. Sentiment analysis models need to handle multilingual content and varied expressions of sentiment across different cultures.
5. **Volume and Velocity:** Twitter generates a vast amount of data in real-time. Processing this volume efficiently and continuously to extract sentiment is challenging.
6. **Handling Imbalanced Data:** Sentiment analysis datasets often suffer from class imbalance, where positive, negative, and neutral sentiments might not be equally represented. Balancing the dataset for training models becomes crucial.

IMPORTANCE:

1. **Real-time Public Opinion:** Twitter reflects public opinion on a wide range of topics, from political views to product reviews. Analyzing sentiment helps understand and gauge the general sentiment towards events, brands, or societal issues in real-time.
2. **Customer Feedback and Business Insights:** Businesses can leverage sentiment analysis to understand customer feedback, identify areas for improvement, and shape marketing strategies based on customer sentiment towards their products or services.
3. **Crisis Management and Social Trends:** Identifying sentiment during a crisis or around social trends aids in crisis management, public safety, and trend analysis, providing insights into public sentiment during crucial moments.
4. **Policy Making and Societal Impact:** Governments and policymakers can use sentiment analysis to understand public sentiment on policies, social issues, and public services, thereby making informed decisions that align with public opinion.
5. **User Engagement and Support:** For social platforms and customer support, sentiment analysis helps in managing user engagement by identifying user sentiment towards the platform or support services, enabling prompt responses and support.
6. **Predictive Analysis and Trend Forecasting:** Sentiment analysis provides a basis for predictive analytics, helping anticipate future trends, market movements, and public opinion shifts.

Methodology:

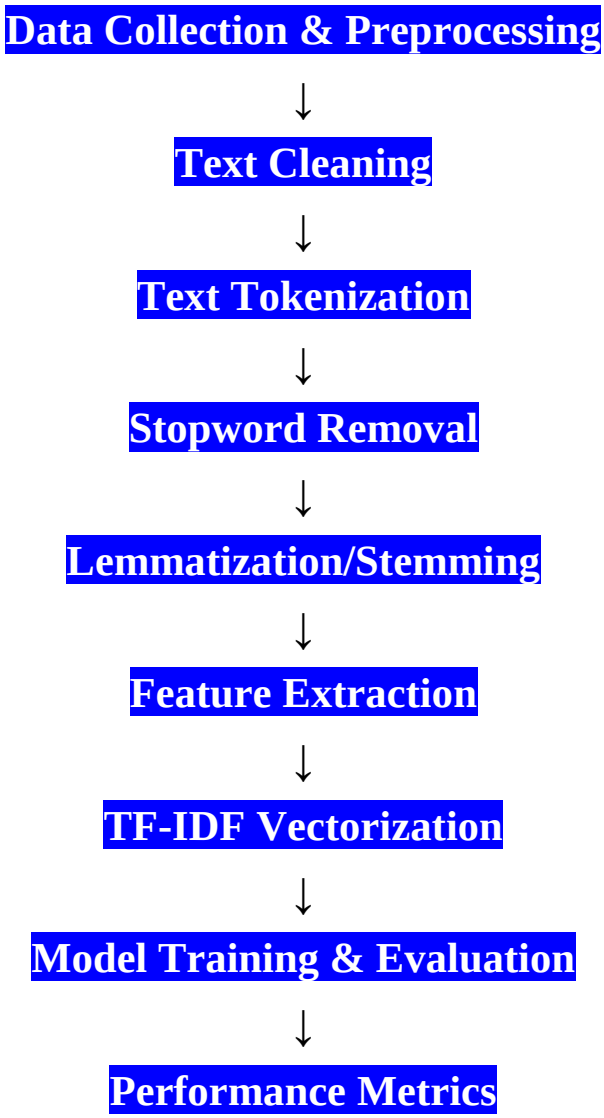
1. **Text Pre-processing:** Converted text to lowercase to ensure uniformity in the text data. Break text into individual words or tokens to prepare for further processing. Eliminated URLs, special characters, mentions, and hashtags as they might not add substantial value to sentiment analysis.
2. **Column Filtering and Justification:** Retained columns containing text content (e.g., 'text') and sentiment labels (e.g., 'airline_sentiment'). Filtered columns ensures focus on essential data relevant to sentiment analysis.
3. **Normalization Steps:**
    - Eliminated common stopwords that do not contribute significantly to sentiment analysis.
    - Lemmatization/Stemming, Reducing words to their base or root form to standardize the vocabulary.
4. **Algorithms/Techniques for Feature Extraction:**
    - TF-IDF (Term Frequency-Inverse Document Frequency): I have used TF-IDF vectorization to convert text data into numerical feature vectors, capturing the importance of words in the corpus.
    - N-grams: Utilize N-grams (e.g., unigrams, bigrams) in TF-IDF vectorization to capture the sequence of words and contextual information.
5. **Feature Set:**
    - **Bag-of-Words (BoW):** Represented text as a bag of words after pre-processing, retaining important words as features.
    - **TF-IDF Features:** Used TF-IDF values as features for each word in the text corpus, considering their importance in individual tweets.

