Analyzing Bangladeshi Political News Sentiment with Machine Learning

A Research Report

Submitted in partial fulfillment of the requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

Submitted by

Anik Kumar Chakrabortty	**2017000000037**
Md. Nishad Khan	**2017000000150**
Shafi Md. Rawfur Raad	**2017000000242**

Supervised by

Mr. Rifat Ahommed

Lecturer

Department of Computer Science and Engineering

Southeast University, Bangladesh

Department of Computer Science and Engineering

Southeast University, Bangladesh

Dhaka, Bangladesh

10 February, 2024

Letter of Transmittal

10 February, 2024

The Chairman,
Department of Computer Science and Engineering
Southeast University, Bangladesh
Tejgaon, Dhaka

Through: Supervisor, Mr. Rifat Ahommed

Subject: Submission of Research Report

Dear Sir,
We are pleased to submit the research report titled "Analysing Political News Sentiment with Machine Learning" in fulfillment of the BSc in CSE at Southeast University. This collaborative effort represents the culmination of 4-months of dedicated research and analysis conducted under the guidance of our esteemed supervisor.

The objective of our research was to analyze the sentiment of Political News Article from various sources with help of Machine Learning.

Thank you for the opportunity to undertake this research, and we look forward to any feedback or discussion that may arise from the examination of this report.

Sincerely Yours,

Anik Kumar Chakrabortty
2017000000037

Md. Nishad Khan
2017000000150

Shafi Md. Rawfur Raad
2017000000242

Supervisor:

Mr. Rifat Ahommed
Lecturer & Supervisor
Department of Computer Science and Engineering
Southeast University, Bangladesh

CANDIDATE'S DECLARATION

We, hereby, declare that the thesis presented in this report is the outcome of the investigation performed by us under the supervision of Mr. Rifat Ahommed, Lecturer, Department of Computer Science and Engineering, Southeast University, Bangladesh. The work was done through CSE4000: Research Methodology course, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this research nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Anik Kumar Chakrabortty
2017000000037

Md. Nishad Khan
2017000000150

Shafi Md. Rawfur Raad
2017000000242

CERTIFICATION

This research titled, **“Analyzing Bangladeshi Political News Sentiment with Machine Learning”**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in 10 February, 2024.

Group Members:

Anik Kumar Chakrabortty	**2017000000037**
Md. Nishad Khan	**2017000000150**
Shafi Md. Rawfur Raad	**2017000000242**

Supervisor:

Mr. Rifat Ahommed
Lecturer & Supervisor
Department of Computer Science and Engineering
Southeast University, Bangladesh

Shahriar Manzoor
Associate Professor & Chairman
Department of Computer Science and Engineering
Southeast University, Bangladesh

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have contributed to the completion of this research report. This endeavor would not have been possible without the support, guidance, and encouragement of various individuals.

First and foremost, We extend our deepest appreciation to our supervisor, Mr. Rifat Ahommed, for his invaluable guidance and mentorship throughout the research process. Their expertise, constructive feedback, and unwavering support significantly enhanced the quality of this report.

We would like to acknowledge the Dept of CSE, Southeast University for providing access to resources and facilities essential for the successful execution of this research. The conducive research environment and collaborative atmosphere greatly enriched the overall experience.

The continuous assessment during the semester and feedback from Faculty members of Dept of CSE, Southeast University helped us overcome many mistakes and helped us learn better to conduct a research. We will be forever grateful for the effort from the department.

Dhaka
10 February, 2024

Anik Kumar Chakrabortty

Md. Nishad Khan

Shafi Md. Rawfur Raad

ABSTRACT

In this research endeavor, we extensively explore sentiment analysis within the realm of Bangladeshi political news, utilizing a diverse dataset sourced from prominent English news outlets. The dataset, meticulously categorized into positive, negative, and neutral sentiments, undergoes rigorous training across various machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, Random Forest, and Logistic Regression. Manual annotation via the Doccano Tool enriches the dataset with nuanced sentiment variations, fostering a profound understanding of the intricate political landscape.

