Mini-Project

Of

Data Mining & Warehousing

(CSN447)

Master Of Computer Applications

Academic Session 2025-2026

Submitted to:

Mr. Rakesh Kumar Saini

Assistant Professor

School Of Computing

Submitted By:

Shashank Bhardwaj

MCA – 2nd Sem

SAP ID- 1000024713

School Of Computing

DIT University, Dehradun

CERTIFICATE

Certified that project report entitled "**Twitter Sentiment Analysis**" submitted by "**SHASHANK BHARDWAJ (SAP ID:-1000024713)"** during the period 2024-2025 in partial fulfilment of the requirements for the award of degree of MCA of DIT University, Dehradun, is a record of work carried out under my guidance and supervision. The project report embodies result of referred work and studies carried out by student themselves and the content of the report do not form the basis for the award of any other degree to the candidate or to anybody of the team.

Mr. Rakesh Kumar Saini

Asst. Professor

DIT University, Dehradun

ACKNOWLEDGEMENT

Apart from the efforts we put, the success of the project depends largely on the encouragement and guidelines of many others. We take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

We take immense pleasure in thanking **Mr Rakesh Kumar Saini** for having permitted us to carry out this project. We would like to express a deep sense of gratitude to **Dr.Bharti Sharma, Head of Department, MCA**, for their able guidance and useful suggestions, which helped us in completing the project in time.

Finally, yet importantly, we would like to express our heartfelt thanks to our parents for their blessings, our friends for their help and wishes for the successful completion of this project.

INDEX

Introduction

In today’s digital age, social media platforms like Twitter have become rich sources of public opinion and sentiment. Analyzing these sentiments can provide valuable insights into consumer behavior, political trends, and social movements. In this project, we will perform sentiment analysis on Twitter data using Weka, a popular open-source tool for data mining and machine learning.

Sentiment analysis involves determining the emotional tone behind a series of words, with applications ranging from brand monitoring to public opinion analysis. By processing a dataset of tweets, we can categorize each tweet as either positive, negative, or neutral based on its content.

Using Weka, we will leverage various machine learning algorithms, such as Naive Bayes, Decision Trees, and Support Vector Machines (SVM), to classify tweets and evaluate the performance of each model. The goal is to identify the most effective approach for analyzing sentiment in short, informal text typical of Twitter posts.

Software Requirements

Performing Twitter sentiment analysis using Weka:

1. **Java Development Kit (JDK)**
   - **Version:** JDK 8 or above
   - **Purpose:** Weka is a Java-based tool, so you'll need the JDK installed on your system to run Weka and its associated functionalities.

2. **Weka**
   - **Version:** Weka 3.8.x or later
   - **Purpose:** Weka is used for building and evaluating machine learning models. It includes algorithms for classification, clustering, and data preprocessing, which are essential for the sentiment analysis task.

ARFF FILE CODE

```
@relation twitter_sentiment


@attribute tweet string

@attribute sentiment {positive, negative, neutral}


@data

"Just got a new phone, love it!", positive

"I hate waiting in long lines.", negative

"Weather is nice today.", neutral

"Excited for the weekend!", positive

"Feeling down, not a good day.", negative

"Had a great lunch with friends!", positive

"Not sure if I should go to the party.", neutral

"Everything went wrong today!", negative
```

Process for analyzing this data:-

1. **Prepare Your Dataset**:
   - Save the example ARFF content (or your own Twitter data) into a `.arff` file using any text editor (e.g., `twitter_sentiment.arff`).
2. **Load the Dataset into Weka**:
   - Open Weka.
   - Go to the **Explorer**.
   - Click on **Open file...** and load your `twitter_sentiment.arff` file.
3. **Preprocessing the Data (Optional)**:
   - In Weka, you can use filters to clean or modify the data, such as removing stop words or converting all text to lowercase.
   - You may also use the **StringToWordVector** filter, which will transform text into a bag-of-words representation.

