Analyzing Sentiment Trends of Ajman University Using Machine Learning Models

Kajal Najeema Sajudeen, Keerthana Lakshmanan
Dapartment of Information Technology

October 4, 2024

Abstract

As universities increasingly rely on online platforms for student feedback and evaluations, understanding the mood underlying these reviews has become crucial for improving educational quality and student happiness. The goal of this research is to create a strong model that can reliably identify the attitudes expressed in feedback from students and other sources after analyzing related works in the area of sentiment analysis to academic and non-academic reviews. We combine machine learning and natural language processing techniques, using a variety of platforms like Indeed, Google, and Glass Door to gather evaluations. Term frequency-inverse document frequency (TF-IDF) is one of the sophisticated feature extraction techniques we use in our approach to efficiently capture the subtle differences in emotion expression. For this study, we analyzed sentiments in reviews from Indeed, Glassdoor, and Google using a variety of classification algorithms. Each model's performance was rigorously evaluated, and the one demonstrating the highest accuracy which was Random Forest, selected for predicting sentiments in new reviews.

Keywords: *sentiment, Random Forest, KNN, Datamining, Textmining, tokenization, lemmatization, TF-IDF, SVM, Naive Bayes, Logistic Regression, Decision Tree.*

1 Introduction

With the rise of digital platforms, universities increasingly rely on online reviews from students, faculty, staff and other customers to gain insights into institutional strengths and areas needing improvement. Reviews on platforms such as Google, Glassdoor, and Indeed offer a wealth of information about various aspects of university life, including course quality, faculty performance, campus facilities, and overall student experience. However, manually analyzing these reviews is both time-consuming and subjective, leading to inconsistencies in understanding overall sentiment.

1.1 Sentiment analysis

Sentiment analysis, a subset of natural language processing (NLP) has gained wide spread acceptance in recent years, not just among researchers but also among businesses, governments, and organizations [7]. This has become a popular tool for extracting and categorizing the sentiment (positive, negative, or neutral) expressed in textual data. The growing popularity of the Internet has lifted the web to the rank of the principal source of universal information. Lots of users use various online resources to express their views and opinions. To constantly monitor public opinion and aid decision-making, we must employ user-generated data to analyze it automatically.
We can perform sentiment analysis mostly on public reviews, social media platforms, and similar sites. In this technology-driven world, a majority portion of the data that we come across is unstructured. Whether it is in the form of emails, texts, or documents, the said data need to be properly structured and then analysed further. This is where sentiment analysis comes into play. It not only helps to store data in an efficient and cost-friendly manner, but you can also solve certain real-time issues with the help of the same.

1. Fine-grained sentiment analysis gives precise results to what the public opinion is about the subject. It classified its results in different categories such as: Very Negative, Negative, Neutral, Positive, Very Positive. This analysis gives you an understanding of the feedback you get from customers.

2. Emotion Detection Sentiment Analysis identifies emotions such as anger, happiness, sadness, and others. This is a more sophisticated way of identifying the emotion in a piece of text. Lexicons, ML algorithms are used to recognize emotions.

3. In aspect-based sentiment analysis, you look at the aspect of the thing people are talking about. Suppose you have reviews of a smartphone; you might want to see what the people are talking about its battery life or its screen size.

4. Multilingual, sometimes organizations need to analyse the text of different languages. This form of sentiment analysis is considerably challenging and requires a lot of effort because of the need of many resources.

5. Intent Analysis is a deeper understanding of the intention of the customer. For example, a company can predict if a customer intends to use the product or not. This means that the intention of a particular customer can be tracked, forming a pattern, and then used for marketing and advertising.

Various Approaches Used in Sentiment Analysis Broadly, there are three main approaches to sentiment analysis. They are-

- Rule-based approach- Unlike the other approaches, the rule-based approach is quite easy to comprehend. It basically counts the total number

of negative and positive words present in the data set. Following this, if the result indicates that the number of positive words is more than the number of negative words, then the sentiment is positive, and vice versa.

- Automatic Approach- In this approach, the data set is initially trained, following which predictive analysis is done. After completion of this stage, words are extracted from the text. This can be done with the help of various techniques, some of which might include Linear Regression, Support Vector, and Naive Bayes, among others.
- Hybrid Approach- As the name suggests, this approach is basically an amalgamation of both the rule-based approach and the automatic approach. It delivers more accurate results when compared to the other approaches.