A crucial aspect of our methodology involves implementing an automated crawler script designed to extract political content from a range of Bangladeshi online English newspapers. The extracted content is organized into paragraphs and stored in CSV format. This dataset, intended for public release on Kaggle, not only includes primary political news data but is enhanced by the incorporation of a supplementary set of common English sentences strategically integrated to elevate the overall accuracy of sentiment analysis. [1]

The adoption of the Bag of Words approach further refines our sentiment analysis methodology, enabling the capture of subtleties in language and context for a more nuanced understanding of political discourse. [2] The implications of this research extend beyond sentiment analysis as we leverage the insights gained to construct an innovative recommendation system tailored specifically for the nuanced landscape of Bangladeshi political news consumption.

By integrating perspectives from diverse sources and viewpoints, our study aims to provide a perspective in machine learning algorithm selection decisions. We have considered different aspects and evaluated the accuracy and performance. The impact of this is directly related to building recommendation system on this particular domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the contemporary landscape of news consumption, the advent of digital platforms has ushered in a paradigm shift, revolutionizing the way information is disseminated and consumed. Nowhere is this transformation more palpable than in the realm of political news, where the proliferation of online platforms and social media has given rise to a dynamic and diverse ecosystem influencing public opinion and discourse.

Within this evolving landscape, Bangladesh serves as a compelling context, characterized by a tapestry of political narratives interwoven with rich sociocultural nuances. The nation's political dynamics are multifaceted, shaped by historical legacies, cultural diversity, and contemporary realities. Consequently, the study of sentiment analysis within the context of Bangladeshi political news becomes not just an academic pursuit but a vital exploration with far-reaching implications for political actors, journalists, and the wider public.

Sentiment analysis, situated at the intersection of natural language processing and machine learning, emerges as a powerful instrument for deciphering the emotional nuances and opinions embedded in textual data. While existing literature extensively explores sentiment analysis across various domains, the unique challenges posed by Bangladeshi political news necessitate a tailored and context-aware approach. The linguistic subtleties, cultural diversity, and a spectrum of political perspectives demand an approach that goes beyond conventional sentiment analysis methodologies.

The aspiration to comprehend public sentiment towards political developments is not new. However, in the context of Bangladesh, where the digital landscape is evolving rapidly, there exists an opportunity to delve deeper into the intricacies of sentiment analysis. As political discourse unfolds across various online platforms, understanding the nuances of public sentiment becomes imperative.

This research endeavors to contribute to the evolving field of sentiment analysis by conducting a comprehensive examination of sentiment within Bangladeshi political news. By leveraging machine learning algorithms, manual annotation, and advanced techniques like the Bag of Words model, the study aims to unravel the complexities inherent in the sentiment dynamics of Bangladeshi political discourse. Through this exploration, the research seeks to deepen our understanding of how sentiment operates in the unique sociopolitical context of Bangladesh and, by extension, contribute to a broader comprehension of news consumption in the digital age.

1.2 Motivation

This research project finds its impetus in a profound acknowledgment of the distinctive challenges encountered in the realm of sentiment analysis within the intricate tapestry of Bangladeshi political news. Recognizing the intricate linguistic nuances, multifaceted cultural contexts, and the rich diversity of information sources, we are confronted with a complex landscape that necessitates a nuanced and tailored approach. This realization serves as the driving force behind our study, motivated by a commitment to bridge the existing gap in understanding and unravel the intricacies inherent in sentiment analysis within this unique sociopolitical context.

The linguistic fabric of Bangladeshi political discourse is woven with subtleties that demand more than conventional sentiment analysis methodologies. Phrases, idioms, and contextual cues, deeply rooted in the nation's history and culture, contribute to the complexity of sentiment expression. Moreover, the diverse sources of information, ranging from traditional news outlets to social media platforms, add layers of intricacy to the sentiment dynamics.

To navigate this intricate landscape effectively, our research adopts a comprehensive and innovative approach. Manual annotation stands as a fundamental component, enabling the incorporation of human insights to decipher sentiment nuances that automated processes may overlook. This human-centric layer ensures the cultural and contextual depth required for a more refined analysis of sentiment in Bangladeshi political news.

