Project Report
CSCI 5502: Data Mining

Project Title: **Exploring Social Media Reactions to Disasters Using Text Mining and Sentiment Analysis**

Prepared By: **Yash Amre (yaam5835@colorado.edu)**

Abstract:

Social media has become a vital platform for public communication and emotional expression during disaster events. This project explores social media reactions to disasters using advanced text mining, sentiment analysis, and clustering techniques. Data was gathered from platforms like API's and YouTube, capturing diverse public opinions and emotions. Preprocessing techniques such as tokenization, lemmatization, and normalization were applied to the clean and structure the data. Sentiment Analysis was done in which the overall sentiments were neutral, while a strong recording of the negative responses demonstrated the public's distress. Emotion detection detected fear and sadness as dominant themes; clustering techniques detected main patterns and repeated topics, such as requests for the aid and the information of dissemination. A few examples of intuitive visualizations included like word clouds, sentiment distributions, and co-occurrence networks. These results indicate actionable suggestions for the disaster management teams in light of improving communication with the public, engaging the community, and addressing concerns effectively. This project underlines the power of social media data in understanding the collective sentiments, purposed at shaping real-time disaster response strategies.

1. Introduction

Disasters often generate copious activities and discussions on social media, displaying emotions, experiences, and information dissemination. This research utilizes natural language processing to interpret insights from such information, drawing inferences regarding the following aspects:

- Sentiment classification-Positive, Negative, Neutral
- Sentiment analysis (fear, anger, happiness, sadness, etc.)
- Thematic grouping of topics; word co-occurrence patterns.

Objective:

To understand public sentiment and the emotional reactions to disasters and afterwards identify key themes and trends in social media discussions.

2. Data Sources and Collection

Data Sources:

Project Report
CSCI 5502: Data Mining

1. **Articles:**
    - Extracted using the APIs (e.g., Guardian News and World News API).
    - Focused mostly on keywords like "earthquake," "disaster," and "relief."

2. **YouTube Comments:**
    - Gathered data using the YouTube Data API from videos discussing disaster events.

3. **Kaggle Datasets:**
    - Gathered data via Web Scraping andiIncluded disaster-related tweets and additional text datasets.

Data Details:

- **Volume:** Around 30,000 entries from Web Scraping, Articles, and YouTube comments.
- **Attributes:** Text content, timestamps, and sentiment scores.

3. Methodology

3.1. Data Preprocessing and Cleaning:

- Duplicates and irrelevant entries were removed.
- Missing Value Handling:
    - Numerical Attributes: Mean value imputation
    - Categorical Attributes: Mode imputation
- Text Standardization:
    - Tokenization, lemmatization, and stopword removal using SpaCy and NLTK.
    - Numerical Attributes Normalization: Z-score normalization

3.2. Sentiment Analysis:

- Tools Used: VADER, TextBlob
- The aim was to classify the text into positive, neutral, or negative sentiments.
- Furthermore, the polarity in sentiment was analyzed for an understanding of the intensity and distribution.

3.3. Emotional Analysis:

- Library Used: NRCLex
- Extracted emotions such as fear, anger, joy, and sadness from the text data.

3.4. Clustering:

- Used **TF-IDF** vectorization to extract features from text.
- Reduced the dimensionality using **TruncatedSVD** by performing dimensionality reduction, hence yielding computational efficiency.
- Applied **K-Means** to cluster data into thematic groups.

3.5. Visualizations:

- **Created informative plots such as:**
  - **Bar Plot** for the frequency of keywords.
  - **Histogram** to visualize the distribution of sentiment score.
  - **Box Plot** to detect the outliers and visualize the range of spread.
  - **QQ Plot** for the normality of numerical distribution.
  - **Word Cloud** for the frequently occurring words in the dataset.
  - **Pie Chart** for showing proportion of sentiment polarity.
  - **Heatmap** for showing co-occurrence in negative sentiment articles.
  - **Bubble Chart** to visualize top 20 words which are classified as positive sentiments.
  - **Silhouette Plot** to evaluate quality of clusters formed using K-mean.
  - **Scatter Plot** to visualize cluster formed using reduced TF-IDF.

3.6. Model Implementation:

- **Created informative model such as:**
  - **Naïve Bayes Classifier** which results in effective baseline model for sentiment classification.
  - **Logistic Regression** which achieves higher accuracy and F-1 score compared to Naïve Bayes.
  - **Support Vector Machine** results in superior performance with higher accuracy and precision.
  - **Decision Tree** which provides moderate accuracy, less robust than SVM and Logistic Regression.
  - **K-Nearest Neighbors(KNN)** results in high accuracy and competitive performance, slightly lagging in F-1 score compared to SVM due to sensitivity to imbalanced data.
  - **Apriori Algorithm** helped unveiled significant patterns e.g. co-occurences of keywords like “boy” and “charge”.

4. Results

4.1. Sentiment Analysis:

- **Distribution:**
  - Positive Sentiments: **25%**
  - Neutral Sentiments: **55%**
  - Negative Sentiments: **20%**
- **Insights:**
  - Neutral comments preponderated to show both factual reportage and detached discussion.
  - Negative sentiments underlined that the community was indeed quite distressed, especially in worst-hit areas.

4.2. Emotional Analysis:

- Fear and sadness dominated other emotions in disaster-related sentiments when outpourings of concern and even pity could be identified.
- Joy in some of the contexts is where the rescue operations have been successful.

4.3. Clustering:

- No. of Clusters: 3 (based on the elbow method and silhouette analysis)
- Themes Identified:
  - Cluster 1: Discussions of relief and support.
  - Cluster 2: Descriptions of events and the sharing of news.
  - Cluster 3: Emotional responses and public opinions.

4.4. Co-occurrence Analysis:

- Words like "help," "disaster," and "pray" tend to co-occur, reflecting shared concern and calls for action.

5. Discussion

- **Key Findings:**
  - Public sentiment during disasters is mostly neutral but with spikes in negative sentiments, which indicate points of distress.
  - Fear and sadness are the main emotional responses, indicating that disaster communication must have a strong need for empathy.
  - Clustering shows themes such as discussions of support, factual reporting of events, and emotional responses.

- **Challenges:**
  - Noisy and unstructured data from social media.
  - Capturing sarcasm or complex emotions from textual data.

- **Recommendations:**
  - Dynamic sentiment monitoring in real time for disaster response strategies.
  - Deepened sentiment classification by transformer-based models such as BERT.

6. Future Work

- Integrate geospatial analysis to map the sentiment distributions across regions.
- Expand data sets to more platforms like Reddit and Instagram.
- Develop an interactive dashboard with live monitoring of sentiment trends.
- Integrate state-of-the-art NLP models to detect sarcasm and more subtle nuances of emotion.

7. Conclusion

This project provides a very strong framework for disaster social media reaction analysis. Important insights into public sentiment and emotions have been unraveled in tune with the results derived from combining text mining, sentiment analysis, and visualization techniques. The latter may help disaster response teams to tailor communication strategies in order to calm public concerns.

8. Appendix

- **Technologies Used**:
  - Programming: Python
  - Libraries: Pandas, NumPy, NLTK, SpaCy, VADER, TextBlob, Seaborn, Matplotlib, Scikit-learn, NetworkX, NRCLex
- **Dataset Sources**:
  - Guardian News API, World News API
  - YouTube Data API
  - Kaggle Datasets(Web Scraping)

WebsiteURL:

https://sites.google.com/view/social-media-reactions?usp=sharing

GitHub:

https://github.com/yashamre/Exploring-Social-Media-Reactions-to-Disaster-using-Text-Mining-and-Sentiment-Analysis