1.2 Datamining

Data mining is the process of using statistical analysis and machine learning to discover hidden patterns, correlations, and anomalies within large datasets. This can be used to evaluate both structured and unstructured data to identify new information[4] and can aid in decision-making, predictive modeling, and understanding complex phenomena[3].

1.2.1 Text Mining

Text mining—also known as text data mining is a sub field of data mining, intended to transform unstructured text into a structured format to identify meaningful patterns and generate novel insights. The unstructured data might include text from sources including social media posts, product reviews, articles, email or rich media formats such as video and audio files. Much of the publicly available data around the world is unstructured, making text mining a valuable practice[3].

1.3 Machine Learning

Machine learning is a subfield of artificial intelligence (AI) that uses algorithms trained on data sets to create self-learning models that are capable of predicting outcomes and classifying information without human intervention[8]

1.3.1 Types of machine learning

1. In supervised machine learning, algorithms are trained on labeled data sets that include tags describing each piece of data. In other words, the algorithms are fed data that includes an “answer key” describing how the data should be interpreted. Supervised machine learning is often used to create machine learning models used for prediction and classification purposes.

2. Unsupervised machine learning uses unlabeled data sets to train algorithms. In this process, the algorithm is fed data that doesn't include tags, which requires it to uncover patterns on its own without any outside guidance. Unsupervised machine learning is often used by researchers and data scientists to identify patterns within large, unlabeled data sets quickly and efficiently.

3. Semi-supervised machine learning uses both unlabeled and labeled data sets to train algorithms. Generally, during semi-supervised machine learning, algorithms are first fed a small amount of labeled data to help direct their development and then fed much larger quantities of unlabeled data to complete the model. Semi-supervised machine learning is often employed to train algorithms for classification and prediction purposes in the event that large volumes of labeled data is unavailable.

4. Reinforcement learning uses trial and error to train algorithms and create models. During the training process, algorithms operate in specific environments and then are provided with feedback following each outcome. Reinforcement learning is often used to create algorithms that must effectively make sequences of decisions or actions to achieve their aims, such as playing a game or summarizing an entire text

1.4 Problem Statement

In today's digital landscape, online reviews significantly influence public perception and decision-making processes related to businesses and workplaces. However, the sheer volume and diversity of sentiment expressed in these reviews from platforms such as Google, Glassdoor, and Indeed present a challenge in effectively extracting and analyzing this information.
The problem in sentiment analysis is classifying the polarity of a given text as the document, sentence, or feature, whether the expressed opinion in a document, a sentence or an entity aspect is positive, negative or neutral [5]. This project aims to address these challenges by developing a robust sentiment analysis model that employs multiple classification techniques to identify and categorize sentiments within these reviews.

1.5 Objectives of Sentiment Analysis

1. Develop a Sentiment Classification Model for the University: Create and train various machine learning models to classify the sentiment (positive, negative, or neutral) of reviews from Google, Glassdoor, and Indeed.

2. Compare Different Classification Techniques: Evaluate the performance of different sentiment analysis models (e.g., SVM, Naive Bayes, Random Forest, and so on)

3. Identification of Best Performing Model: to identify the best-performing machine learning model based on key performance metrics.

4. Freely available, annotated corpus, pre-written classifier codes in Python using NLTK that can be used in NLP to promote research that will lead to a better understanding of how sentiments conveyed in online platforms through texts

1.6 Scope of Sentiment Analysis

1. Gather reviews from various sources where students, faculty, staff and other customers provide feedback about universities, such as Google, Glassdoor, Indeed, and university-specific review platforms.

2. Clean and preprocess the reviews by removing irrelevant content (e.g., special characters, stopwords), handling negated words, and converting text into a machine-readable format by applying feature extraction techniques such as TF-IDF.

3. Apply various machine learning models to classify the sentiment of reviews as positive, negative, or neutral.

4. Visualize the results through charts and graphs, displaying sentiment distribution of university life.

5. The project will focus on text-based reviews and will not consider multimedia content. The scope is also limited to reviews in English.