Machine learning algorithms, including Support Vector Machines, Naive Bayes, Random Forest, and Logistic Regression, complement the manual annotation by providing a systematic and quantitative framework. Trained on the annotated dataset, these algorithms empower the model to discern patterns, relationships, and sentiment trends within the vast corpus of political news.

Additionally, advanced techniques such as the Bag of Words model play a pivotal role in refining the analysis. This approach, which dissects the textual content into a matrix of word frequencies, allows for a more granular examination of linguistic patterns and their correlation with sentiment. By integrating these diverse methodologies, our study aspires not only to overcome the challenges posed by Bangladeshi political news but also to pave the way for a more holistic understanding of sentiment dynamics in political discourse.

In essence, our research motivation extends beyond mere acknowledgment of challenges; it encapsulates a commitment to unraveling the layers of sentiment intricacies in Bangladeshi political news. Through a multidimensional and comprehensive methodology, we aim to contribute valuable insights that transcend the conventional boundaries of sentiment analysis and establish a foundation for informed discourse and decision-making within the unique context of Bangladeshi politics.

1.3 Objective

The overarching objective of this extensive study is to undertake a nuanced exploration of sentiment analysis within the intricate landscape of Bangladeshi political news. The study is motivated by a profound recognition of the challenges posed by linguistic subtleties, diverse cultural contexts, and the myriad of information sources present in the political discourse of Bangladesh. The primary aim is to contribute substantively to the understanding of sentiment dynamics in this unique sociopolitical environment.

To achieve this goal, the study will curate and expand a diverse dataset of Bangladeshi political news articles, ensuring representation from various sources and perspectives. This dataset will serve as the foundation for training and validating machine learning algorithms, incorporating both manually annotated sentiments and those derived through automated processes. The integration of manual annotation into the dataset is a crucial aspect, leveraging human insights to discern nuanced sentiment variations that may be challenging for automated algorithms to capture. This approach seeks to infuse the analysis with cultural and contextual depth, enhancing the model's capacity for nuanced sentiment interpretation.

The study will apply a repertoire of machine learning algorithms, including Support Vector Machines, Naive Bayes, Random Forest, and Logistic Regression, to systematically analyze the annotated dataset. The algorithms will be trained and fine-tuned to effectively predict sentiment patterns in Bangladeshi political news articles. Additionally, advanced techniques such as the Bag of Words model will be employed to refine the sentiment analysis further. This method dissects textual content into matrices of word frequencies, allowing for a more granular examination of linguistic patterns and their correlation with sentiment, providing insights beyond traditional methodologies.

Beyond algorithmic analysis, the study will uncover patterns in political news consumption by analyzing sentiments across diverse sources and platforms, including traditional news outlets and social media. The goal is to explore how sentiment varies based on the source, political stance, and the level of public engagement. Leveraging insights derived from sentiment analysis, the study aims to construct an efficient recommendation system for Bangladeshi political news, tailoring the system to individual user preferences and fostering a more personalized and informed news consumption experience.

In its entirety, this research aspires to contribute both academically and practically by addressing the unique challenges posed by Bangladeshi political news sentiment analysis. The intention is not only to unravel the layers of sentiment intricacies within this context but also to establish a foundation for informed discourse and decision-making, shaping the landscape of political news consumption in Bangladesh. As part of the commitment to transparency and collaboration, the study plans to publicize the labeled dataset, along with insights gained, on widely accessible platforms, such as Kaggle, encouraging further advancements in the field of sentiment analysis within the Bangladeshi political context.

Chapter 2

Literature Review

The literature review, an expansive journey into the complex interplay of sentiment analysis, machine learning algorithms, and recommendation systems within the realm of Bangladeshi political news, now extends its scrutiny even further. This comprehensive investigation not only illuminates critical themes and seminal studies but broadens its scope to encompass the broader landscape of political news consumption, recognizing the symbiotic relationship between sentiment analysis and recommendation systems. By delving into the intricacies of these interconnected domains, the review not only elucidates established methodologies but also spotlights the dynamic evolution of these frameworks, accentuating the complex and multifaceted nature of sentiment dynamics in this distinctive sociopolitical context.