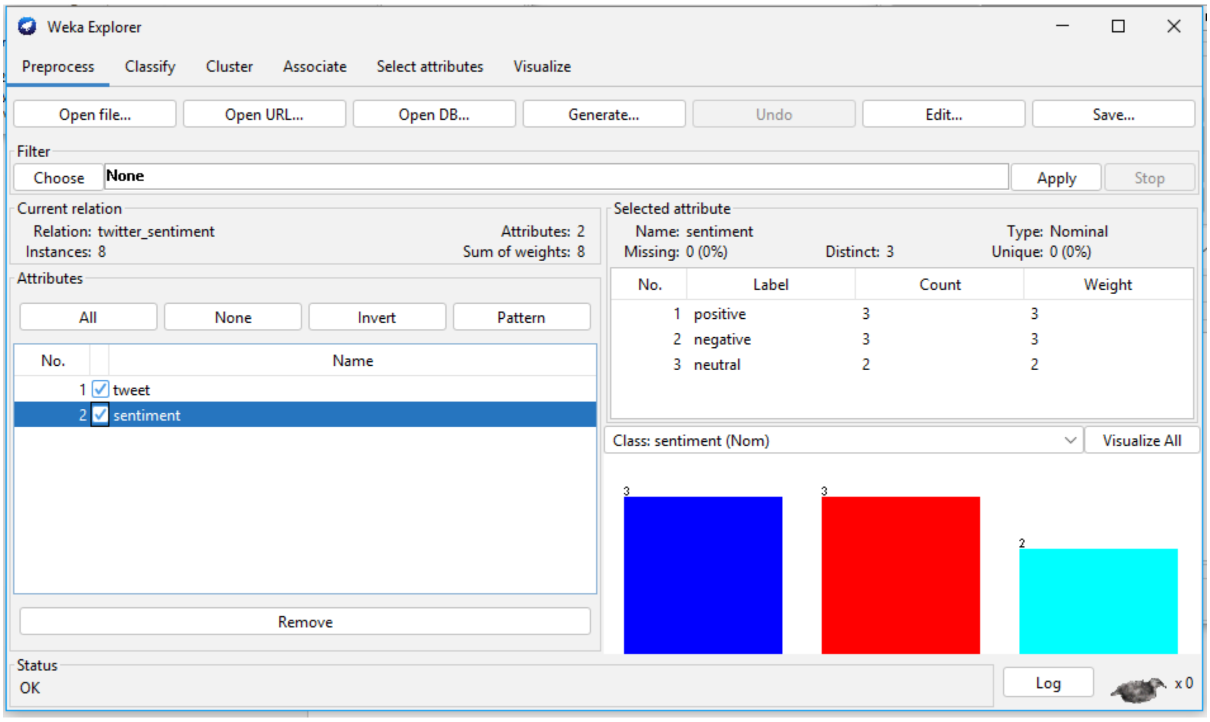
   Steps to apply the **StringToWordVector** filter:

   - Click on the **Preprocess** tab.

- In the **Filter** section, choose `Supervised > Attribute >StringToWordVector`.
- Configure the filter (e.g., set the maximum number of words, remove stopwords, etc.).
- Click **Apply**.

4. **Build a Classifier**:
   - Once the dataset is loaded, switch to the **Classify** tab.
   - Select the classifier you want to use (e.g., **Naive Bayes**, **J48**, **SVM**, etc.).
   - Set the class attribute (this is the sentiment attribute).
   - Click **Start** to train the model.
5. **Evaluate the Model**:
   - After training, Weka will display the evaluation results (such as accuracy, precision, recall, F1 score).
   - You can also use **Cross-validation** to assess the model's performance.
6. **Use the Model for Prediction**:
   - To predict the sentiment of new tweets, you can click on **Classify**>**Supplied test set**, and provide a new ARFF file with the tweets you want to classify.

OUTPUT



Filter: None

Current relation
Relation: twitter_sentiment
Instances: 8
Attributes: 2
Sum of weights: 8

