Chapter 20
Machine Learning-Based Sentiment Analysis of Twitter Using Logistic Regression

D. Kavitha
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Shyam Venkatraman
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Karthik CR
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Navtej S Nair
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

ABSTRACT

Twitter sentiment analysis is crucial for understanding public opinion in the digital age. This project employs logistic regression, a machine learning approach, to identify emotions in tweets from the Sentiment 140 dataset. Exploratory data analysis (EDA) identifies patterns in emotion distribution. Various machine learning algorithms, such as logistic regression, etc., are then used to classify tweets as good, negative, or neutral. Text preprocessing techniques prepare data, but TF-IDF weights words based on their significance. The challenges include capturing the complexities of human emotions while also keeping up with the ever-changing nature of Twitter data. Despite these limitations, data analysis and logistic regression provide important insights into public sentiment, assisting decision-making in a range of businesses. Looking ahead, the study emphasises the need for additional research to strengthen sentiment analysis methodologies. This includes addressing context-dependent emotions, adapting to diverse domains, and considering ethical issues such as partiality.

DOI: 10.4018/979-8-3693-3502-4.ch020

1. INTRODUCTION

Twitter is a highly regarded and prominent social media network that enables users to freely share their thoughts and feelings on a wide range of subjects, occasions, goods, or companies. Sentiment analysis on Twitter involves automatically recognizing and classifying the different types of emotions that people express through their tweets, such as neutral, negative, and positive. It is a significant use of natural language processing and machine learning since it can reveal insightful information about the beliefs and perspectives of Twitter users regarding various topics of interest. Additionally useful for monitoring brand reputation, managing crises, conducting marketing research, monitoring customer feedback, and analyzing public opinion is Twitter sentiment analysis, which can be utilized by businesses, politicians, researchers, and social media managers. But figuring out how to interpret the emotions in tweets is difficult since it necessitates a comprehension of user tone, context, and natural language. The lexicon-based, machine learning-based, and deep learning-based systems that are currently in use for Twitter sentiment analysis have several drawbacks, including aspect-level analysis, domain adaption, data scarcity, and complicated sentiment expressions. To overcome these obstacles and produce more accurate and trustworthy results, a more thorough and robust method for Twitter sentiment analysis is required. The goal of this project is to use a combination of data analysis and machine learning tools and techniques to create and assess a unique approach for Twitter sentiment analysis. The Sentiment140 dataset—which comprises 1.6 million tweets classified as good, negative, or neutral—will serve as the primary data source for the project. In addition, the primary data analysis tools for the project will be data visualization, sentiment lexicons, feature extraction, descriptive and inferential statistics, and text preprocessing. Five machine learning classifiers—Logistic Regression, Decision Tree, K-Nearest Neighbors, Multinomial Naive Bayes, and Support Vector Machine—will be compared for performance in this project, and the best one will be chosen based on measures such as accuracy and F1-score. The matplotlib and word cloud libraries will be used to show the word cloud as well as the distribution of positive, negative, and impartial words in tweets.

2. PROBLEM STATEMENT

Twitter is a popular social media network that lets users express their thoughts and feelings on a variety of topics, events, goods, and businesses. However, assessing Twitter sentiment is a complex and difficult operation because it needs comprehending the user's natural language, context, and tone. Existing Twitter sentiment analysis methods, such as lexicon-based, machine learning-based, and deep learning-based approaches, have drawbacks that include data quality, data scarcity, domain adaptability, aspect-level analysis, and complicated sentiment expressions. As a result, a more complete and robust data analysis technique for Twitter sentiment analysis is required, one that can overcome these problems while also providing important insights into Twitter users' opinions and attitudes. Using a variety of instruments and methods, including text preprocessing, feature extraction, sentiment lexicons, descriptive statistics, inferential statistics, and data visualization, the project's goal is to do a sentiment analysis of Twitter data. The following are the project's intended results:

- An examination of the tweets' sentiment labels: positive, negative, or neutral, that is both exploratory and descriptive