Diverse methodological approaches have been observed across various studies in the field. For example, the work of Yu and Hatzivassiloglou (2003) [3] places emphasis on the distinction between subjective texts and those conveying factual information. In this context, the latter approach assumes the inherently opinionated nature of text, leading to its classification into distinct sentiment categories—typically positive or negative—as outlined by Pang and Lee (2008) [4].

On a parallel track, researchers have directed their focus towards the intricate task of discerning an author's sentiment, specifically whether it leans positively or negatively, regarding a particular topic or object. This line of inquiry involves a nuanced exploration of the subtleties embedded within language to capture the sentiment nuances associated with specific entities. [5]

To execute these varied tasks, a combination of both supervised and unsupervised approaches has been employed, as noted by Feldman (2013). The utilization of supervised techniques involves training models on labeled datasets, allowing them to learn and predict sentiment based on predefined categories. In contrast, unsupervised approaches leverage inherent patterns and structures within the data to uncover sentiments without the need for preexisting labels. The choice between these approaches depends on the specific nuances of the sentiment analysis task at hand and the availability of labeled data.

In essence, the literature reveals a rich tapestry of approaches, each tailored to address specific aspects of sentiment analysis, from distinguishing subjective and objective content to unraveling an author's nuanced opinions. This diversity not only underscores the complexity of sentiment analysis but also highlights the importance of selecting methodologies that align with the specific objectives of a given study.

The conventional structure of sentiment analysis, marked by its two-phase methodology of data collection and subsequent sentiment extraction, remains foundational. As evidenced by prior studies [6], this systematic approach provides a foundational understanding of how sentiments unfold in response to the nuanced developments in Bangladeshi political discourse. Acknowledging the limitations of this traditional framework, the literature review emphasizes the pressing need for advanced methodologies, particularly machine learning algorithms. This acknowledgment is crucial to navigating the intricate nuances deeply embedded within the complex fabric of Bangladeshi political discourse, reflecting the evolution of sentiment analysis in the ever-dynamic digital age.

While conventional recommendation systems predominantly focused on generic news consumption patterns, the literature review extends its narrative to underscore the symbiotic relationship between sentiment analysis and recommendation systems. This fusion transcends a mere strategy; it signifies a transformative approach to elevate user-specific recommendations, leveraging nuanced insights derived from sentiment analysis. In recognizing global research efforts in this domain, with examples spanning Brazil [7], Turkiye [8], and various international landscapes, the literature underscores the universality of the challenges faced in deciphering sentiment dynamics. This emphasis not only reinforces the need for context-specific solutions but also paves the way for a more holistic understanding of how sentiment influences news consumption patterns on a global scale.

Within the sentiment analysis framework, the literature elevates the discussion by focusing on the prevalent utilization of the Bag of Words approach. Acknowledged as a widely adopted model celebrated for its efficacy in capturing word frequencies and patterns within political news discourse [9], the Bag of Words model emerges as a practical and adaptable tool. Its simplicity and effectiveness facilitate the distillation of complex textual data into manageable components, providing a nuanced interpretation of sentiments expressed within the dynamic political sphere. Additionally, the review further scrutinizes the consistent use of specific machine learning algorithms, such as Support Vector Machines (SVM) and Naive Bayes, across diverse sentiment analysis projects [10]. This observed consistency not only underscores the reliability and adaptability of these algorithms but also prompts a deeper inquiry into their efficacy within the multifaceted landscape of political news sentiments.

Expanding the scope even further, the literature review illuminates the challenges posed by linguistic subtleties, cultural context, and the diverse array of information sources within Bangladeshi political news. This nuanced understanding is crucial for refining sentiment analysis methodologies and tailoring them to the specific intricacies of the sociopolitical landscape. By acknowledging these challenges, the literature review sets the stage for a more context-aware approach to sentiment analysis.