Attributes

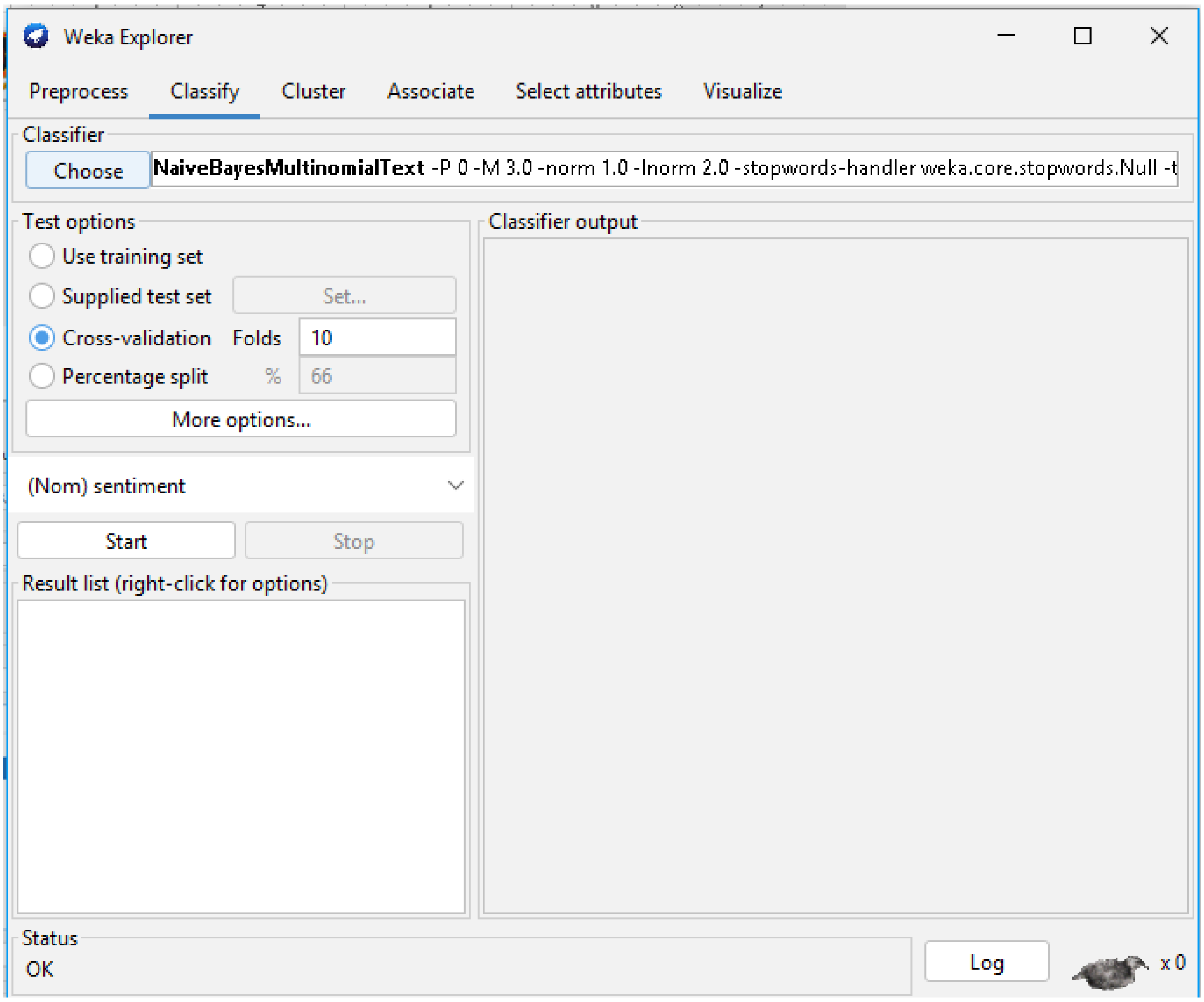
No.	Name
1 ☑	tweet
2 ☑	sentiment

Selected attribute
Name: sentiment
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	positive	3	3
2	negative	3	3
3	neutral	2	2

Class: sentiment (Nom)

