SURVEY ON SENTIMENT ANALYSIS

Jay Sankhe[1], Kaushal Batavia[2], Himanshu Borse[3], Prof. Shweta Sharma[4]

[1,2,3]Student, Dept. of Computer Engineering, Atharva College of Engineering, Mumbai, India
[4]Prof, Dept. of Computer Engineering, Atharva College of Engineering, Mumbai, India

---

Abstract - *Sentiment analysis is a method of contextual mining of reviews that extracts information that helps businesses to understand social reviews of their product or services. With advancements in machine learning technologies, we can analyze customer reviews and identify whether they are positive, negative, or neutral. Businesses use this information to know about customers' concerns about their products or services and take appropriate decisions to improve their services. In this paper, we have described the sentiment analysis process including its definition, datasets, preprocessing, algorithms used, evaluating algorithms, and conclusion. Evaluation metrics such as precision, recall, accuracy, f1 score are used to check the performance of the algorithms used. This research-based survey is divided into different sections where each section describes a particular step of sentiment analysis. A sentiment analysis survey on different platforms like social media, education, e-commerce, and Google play store is performed in this paper. Various Machine Learning algorithms like Naïve Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest are used to perform sentiment analysis.*

Key Words: Sentiment Analysis, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest.

1.INTRODUCTION

Sentiment Analysis is a way of knowing subjective information related to a particular product or service from text data. This analysis tells us whether the user reviews are positive, negative, or neutral for that product or service.

1.1 Motivation for Sentiment Analysis

While using any service or buying any product, a customer checks previous customers' reviews to know about the quality of the product. This means the customer decides the quality of the service based on previous service's reviews. So, businesses must study customer reviews to know their opinion about service. If negative reviews are more, then it is important to find out why their services have these types of reviews. For businesses, sentiment analysis is important to know about users' concerns about their product and take appropriate decisions to improve their services.

1.2 Basic Concept

The base concept is to collect reviews of particular products or services from different platforms like Amazon, Google Play store, Coursera, and Twitter. Perform sentiment analysis on these datasets using various Machine Learning algorithms. Use these results to know about customer's opinions about a particular product or service.

Customer reviews can be either be positive, negative, or neutral. This can be decided by analyzing words in the comment. From negative reviews, we can find out problems and solutions to these problems. In this way, businesses can improve their customer service and product sales.

1.3 Applications

A. E-commerce Websites:

E-commerce websites can increase their product sales by analyzing customer reviews. From reviews, owners can take decisions to improve their customer service and improve their product sales. Also, with the help of reviews, businesses can recommend products to users. If the product has more positive reviews, then there is a possibility that customers will buy that product as they feel trust by reading previous customers' reviews [1][2].

B. Social Media Platforms:

People share their opinions using various social media platforms. Twitter is one of the most popular used social media platforms. It is used by all level people including higher authorities of the country. At this time, every person may have different opinion about particular thing. Some people try to spread negativity around the society. It is important to detect hate speech in tweets. For simplicity, we can say hate speech means it has a racist or sexist sentiment associated with it. So, we have to classify racist or sexist tweets from other tweets [3].

C. Educational Platforms:

When a student goes for buying any course, they first check reviews of that course. Based on these reviews, they decide whether this course is easy to learn or not, whether to buy the course or not. If the course has more positive reviews, then more students will tend to buy that course. Education platforms can recommend the most populated courses based on previous reviews [4].

D. Mobile Applications:

While installing any mobile application from Google Play Store, user first check what past user's experience says about

this application. If application has bad reviews, then it results into user churn [5][6][7].

2. WORKFLOW OF SENTIMENT ANALYSIS

Every analysis starts with the data collection process. Nowadays, social media platforms give access to a wide range of data for research and analysis. Real-time data can be extracted from social media platforms using web scraping techniques. In Python, web scrapping is performed using the BeautifulSoup library. This extracted raw data is preprocessed to get cleaned data which simplifies further analysis process. Feature extraction is performed on preprocessed data to get more accurate results. Various algorithms are used to perform sentiment classification. This is the most important step in the whole process. The performance of algorithms is measured by evaluation metrics such as precision, recall, accuracy, and f1-score.