Moreover, it is imperative to recognize the ethical dimensions that accompany sentiment analysis, particularly in the politically charged context of news consumption. The literature review delves into discussions surrounding the responsible use of sentiment analysis, considering issues such as bias, privacy, and the potential impact on democratic processes. This ethical dimension adds depth to the understanding of sentiment analysis within the broader framework of political news consumption.

The literature review played a pivotal role in shaping and informing various aspects of this study on sentiment analysis in Bangladeshi political news, incorporating machine learning algorithms and a recommendation system. The review provided a comprehensive understanding of conventional sentiment analysis methodologies, emphasizing the two-phase structure involving data collection and sentiment extraction. This foundational knowledge guided the study's methodological approach, ensuring a systematic and structured analysis of sentiment in Bangladeshi political news.

The literature review highlighted the evolving nature of sentiment analysis, advocating for the incorporation of advanced methodologies, particularly machine learning algorithms. This insight influenced the study's decision to employ algorithms such as SVM, Naive Bayes, Random Forest, and Logistic Regression, enhancing the analytical depth and accuracy.

In this research, we adopted the first approach, and relied on a group of annotators to classify the polarity of news, along with the entity to which it refer. To do so, textswere segmented in paragraphs, instead of sentences, so as to offer annotators a wider context in which to work. Given that our intention was to cover political news in Bangladesh from a greater variety of news producers (so as to allow for a reasonable comparison amongst them), we had to collect a corpus of our own, for example, however important the existing dataset are available, do not fit perfectly our purposes, either because the amount of political news dataset is still small, or because the corpus focus in different category or multi category.

Alternatively to the use of human annotators, another approach found is the use of external sources of information to classify the news. This is the approach taken by [11], who used stock market fluctuations to determine the polarity of news related to some specific stock. As such, if the stock price raised after the news, then that news is regarded as positive, otherwise, it is negative. However solving the problem of low inter annotator agreement scores, this kind of approach raises issues of its own. In this specific case, one can never be too sure about the time that it takes for the news to produce An Annotated Corpus for Sentiment Analysis in Political News any measurable impact on the stock market, there being a potential confounding between the news content and other external variables that might have influenced the sock prices,but which are unrelated to the news itself

Besides defining the methodology underlying the classification of news, another related question is at what level the news are to be annotated. Since news can referto multiple facts and, consequently, have multiple polarities, splitting them in smaller units of annotation might help capture each of these individual facts. This, however, is still an unsettled issue, with current approaches ranging from segmenting news in sentences (e.g. [12]) to separating out textspans.

In summation, the literature review not only serves as a comprehensive guide to the multifaceted nature of sentiment analysis within Bangladeshi political news but also broadens its horizons to encompass the ethical considerations inherent in this analytical process. By integrating insights from global research endeavors, advocating for advanced methodologies, recognizing the symbiotic relationship between sentiment analysis and recommendation systems, and addressing ethical dimensions, the review establishes a robust foundation. It not only opens avenues for future research endeavors and practical applications but also underscores the critical role sentiment analysis plays in shaping the landscape of political news consumption, offering a balanced perspective that accounts for both technological advancements and ethical considerations.

Chapter 3

Methodology

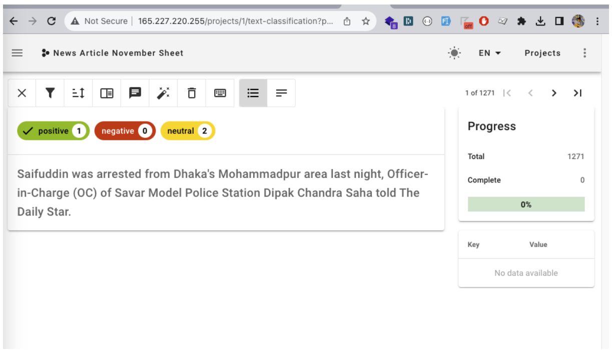
3.1 Data Collection:

1. Selection of News Outlets: A diverse set of Bangladeshi online English newspapers is chosen for data collection to ensure representation across various perspectives and sources.

2. Automated Crawler Script: An automated script is developed to crawl political content from selected news outlets. The script parses the content and organizes it into paragraphs, generating a CSV format for further processing.