Status
OK

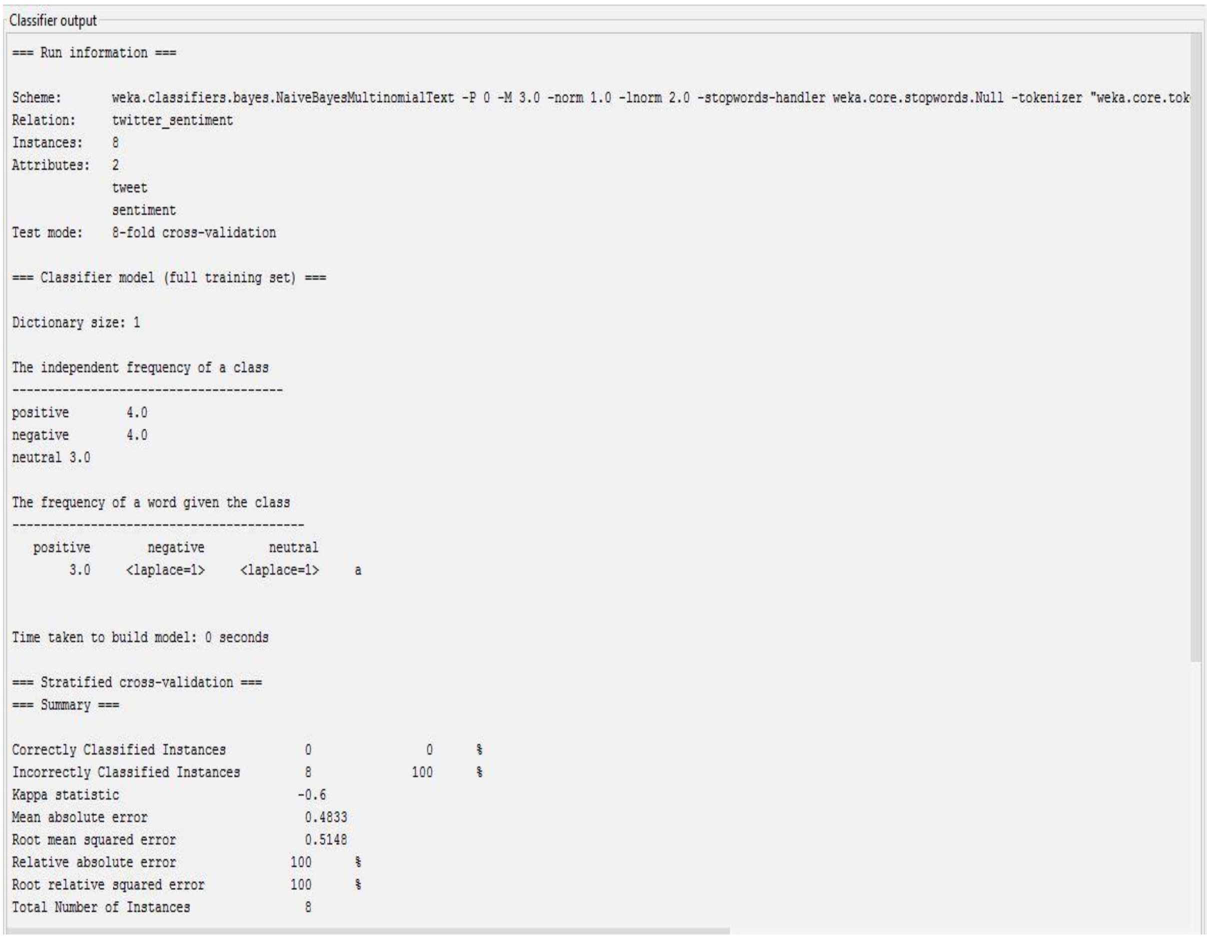


NaiveBayesMultinomialText -P 0 -M 3.0 -norm 1.0 -lnorm 2.0 -stopwords-handler weka.core.stopwords.Null -t