2 Literature Survey

1. **Sentiment Analysis of Twitter Data[Yili Wang, Jia Xuan Guo, Cheng sheng Yuan and Baozhu,*-2022]**[9]

   This research analysis presents an analysis of Twitter-based sentiment data, particularly focusing on new algorithms and techniques used in Twitter Sentiment Analysis (TSA). The study categorises and reviews various approaches and methodologies applied in TSA, highlighting the growing importance of analysing user-generated data from social media to gauge public opinions and trends. The authors follow a multi-step methodology typical of sentiment analysis research. The first stage is data collection, where the data of tweets are extracted using the Twitter API, focusing on relevant keywords, hashtags, or topics. Then, cleaning the data by removing irrelevant characters and stopwords and performing tokenisation, stemming, and lemmatisation in the data preprocessing stage. Machine learning models or natural language processing (NLP) techniques are applied to classify tweets (positive, negative, or neutral). The model's performance is evaluated using accuracy, precision, recall, and F1 score metrics. The research also provides insights into various classification algorithms used in TSA, such as Support Vector Machines (SVM), Naive Bayes, Random

Forest, and neural network-based models like Long Short-Term Memory (LSTM). The choice of algorithm depends on the dataset and the complexity of the analysis. The study concludes that neural network-based models, especially deep learning approaches like LSTM, provide better accuracy and performance in classifying sentiments compared to traditional machine learning techniques. The paper emphasises the importance of combining sentiment analysis with real-time data to predict trends in domains such as stock markets, product reviews, and political events.

2. **Research progress of sentiment analysis based on Deep learning[Jiameng Li -2024]**

   This paper explore the use of deep learning techniques in sentiment analysis, especially in handling complex natural language data from social media and e-commerce platforms. The study focuses on using large-scale datasets from social media, review platforms, and e-commerce sources. Data is collected using APIs like Twitter's and web scraping techniques to gather user comments and reviews. Text data is cleaned using common NLP techniques like tokenization, stopword removal, and lemmatization. The goal is to prepare the data for efficient feature extraction. The paper highlights the use of word embeddings like Word2Vec and GloVe to capture semantic relationships between words. Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are used to handle sequential dependencies in text data. CNNs are applied to extract features from text, especially for sentence classification tasks. The research explores the combination of CNNs and RNNs to capture both local features and long-term dependencies in the text. The models are trained using labeled sentiment datasets, and techniques like cross-validation are applied to ensure robustness.Finally the result shows hat deep learning models, particularly hybrid architectures combining CNN and LSTM, outperform traditional machine learning approaches (like SVM and Random Forest) in accuracy, precision, and recall for sentiment classification. Despite the improvements, issues such as handling sarcasm, domain-specific language, and real-time sentiment analysis are highlighted as ongoing challenges. This work concludes that while deep learning has advanced sentiment analysis significantly, there is still a need for improvements in efficiency and the ability to handle complex language structures.

3. **Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions[Huang Huang, Adeleh Assemi Tavares, (member, IEEE), and Mumtaz Begum Mustafa, (member, IEEE)-2023]**

   in this paper, the authors discussing the current techniques and future direction for sentiment analysis. The current techniques discuss several key techniques, including:

- Machine Learning Approaches: Traditional methods such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees are used to classify sentiments
- Deep Learning Techniques: More advanced techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, are explored for handling complex sentiment data.
- Lexicon-Based Approaches: This method relies on predefined sentiment lexicons to analyse the polarity of text, often combined with other techniques for enhanced accuracy.

. Challenges in E-commerce Sentiment Analysis:

- Data Volume and Diversity: E-commerce platforms deal with massive amounts of unstructured data, including product reviews in multiple languages, making accurate analysis difficult.
- Sarcasm and Contextual Understanding: Detecting sarcasm, irony, and the nuanced context of reviews remains a significant challenge.

. Future Directions

- Integration of Multimodal Data: Future techniques may involve integrating text with other data types, such as images or videos.
- Real-Time Sentiment Analysis: The need for more robust, real-time systems capable of analysing and responding to customer feedback as it occurs.
- Hybrid Models: Combining machine learning with rule-based or lexicon-based systems for enhanced performance and context-aware analysis is seen as a potential area for growth.

4. **A survey on sentiment analysis methods, applications,and challenges[Mayur Wankhade, Annavarapu Chandra Sekhara Rao, Chaitanya Kulkarni-2022]**