3.2 Data Categorization and Annotation

Manual Annotation using Doccano: The collected paragraphs are manually labeled with sentiment flags using the Doccano Tool. The annotation process involves categorizing sentiments into positive, negative, or neutral to create a labeled dataset for training machine learning algorithms.



Not Secure | 165.227.220.255/projects/1/text-classification?p...
News Article November Sheet
EN
Projects
1 of 1271
✓ positive 1
negative 0
neutral 2
Progress
Key
Value
No data available

Saifuddin was arrested from Dhaka's Mohammadpur area last night, Officer-in-Charge (OC) of Savar Model Police Station Dipak Chandra Saha told The Daily Star.

Total 1271

Complete 0

0%

Figure 3.1: Docanno Annotation Tool

3.3 Machine Learning Algorithm Selection

Support Vector Machines (SVM), Naive Bayes, Random Forest, and Logistic Regression are chosen as machine learning algorithms for sentiment analysis due to their proven effectiveness in classification tasks. [13] The labeled dataset is used to train each algorithm, optimizing model parameters and ensuring adaptability to the nuances of Bangladeshi political news sentiments.

3.4 Approach : Bag of Words

The "Bag of Words" approach is a commonly used technique in natural language processing and sentiment analysis. This approach simplifies the complexity of textual data by representing it as an unordered set of words, disregarding grammar and word order but maintaining the frequency of word occurrences. The name "Bag of Words" implies that the focus is on the presence and frequency of words in a document, treating the document as an unordered collection or "bag" of words.

1. Tokenization:

   The first step is to break down a piece of text (document or paragraph) into individual words or tokens. This process, known as tokenization, separates the text into meaningful units, typically words. Punctuation and other non-alphabetic characters are often removed during this stage.

2. Vocabulary Creation:

   A vocabulary is then created by compiling a unique set of all the words present in the entire corpus (collection of documents). Each word in the vocabulary is assigned a unique identifier or index. This step establishes the basis for representing documents in a numerical format. Document Representation:

   Each document is represented as a vector where each element corresponds to a word in the vocabulary, and the value of each element is the frequency of that word in the document. This results in a high-dimensional, sparse vector, where most elements are zero because a document typically contains only a small subset of the entire vocabulary.

3. Vectorization:

   The process of converting each document into a vector representation is known as vectorization. This step is crucial for applying machine learning algorithms, as it transforms the text data into a numerical format that algorithms can process.

4. Feature Matrix:

   The collection of document vectors forms a feature matrix, where each row corresponds to a document, and each column corresponds to a unique word in the vocabulary. This matrix is then used as input for machine learning algorithms.

5. Analysis and Modeling:

   The feature matrix is utilized to train machine learning models for various tasks, such as sentiment analysis. The models learn to identify patterns in the frequency of words associated with different sentiments (positive, negative, or neutral).

6. Prediction:

   Once the model is trained, it can be applied to new, unseen documents. These documents are vectorized using the same vocabulary, and the trained model predicts the sentiment based on the patterns it learned during training.

3.5 Recommendation System Construction

Insights from sentiment analysis are harnessed to construct an efficient recommendation system for Bangladeshi political news. Our analysis provides insights into the computational efficiency and predictive performance of different machine learning algorithms in the context of building a recommendation system. The analysis helps in selecting an algorithm based on the trade-off between training time and accuracy. Understanding the time required for model training assists in resource allocation. Depending on the available computational resources and the real-time nature of the recommendation system, developers can make informed decisions on algorithm implementation. The insights gained from the analysis can guide decisions regarding the scalability of the recommendation system. If the system needs to handle a large volume of data or real-time updates, algorithms with shorter training times might be preferred.

Efficient algorithms reduce the time and resources required for model maintenance and updates. Regular updates to recommendation models are common to adapt to changing user preferences, and understanding the training time helps in planning these maintenance activities. The analysis facilitates an iterative development approach. Developers can experiment with different algorithms, monitor their impact on system performance, and make continuous improvements based on the evolving needs of the recommendation system.