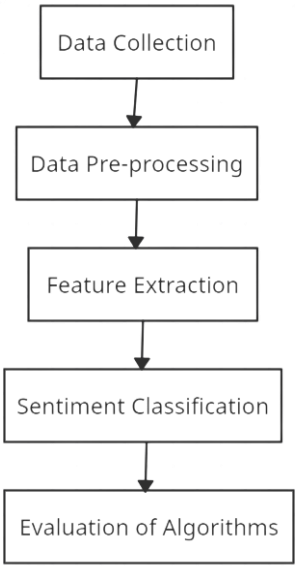

Fig -1: Workflow of Sentiment Analysis

Refer Fig. 1 for workflow of sentiment analysis.

Steps in Sentiment Analysis:

1. Data collection
2. Preprocessing
3. Feature Extraction
4. Sentiment Classification
5. Evaluation

3. MACHINE LEARNING APPROACH

Machine Learning comes under Artificial Intelligence where mathematical models are trained over text data. Here, learning is based on prior experience. Various machine learning algorithms are used for classification and regression. Data is collected, preprocessed, features extracted, models are trained and after evaluation, we get the required results.

4. PREPROCESSING AND FEATURE EXTRACTION

Data we get after the data collection process is raw data. It contains numbers, punctuation marks, URLs, emojis that will not help in the analysis process. So, we have to remove all these in preprocessing. Stop words are connective words that occur most of the time. Remove stop words as they do not help in classification. Finally, convert all text into lowercase.

Tokenization, Stemming, Lemmatization

In the tokenization process, we break the preprocessed data into a list of separate words called tokens. This is done to convert raw input data into numbers or vectors that will be understood by the model to achieve correct feature extraction.

Stemming is the process of converting words into their fundamental structure or root structure. One word can be used in different formats by adding suffix or prefix to it. Remove this suffix or prefix to get the root word.

Lemmatization is similar to stemming. Here we get roots without losing their dictionary meaning. It is a dictionary-based approach.

5. ALGORITHMS

Various algorithms can be used to perform sentiment analysis. They are Naïve Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest.

5.1 Naïve Bayes

Naive Bayes classifier is a supervised learning algorithm. It is based on Bayes' theorem. Naive Bayes is a probabilistic classifier as it predicts based on the probabilities of the object.

Bayes' theorem calculates the probability of hypothesis with prior knowledge. It calculates conditional probability called a posterior or revised probability. According to Bayes' rule, the probability of any two random variables A and B is given by,

$$P(A/B) = (P(B/A)\ P(A)) / (P(B))$$

Where,

P(A/B) = Posterior probability,

P(A) = Prior probability,

P(B) = Marginal probability,

P(B/A) = Likelihood probability

Naive Bayes theorem predicts a class value for a given set of attributes. For each known class value,

- Naive Bayes calculates probabilities for each attribute, conditional on the class value.
- It uses product rule to obtain a joint conditional probability for the attributes.
- It uses the Bayes rule to derive conditional probabilities for the class variable.

When all class values are calculated, output the class with the highest probability.

5.2 K-Nearest Neighbor (KNN)

KNN is a supervised learning technique that assumes the similarity between new data and available data. It put new data into the category that is most similar to it. It is used for both classification and regression.

5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that learns from the dataset and is used for classification. The goal of SVM is to find a decision boundary between two classes that are maximally far from any point in the training data. SVM assigns new data elements to one of the labeled categories.

For a given set of training examples, SVM builds a model that predicts whether a new example falls into one class or the other. SVM is also used in predictive analysis.