This paper offers a comprehensive overview of sentiment analysis techniques, their practical applications, and the challenges they face. The survey highlights several methods for collecting sentiment data, including web scraping from sources like social media, blogs, e-commerce websites, and forums. Twitter, Facebook, and review platforms are common data sources. These platforms provide real-time data crucial for dynamic sentiment analysis. Data preprocessing techniques are critical for cleaning the raw text data. The authors emphasize methods like tokenization, stopword removal, and lemmatization to prepare data for analysis. Preprocessing ensures that only the most relevant data is fed into sentiment classification models. Effective feature extraction is essential for building robust models. Features can include word embeddings, sentiment lexicons, and linguistic variables. The authors review various feature selection techniques that optimize model performance by eliminating irrelevant or redundant data. Various machine learning and deep learning models are employed to classify data into positive, negative, or neutral sentiments. The authors review algorithms such as Naive Bayes, Support Vector Machines (SVM), and advanced deep learning models like LSTM and CNN. Sentiment classification can occur at different levels: document, sentence, and aspect-based analysis. The survey concludes that deep learning models, especially LSTM and CNN, have shown superior performance in handling complex and large datasets. However, the authors also note that these models require significant computational resources and large amounts of labeled data. Additionally, aspect-level sentiment analysis is gaining traction as it allows for more fine-grained sentiment detection within a sentence or document. One key challenge identified is the need for more domain-specific datasets and improved algorithms capable of understanding context and sarcasm. The paper also highlights future directions, such as improving multilingual sentiment analysis and enhancing models' ability to handle noisy, unstructured data from social media platforms.

5. **A Review Of Sentiment Analysis Methodologies, Practices And Applications[Pooja Mehta, Dr.Sharnil Pandya-2020]**

In this paper, there are mainly focus on the basics of sentiment opinion mining and its levels. There are various approaches and methods to identify sentiment from content. In this paper, their examination represents machine learning procedures. From various classification methods, Sentiment Analysis indicates the results into positive, negative and neutral scores. The study shows that machine learning methods, such as SVM, Naive Bayes, and neural networks, have the highest accuracy and can be

considered as the baseline learning methods as well, as in some cases, lexicon-based methods are very effective. In future work, discovering the result of various other combinations of text data and other prediction accuracy can be done. More work in the future is needed to improve performance measures. The authors conducted an extensive review of existing sentiment analysis methodologies by analysing research papers and case studies in the field. They classified sentiment analysis into different categories, such as machine learning-based, lexicon-based, and hybrid techniques. The review focuses on preprocessing steps, including tokenization, stop-word removal, stemming, and lemmatization, to improve the quality of textual data for analysis. The authors highlight various traditional machine learning algorithms used for sentiment classification, such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest. Deep learning methods such as recurrent neural networks (RNNs) and Convolutional Neural Networks (CNNs) are also discussed for handling more complex textual data. The review explores approaches that rely on predefined sentiment lexicons (e.g., SentiWordNet, VADER) to determine the polarity of text and how these are often integrated with machine learning models for more accurate predictions. Also, this paper discusses how sentiment analysis is applied in various domains, such as social media analysis, customer feedback for e-commerce, and product review systems. They provide use cases that show the impact of sentiment analysis on business decision-making. The author suggest future research should focus on improving contextual sentiment understanding and building more efficient hybrid models that combine the strengths of machine learning and lexicon-based approaches. Hey also recommends that more work is needed in the area of multimodal sentiment analysis, integrating text with other forms of data, such as images and videos.

3 Methodology of Proposed System

3.1 Data Collection

3.1.1 Data Source

The data for this project is sourced from reviews on platforms such as Google Reviews, Glassdoor, and Indeed, where users share feedback on various aspects of their university experience. These reviews include a star rating system, which provides an indication of sentiment. To create a labeled dataset for sentiment analysis, the sentiment of each review was manually labeled based on the star ratings provided: higher star ratings (e.g., 4-5 stars) were classified as positive, lower star ratings (e.g., 1-2 stars) as negative, and middle ratings (e.g., 3 stars) as neutral. This labeling process ensures that the sentiment classification model has accurate, reliable training data.

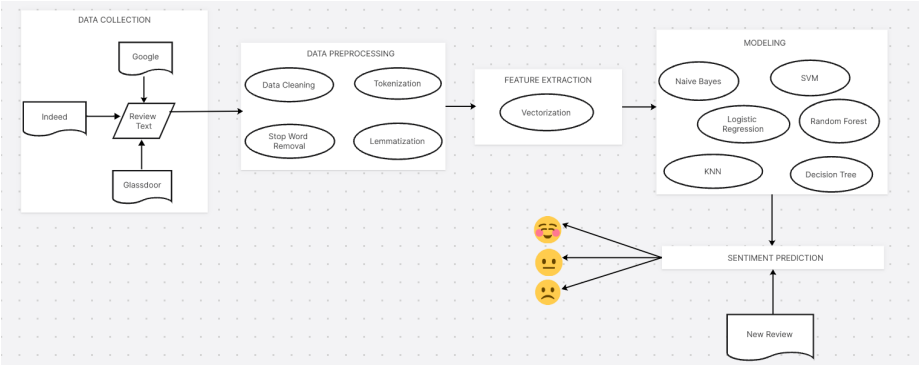

Figure 1: Flow Diagram

Figure 2: Review Collection

3.1.2 Dataset Description

The dataset used for this sentiment analysis consists of 206 reviews sourced from platforms such as Google, Glassdoor, and Indeed. Each review is manually labeled to reflect its sentiment, with three possible labels:

- 1 for positive sentiment,
- 0 for neutral sentiment,
- -1 for negative sentiment.