Chapter 4

Algorithms and Techniques

4.1 Algorithm

The following algorithm has been used to make the training model and analysing the model for accuracy. The example includes SVM but the process is pretty much similar for other algorithms too.

Algorithm 1: Sentiment Analysis using SVM

```
Input: Manually annotated data $D_1$ (file_path), Training data $D_2$ (file_path_set_2);
Output: SVM Model;
Load data from $D_1$ and $D_2$;
Process data from $D_1$: Extract features and labels from $D_1$;
Convert text labels to numerical labels;
Split $D_1$ into training and testing sets;
Process data from $D_2$: Extract features and labels from $D_2$;
Merge processed data from $D_1$ and $D_2$ to get the final training data;
Training: Create a CountVectorizer and transform the texts to a bag-of-words
  representation;
Train a SVM classifier with a linear kernel;
Evaluation: Evaluate the model on the test set from $D_1$;
Example Prediction: Predict sentiment for the example text;
```

4.2 Technique

We have used the Bag of Word approach for training the models. The BoW approach provides a way to represent text data in a format that machine learning models can handle, and it's particularly effective for tasks like text classification. The simplicity of BoW makes it computationally efficient, but it may not capture more complex semantic relationships.

Chapter 5

Implementation and Code

The provided Python code performs sentiment analysis using a Support Vector Machine (SVM) classifier with a linear kernel. It utilizes a bag-of-words (BoW) approach for text representation, employing the scikit-learn library's CountVectorizer to convert the training texts into numerical vectors.

5.1 SVM Code

Listing 5.1: SVM Python code

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder
import pandas as pd
import time
import datetime

# dataset 1 our manually annotated data.
file_path = '/path/to/set1.csv'

# dataset 2 training (Basic sentiment examples of Common English Sentences)
file_path_set_2 = '/path/to/set2.csv'

df = pd.read_csv(file_path)
df_set2 = pd.read_csv(file_path_set_2)
```

```
# array processing set 1
list_data = df.values

train_data = []
for x in list_data:
    train_data.append((x[1], x[2]))

# Separate features (texts) and labels
train_texts = [text for text, label in train_data]
train_labels = [label for text, label in train_data]

# Convert text labels to numerical labels
label_encoder = LabelEncoder()
train_labels_encoded = label_encoder.fit_transform(train_labels)

# Split the data into training and testing sets
X_train, X_test_before, y_train, y_test_before = train_test_split(train_texts
                                                                  random_sta

# array processing set 2 and merging to train_data
list_data_set2 = df_set2.values

for x in list_data_set2:
    train_data.append((x[2], x[3]))

# train_data variable is your final processed array for training.

print("Started␣training␣model␣with␣SVM:␣", datetime.datetime.fromtimestamp(ti
# Separate features (texts) and labels
train_texts = [text for text, label in train_data]
train_labels = [label for text, label in train_data]

# Convert text labels to numerical labels
train_labels_encoded = label_encoder.transform(train_labels)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(train_texts, train_labels
# Create a CountVectorizer and transform the texts to a bag-of-words represe
vectorizer = CountVectorizer()
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test_before)  # Use the same vect

# Train a SVM classifier
model = SVC(kernel='linear')
model.fit(X_train_vectorized, y_train)

print("Training␣completed.␣", datetime.datetime.fromtimestamp(time.time()))

# Training Accuracy
print("Accuracy␣of␣Train␣Data")
print(model.score(X_train_vectorized, y_train))

# Test Accuracy

# Evaluate the model on the test set
y_pred = model.predict(X_test_vectorized)

print("Accuracy␣of␣Test␣Data:␣")
print(accuracy_score(y_test_before, y_pred))
```

5.2 Logistic Regression Code

For the rest of the algorithms the code changes slightly. For logistic regression, we have to set a max iteration count.

Listing 5.2: Logistic Regression