5.4 Random Forest

Confusion Random Forest is a supervised learning technique that can be used for both Classification and Regression. It combines multiple classifiers to solve a complex problem and to improve the performance of the model. Random forest combines multiple decision trees for prediction

Steps used in Random Forest:

1. Selection of random K data points from the training set.
2. Constructing decision trees associated with the selected data points.
3. Select N to construct N decision trees and repeat the above two steps.
4. Find the predictions of each decision tree. Assign these predictions as new data points to the category that wins the majority votes.

6. EVALUATION OF ALGORITHMS

All the algorithms discussed in the previous section are evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics are measured using a confusion matrix.

6.1 Confusion Matrix

Confusion matrix is a tool for analyzing how well our classifier can recognize tuples of different classes. Fig. 4 shows the confusion matrix. Refer Fig. 2.

		Predicted Class: P	Predicted Class: N
Actual Class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Fig -2: Confusion Matrix

6.2 Accuracy

Accuracy of any classification algorithm is the percentage of test set tuples that are correctly classified by the model.

$$\text{Accuracy} = ((TP + TN)) / ((TP + TN + FP + FN))$$

TP = True Positive, TN = True Negative

FP = False Positive, FN = False Negative

6.3 Precision

Precision means exactness of algorithm. It is the percentage of tuples that are correctly classified as positive are actual positive.

$$\text{Precision} = TP/ (TP + FP)$$

TP = True Positive, FP = False Positive

6.4 Recall

Recall is the measure of completeness. It is the percentage of positive tuples which the classifier labeled as positive.

$$Recall = TP/ (TP + FN)$$

TP = True Positive, FN = False Negative

After calculating the values of all these metrics for each algorithm, we decide which algorithm is best for predictions based on any of the metrics selected by the user. Using this best algorithm, we proceed for sentiment analysis.

7. CONCLUSIONS

Sentiment analysis contains some sentiments whose classification of text must be dealt with. This survey paper describes the steps in the sentiment analysis process that are named as data collection, data preprocessing, feature extraction, sentiment classification, and evaluation. We performed sentiment analysis for mobile applications available on Google Play Store, Amazon, Coursera, and Twitter. We have used Machine Learning algorithms to classify user reviews as positive, negative, or neutral. Depending on negative reviews, businesses can analyze the customers' concerns about their service. We have used Naive Bayes Classifier, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest algorithms for sentiment analysis. We evaluated the performance of all these algorithms based on performance metrics such as precision, recall, accuracy, and F1 score.

REFERENCES

[1] Dr S. N. Singh, T. Sarraf, "Sentiment Analysis of a Product based on User Reviews using Random Forests Algorithm", IEEE, 2020.

[2] B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O.N. Kushwaha, N. K. Shah, "Sentiments Detection for Amazon Product Review", IEEE, 2021.

[3] Chirag Kariya, Priti Khodke, "Twitter Sentiment Analysis", IEEE, 2020.

[4] R. S. Shudapreyaa, T. Yawanikha, A. Mahalakshmi, J. Mary Jenifer, G. Kavipriya, "Analysis of Udemy Courses based on Machine Learning Algorithm", IJCA, 2020.

[5] Indriati, Ari Kusyanti, Dea Zakia, "Sentiment Analysis in the Mobile Application Review Document Using the Improved K-Nearest Neighbor Method", IEEE, 2019.

[6] R. R. Putra1, M. E. Johan, E. R. Kaburuan, "A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia", IJATCSE, 2019.

[7] S. Ranjan, S. Mishra, "Comparative Sentiment Analysis of App Reviews", arXiv, 2020.

BIOGRAPHIES

Jay Sankhe
(BE Student) Atharva College of Engineering, Department of Computer Engineering, Mumbai, India

Himanshu Borse
(BE Student) Atharva College of Engineering, Department of Computer Engineering, Mumbai, India

Shweta Sharma
(Professor) Atharva College of Engineering, Department of Computer Engineering, Mumbai, India

Kaushal Batavia
(BE Student) Atharva College of Engineering, Department of Computer Engineering, Mumbai, India