The dataset includes two main attributes:

1. Review Text: The content of the review, which describes user experiences and opinions.
2. Sentiment Label: The corresponding sentiment label (1, 0, or -1) based on the star rating or manual classification.

```
data = pd.read_csv('University_review.csv')#Load dataset of University review
print(data.head())# Show first few rows
```

```
   RATING                                             REVIEW
0       1  Ajman University is highly regarded for its di...
1       1  as soon as i stepped through the university do...
2       1  Ajman university is a good place to study, pro...
3       1  A wonderful university. I'm a new student and ...
4       1  I am so glad to be accepted at my dream univer...
```

Figure 3: Part of Dataset

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 206 entries, 0 to 205
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   RATING  206 non-null    int64 
 1   REVIEW  206 non-null    object
dtypes: int64(1), object(1)
memory usage: 3.3+ KB
```

```
print(data.shape)
```

```
(206, 2)
```

Figure 4: Description of Dataset

```
print('Number of positive, neutral and negative reviews: ', data.RATING.value_
```

```
Number of positive, neutral and negative reviews:  RATING
 1    112
 0     49
-1     45
Name: count, dtype: int64
```

```
print('Percentage of positive, neutral and negative reviews: ', data.RATING.va
```

```
Percentage of positive, neutral and negative reviews:  RATING
 1    54.368932
 0    23.786408
-1    21.844660
Name: count, dtype: float64
```

Figure 5: Count and proportion of Positive, Neutral and Negative Reviews

3.2 Data Preprocessing

Data quality significantly influences the performance of a machine-learning model. Inadequate or low-quality data can lead to lower accuracy and effectiveness of the model.
In general, text data derived from natural language is unstructured and noisy. So text preprocessing is a critical step to transform messy, unstructured text data into a form that can be effectively used to train machine learning models, leading to better results and insights. Data preprocessing typically involves the following steps[2]:

- Lowercasing

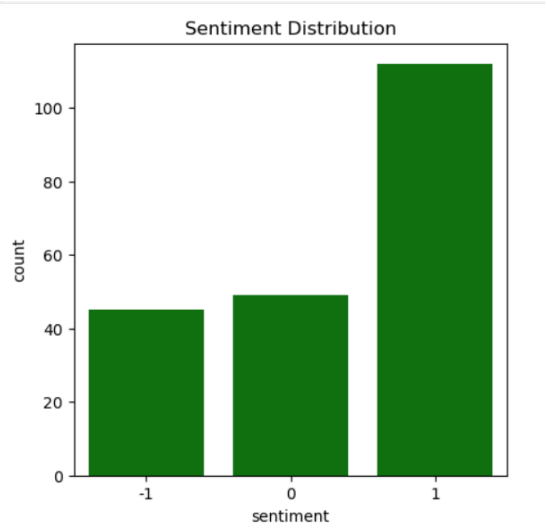

Figure 6: Sentiment Distribution

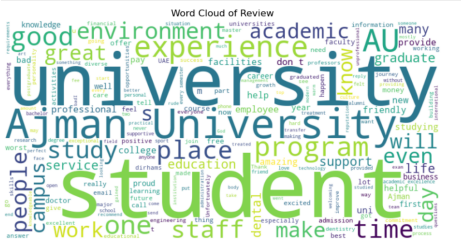

Figure 7: Review Wordcloud

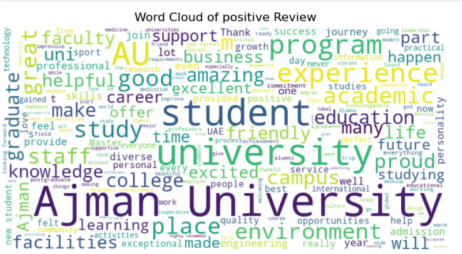

Figure 8: Positive Review Wordcloud Visualization

Figure 9: Neutral Review Wordcloud Visualization

Figure 10: Negative Review Wordcloud Visualization

- Removing Punctuation, Special Characters and Digits
- Handling Negations
- Tokenization
- Stop-Words Removal
- Stemming & Lemmatization

3.2.1 Lowercasing

Lowercasing is a text preprocessing step where all letters in the text are converted to lowercase. This step is implemented so that the algorithm does not treat the same words differently in different situations.

3.2.2 Removing Punctuation, Special Characters and Digits

Punctuation removal is a text preprocessing step where we remove all punctuation marks (such as periods, commas, exclamation marks, emojis etc.), special characters and digits from the text to simplify it and focus on the words themselves. Because they are irrelevant for sentiment analysis.