```
# Train a Logistic Regression classifier
model = LogisticRegression(max_iter=1000)
model.fit(X_train_vectorized, y_train)
```

Chapter 6

Result and Analysis

6.1 Training and Test Accuracy

We trained the model as outlined in the code across four different algorithms. The accuracy values for both training and test datasets are given for four different models: Logistic Regression, SVM (Support Vector Machine), RandomForest, and Naive Bayes. The following formula has been used to calculate the accuracy:

$Accuracy = \frac{Number of Correct Predictions}{Total Number of Predictions}$

Table 6.1: Accuracy across Models with 1000 dataset

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.932	0.822
SVM	0.948	0.862
Random Forest	0.995	0.826
Naive Bayes	0.850	0.776

Table 6.2: Accuracy across Models with 3000 dataset

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.910	0.791
SVM	0.933	0.835
Random Forest	0.925	0.760
Naive Bayes	0.834	0.755

6.2 Classification Report

The classification report provides a comprehensive evaluation of the performance of a classification model. It goes beyond a single metric (such as accuracy) and includes metrics like precision, recall, F1-score, and support for each class.

Table 6.3: Weighted Average Metrics for Each Algorithm

Algorithm	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
SVM	0.88	0.86	0.86
RandomForest	0.88	0.83	0.84
Logistic Regression	0.82	0.82	0.82
Naive Bayes	0.80	0.78	0.77

These metrics (precision, recall, and F1-score) provide a balanced view of the model's performance. Precision measures the accuracy of positive predictions, recall measures the ability to capture true positives, and the F1-score is the harmonic mean of precision and recall, offering a balanced measure between the two.

6.3 Performance

The Naive Bayes algorithm demonstrated the fastest training time among the models, taking only 0.15 seconds, while SVM, RandomForest, and Logistic Regression required 25.71 seconds, 16.79 seconds, and 2.61 seconds, respectively. However, we have seen higher accuracy with SVM even though it performed very slow.

Table 6.4: Time Taken to Build Models

Algorithm	Time Difference (seconds)
SVM	25.706279
RandomForest	16.788389
Logistic Regression	2.613934
Naive Bayes	0.146225

In conclusion, the Naive Bayes algorithm exhibited the quickest training time, showcasing efficiency in model building. However, the SVM algorithm, despite requiring the longest training time, achieved the highest test accuracy, highlighting a trade-off between training duration and predictive performance. The choice of algorithm may depend on the specific requirements of the application, balancing computational resources with desired model accuracy.

References

[1] "Kaggle common english sentence dataset from twitter." https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis. Accessed:2024-12-22.

[2] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, pp. 43–52, 2010.

[3] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 129–136, 2003.

[4] B. Pang, L. Lee, *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[5] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373, 2004.

[6] C. Soo-Guan Khoo, A. Nourbakhsh, and J.-C. Na, "Sentiment analysis of online news text: A case study of appraisal theory," *Online Information Review*, vol. 36, no. 6, pp. 858–878, 2012.

[7] G. D. de Arruda, N. T. Roman, and A. M. Monteiro, "An annotated corpus for sentiment analysis in political news," in *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pp. 101–110, SBC, 2015.

[8] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 174–180, IEEE, 2012.

[9] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, “More than bags of words: Sentiment analysis with word embeddings,” *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 140–157, 2018.

[10] A. Mukwazvure and K. Supreethi, “A hybrid approach to sentiment analysis of news comments,” in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*, pp. 1–6, IEEE, 2015.

[11] M. Siering, “" boom" or" ruin"–does it make a difference? using text mining and sentiment analysis to support intraday investment decisions,” in *2012 45th Hawaii International Conference on System Sciences*, pp. 1050–1059, IEEE, 2012.

[12] E. Refaee and V. Rieser, “Benchmarking machine translated sentiment analysis for arabic tweets,” in *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: student research workshop*, pp. 71–78, 2015.

[13] S. Rana and A. Singh, “Comparative analysis of sentiment orientation using svm and naive bayes techniques,” in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pp. 106–111, IEEE, 2016.

Generated using Research LaTeX Template, Version 1.0. Department of Computer Science and Engineering, Southeast University, Bangladesh.

The Original Version was Developed by Tashreef Muhammad on Friday 17th November, 2023

This report was generated on Monday 19th February, 2024 at 02:13:27.