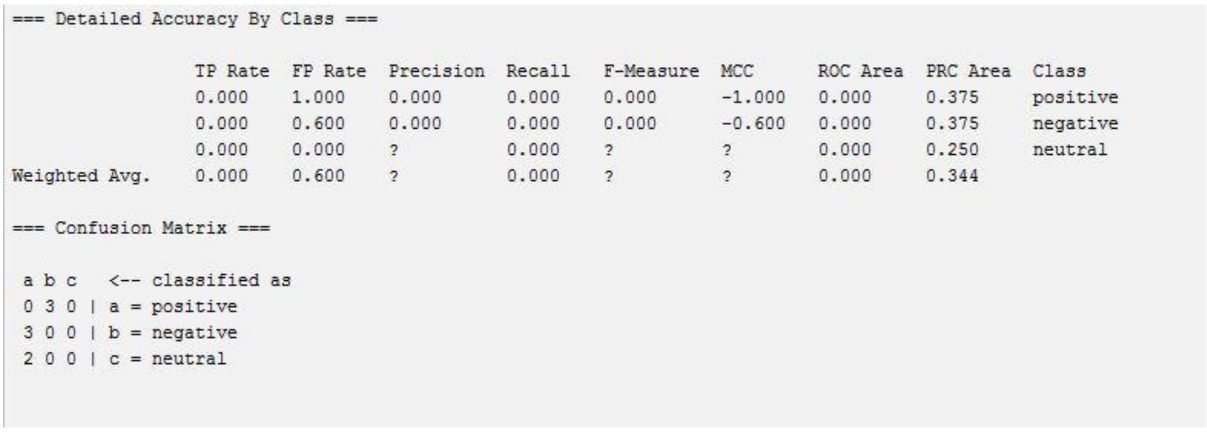
● Cross-validation Folds 10

(Nom) sentiment

OK



```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomialText -P 0 -M 3.0 -norm 1.0 -lnorm 2.0 -stopwords-handler weka.core.stopwords.Null -tokenizer "weka.core.tok
Relation:     twitter_sentiment
Instances:    8
Attributes:   2
              tweet
              sentiment
Test mode:    8-fold cross-validation

=== Classifier model (full training set) ===

Dictionary size: 1

The independent frequency of a class
--------------------------------------
positive        4.0
negative        4.0
neutral 3.0

The frequency of a word given the class
-----------------------------------------
   positive        negative         neutral
       3.0     <laplace=1>     <laplace=1>     a


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         8              100      %
Kappa statistic                         -0.6
Mean absolute error                      0.4833
Root mean squared error                  0.5148
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances                8
```



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    1.000    0.000      0.000    0.000      -1.000   0.000     0.375     positive
                 0.000    0.600    0.000      0.000    0.000      -0.600   0.000     0.375     negative
                 0.000    0.000    ?          0.000    ?          ?        0.000     0.250     neutral
Weighted Avg.    0.000    0.600    ?          0.000    ?          ?        0.000     0.344

=== Confusion Matrix ===

 a b c   <-- classified as
 0 3 0 | a = positive
 3 0 0 | b = negative
 2 0 0 | c = neutral
```

Conclusion

The sentiment analysis of Twitter data, as represented by the given ARFF file, allows for the classification of tweets into three distinct sentiment categories: positive, negative, and neutral. Through the application of machine learning algorithms in Weka, the model can learn to identify the underlying sentiment in new, unseen tweets based on the words and phrases used.

From the sample data provided, it is evident that the sentiment of a tweet can be inferred from context, vocabulary, and tone. For example:

- Positive sentiments are often expressed through words like "love," "great," or "excited."
- Negative sentiments typically involve words like "hate," "down," or "wrong."
- Neutral tweets appear to be more neutral in tone, lacking strong emotional language, such as in "Weather is nice today" or "Not sure if I should go to the party."

With the dataset being small and simple, further training with a larger, more complex dataset is likely to improve the model's accuracy and predictive capabilities.

Application

The sentiment analysis model developed in this project can be applied to several real-world use cases, including:

1. **Brand Monitoring:** Companies can use sentiment analysis to monitor Twitter conversations about their brand. By identifying positive and negative sentiments in tweets mentioning the brand, businesses can track their public image and respond proactively to customer feedback.
2. **Customer Feedback Analysis:**Analyzing tweets about products or services provides valuable insights into customer satisfaction. This can guide product development, customer service strategies, and marketing campaigns.
3. **Market Research:** By analyzing public sentiment around certain topics, products, or events, companies and researchers can gauge consumer interest, identify trends, and make informed business decisions.
4. **Political and Social Movements Analysis:** Sentiment analysis can be used to track the sentiment around political events, candidates, or social movements. This provides insight into public opinion and can assist in gauging public reaction during elections, protests, or major social events.
5. **Mental Health Monitoring:** Twitter sentiment analysis could also be applied to detect signs of mental health distress by identifying negative sentiments in tweets. This could potentially be used by social organizations to reach out and offer support to individuals exhibiting signs of distress.

Reference

•**Hall, M. A., & Frank, E. (2008).***Applied Machine Learning: Weka for Data Mining.* In *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

- This book provides a comprehensive introduction to machine learning techniques using Weka, which is useful for understanding the algorithms applied in this sentiment analysis project.

•**O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010).***From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010.

- This paper discusses how sentiment analysis can be used to analyze public opinion on Twitter, providing context and methodology for sentiment analysis in social media data.