```
print(data['lower_review'])
```

```
0      ajman university is highly regarded for its di...
1      as soon as i stepped through the university do...
2      ajman university is a good place to study, pro...
3      a wonderful university. i'm a new student and ...
4      i am so glad to be accepted at my dream univer...
                             ...
201    a very bad way to contact students none of the...
202    if you wanted to have a free dental check up a...
203          great place to learn and aquire new skills.
204    thank you for your information reply for my in...
205                         amazing service, lovely people
Name: lower_review, Length: 206, dtype: object
```

Figure 11: Lowercasing Output

```
0      ajman university is highly regarded for its di...
1      as soon as i stepped through the university do...
2      ajman university is a good place to study prov...
3      a wonderful university im a new student and i ...
4      i am so glad to be accepted at my dream univer...
                             ...
201    a very bad way to contact students none of the...
202    if you wanted to have a free dental check up a...
203          great place to learn and aquire new skills
204    thank you for your information reply for my in...
205                          amazing service lovely people
Name: no_special_review, Length: 206, dtype: object
```

Figure 12: Removal of Special Charcters, Punctuation and Digits

3.2.3 Handling Negations

Negations can completely change the sentiment of a sentence, for example, "not good", is negative even though "good" is positive. One way to handle this is by concatenating the negation with the following word.

```
0      ajman university is highly regarded for its di...
1      as soon as i stepped through the university do...
2      ajman university is a good place to study prov...
3      a wonderful university im a new student and i ...
4      i am so glad to be accepted at my dream univer...
                             ...
201    a very bad way to contact students none of the...
202    if you wanted to have a free dental check up a...
203          great place to learn and aquire new skills
204    thank you for your information reply for my in...
205                          amazing service lovely people
Name: negated_review, Length: 206, dtype: object
my experience in this university was life changing maybe not_only because of the university life but mostly the exchange group
```

Figure 13: Concatenating Negated words

3.2.4 Tokenization

Tokenization is refers to the process of converting a sequence of text into smaller parts, known as tokens. These tokens can be as small as characters or as long as words. The primary reason this process matters is that it helps machines understand human language by breaking it down into bite-sized pieces, which are easier to analyze.

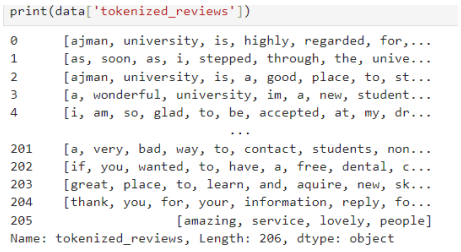

Figure 14: Tokenization of Reviews

3.2.5 Stop-Words Removal

Stopwords (”the”, ”is”, ”in”, etc,.) are words that don’t contribute to the meaning of a sentence. So they can be removed without causing any change in the meaning of the sentence. The NLTK library has a set of stopwords and we can use these to remove stopwords from our text and return a list of word tokens. Removing these can help focus on the important words.

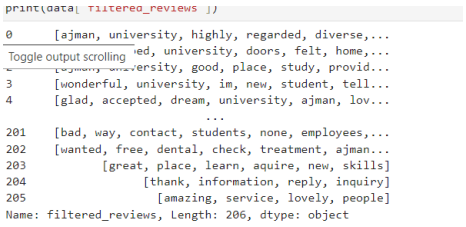

Figure 15: Removing Stop-words

3.2.6 Stemming and Lemmatization

Stemming is a process to reduce the word to its root stem for example run, running, runs, runed derived from the same word as run. basically stemming do is remove the prefix or suffix from word like ing, s, es, etc. NLTK library is used to stem the words. The stemming technique is not used for production purposes because it is not so efficient technique and most of the time it stems the unwanted words. So, to solve the problem another technique came into the market as Lemmatization.
Lemmatization is similar to stemming, used to stem the words into root word but differs in working. Actually, Lemmatization is a systematic way to reduce the words into their lemma by matching them with a language dictionary.