Machine Learning Models:

- LinearSVC: Linear Support Vector Classifier (LinearSVC) is being utilized for classification tasks due to its effectiveness in handling high-dimensional data and linear separation.
- Random Forest Classifier: Random Forest as it works well with text data and handles non-linear relationships between features.
- SVM
- MultinomialNaiveBayes

Evaluation Metrics: Accuracy, Precision, Recall, F1-score are the metrics used to evaluate model performance. Accuracy measures the overall correctness, while precision, recall, and F1-score provide insights into the model's performance concerning positive, negative, and neutral sentiments.

6. **A framework diagram for our methodology:**



Dataset:

1. **Sentiment Analysis:**
   Different sentiments expressed in tweets: positive, negative, neutral. Confidence levels associated with sentiment predictions.
2. **Negative Sentiment Reasons:**
   Reasons provided for negative sentiment, if available. Confidence levels associated with negative sentiment reasons.
3. **Tweet Information:**
   - Tweet ID: Unique identifier for each tweet.
   - Text: Actual content of the tweet.
   - Retweet count: Number of times the tweet was retweeted.
   - Tweet creation date and time.
   - User-related details: Twitter username, user timezone, and tweet location.
4. **Airline Information:**
   The airline involved in the tweets (Virgin America).Sentiment labeled as "gold standard" (possibly human-labeled).

5. **Miscellaneous:**
   Coordinates associated with tweets (tweet_coord).
6. **Geographical and Temporal Aspects:** Tweet location and user timezone, providing insights into the geographical distribution of tweets and user time zones.

Results and Discussion

Describing the confusion matrix of our results:

The confusion matrix helps in understanding the distribution of predicted and actual classes, assisting in the calculation of various performance metrics.

	Predicted Negative	Predicted Neutral	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)	False Positive (FP)
Actual Neutral	False Negative (FN)	True Neutral (TN)	False Negative (FN)
Actual Positive	False Negative (FN)	False Negative (FN)	True Positive (TP)

Results of experiments in tabular form:

Model	Accuracy	Precision	Recall	F1-Score
LinearSVC()	0.7971	0.7864	0.7971	0.7873
RandomForestClassifier()	0.7715	0.79	0.94	0.86
SVM()	0.7838	0.7743	0.7838	0.7755
MultinomialNaiveBayes()	0.7622	0.7532	0.7622	0.7423

Future Work

These are some Future advancements that can be applied.

1. **Deep Learning Architectures:** Employing more complex deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), or Transformer models (such as BERT, GPT) for improved context understanding in text data.
2. **Real-Time Sentiment Monitoring:** Developing systems for real-time sentiment monitoring and analysis of Twitter data related to airlines, allowing prompt responses to customer feedback and emerging issues.