```
print(data['lemmatized_reviews'])

0      [ajman, university, highly, regarded, diverse,...
1      [soon, stepped, university, door, felt, home, ...
2      [ajman, university, good, place, study, provid...
3      [wonderful, university, im, new, student, tell...
4      [glad, accepted, dream, university, ajman, lov...
                             ...
201    [bad, way, contact, student, none, employee, t...
202    [wanted, free, dental, check, treatment, ajman...
203            [great, place, learn, aquire, new, skill]
204                 [thank, information, reply, inquiry]
205                   [amazing, service, lovely, people]
Name: lemmatized_reviews, Length: 206, dtype: object
```

Figure 16: Lemmatized Reviews

3.3 Feature Extraction

A crucial part of sentiment classification is feature extraction because it involves extracting valuable information from text data, which affects the model's performance. Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set[1]. It yields better results than applying machine learning directly to the raw data.

3.3.1 Text Vectorization using TF-IDF

Vectorization in NLP is used to convert raw text data into a numerical format that machine learning algorithms can understand and process.

$$TFIDF(term) = TF(term) * IDF(term)$$

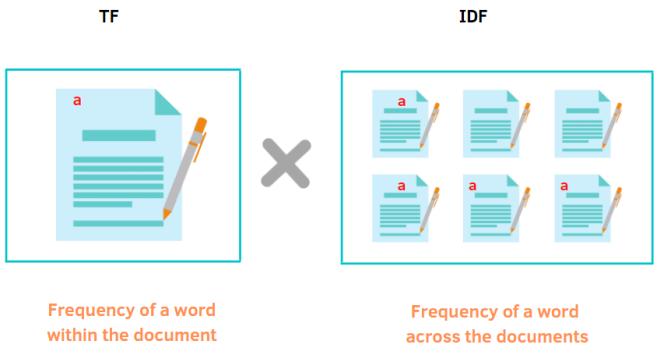

Figure 17: TF-IDF (google source)

1. **Term Frequency**

   In document d, the frequency represents the number of instances of a given word t. Therefore, we can see that it becomes more relevant when a word appears in the text, which is rational. Since the ordering of terms is not significant, we can use a vector to describe the text in the bag of term models. For each specific term in the dataset, there is an entry with the value being the term frequency. Or, It is the percentage of the number of times a word (t) occurs in a particular document (d) divided by the total number of words in that document.
   $TF(term) = \frac{Number\ of\ times\ term\ appears\ in\ document\ d}{Total\ number\ of\ items\ in\ the\ document}$

2. **Document Frequency**

   This tests the meaning of the text, which is very similar to TF, in the whole corpus collection. The only difference is that in document d, TF is the frequency counter for a term t, while df is the number of occurrences in the document set N of the term t. In other words, the number of papers in which the word is present is DF. $DF(term) = Occurences\ of\ term\ in\ documents$

3. **Inverse Document Frequency**

   It measures the importance of the word in the corpus. It measures how common a particular word is across all the documents in the corpus. It is the logarithmic ratio of no. of total documents to no. of a document with a particular word. $IDF(term) = log(\frac{Total\ number\ of\ documents}{Number\ documents\ with\ term\ in\ it})$

```
print(X_tf)
```

	abdullah	ability	able	abroad	absent	academic	aca
0	0.0	0.0	0.0	0.0	0.0	0.21252	
1	0.0	0.0	0.0	0.0	0.0	0.00000	
2	0.0	0.0	0.0	0.0	0.0	0.00000	
3	0.0	0.0	0.0	0.0	0.0	0.00000	
4	0.0	0.0	0.0	0.0	0.0	0.00000	
..	...	...	...	...	...	...	
201	0.0	0.0	0.0	0.0	0.0	0.00000	
202	0.0	0.0	0.0	0.0	0.0	0.00000	
203	0.0	0.0	0.0	0.0	0.0	0.00000	
204	0.0	0.0	0.0	0.0	0.0	0.00000	
205	0.0	0.0	0.0	0.0	0.0	0.00000	

	accelerate	acceptance	accepted	...	worse	worst	w
0	0.0	0.0	0.000000	...	0.0	0.0	
1	0.0	0.0	0.000000	...	0.0	0.0	
2	0.0	0.0	0.000000	...	0.0	0.0	
3	0.0	0.0	0.000000	...	0.0	0.0	
4	0.0	0.0	0.365939	...	0.0	0.0	

Figure 18: TFIDF Vectorization Output

3.4 Model Selection

In this project, for model selection, we employed various machine learning algorithms[4][6] to classify the sentiment of reviews. The following models were used for comparison:

1. Naive Bayes
2. Support Vector Machine (SVM)
3. Logistic Regression
4. Random Forest
5. K-Nearest Neighbors (KNN)
6. Decision Tree

3.4.1 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, which assumes independence between the features (words in the case of text classification). It is particularly well-suited for text classification problems due to its simplicity and efficiency.

3.4.2 Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm that finds the optimal hyperplane separating different classes in a high-dimensional space. For sentiment analysis, it transforms text data into a numerical vector and separates positive, neutral, and negative sentiments based on this data.

3.4.3 Logistic Regression

Logistic Regression is a linear model that predicts the probability of an instance belonging to a particular class (positive, neutral, or negative). It works by modeling the relationship between the features and the probability of a given label.

3.4.4 Random Forest

Random Forest is an ensemble model that builds multiple decision trees during training and averages their results to improve accuracy. Each tree is trained on a subset of the data, helping to reduce overfitting.

3.4.5 K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric algorithm that classifies a new data point based on its distance to the k-nearest points in the training dataset. In sentiment analysis, the distance is measured between feature vectors representing the review text.

3.4.6 Decision Tree

Decision Tree models split the dataset based on feature conditions to create a tree where each branch represents a decision rule and each leaf represents an output label. For sentiment analysis, this means splitting the data based on the occurrence of specific words or phrases.

3.5 Model Training and Evaluation

For this sentiment analysis project, we used multiple machine learning classifiers to predict the sentiment of reviews.

3.5.1 Splitting the Dataset

The dataset of 206 reviews was split into training and testing sets (typically 80 % for training and 20% for testing). This ensures that the model is trained on one portion of the data and tested on another portion to measure its performance on unseen reviews.

3.5.2 Model Training

For each classifier (Naive Bayes, SVM, Logistic Regression, Random Forest, KNN, and Decision Tree), the training set was used to teach the model how to classify reviews into positive, neutral, or negative categories. The training process involves feeding the features (e.g., review text, sentiment rating) into the model and adjusting parameters to minimize classification errors.

3.5.3 Evaluation Metrics Used

After training the models, we evaluated their performance using several key metrics to compare and select the best model for sentiment classification:

1. Accuracy: measures the proportion of correctly classified reviews (positive, neutral, or negative) out of the total number of reviews.

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

2. Precision: measures how many reviews classified as a particular sentiment (positive, neutral, negative) are actually correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. The F1-score: is the harmonic mean of precision and recall, offering a single metric that balances both.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Recall: measures how many actual reviews of a particular sentiment were correctly identified by the model.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

5. Confusion matrix: provides a summary of correct and incorrect predictions by showing the counts of true positives, false positives, false negatives, and true negatives for each sentiment class. It gives a more detailed picture of where the model is making mistakes.

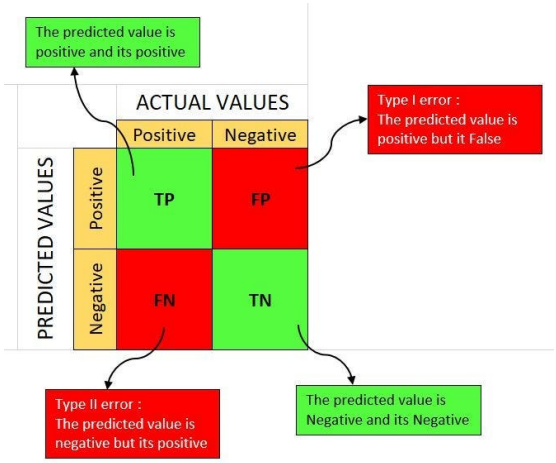

Figure 19: Sample Conusion Matrix

4 Results and Discussion

In this section, we present the performance of the models evaluated in this sentiment analysis project. The results include accuracy, precision, recall, F1-score, and confusion matrices. These metrics allow us to assess and compare how well each model classifies sentiment into positive, neutral, and negative categories. Below are the key results for each classifier.

4.1 Model Performance

Since Random Forest showed the best performance, we provide its confusion matrix and classification report to better understand the classification results.

- The Random Forest classifier outperformed the other models with an accuracy of 76.19%, and strong precision, recall, and F1-scores across all sentiment categories.
- Logistic Regression also performed well, achieving 73.81% accuracy, proving it to be a good baseline model for text classification tasks like sentiment analysis.

Model	Accuracy	Class label	Precision	Recall	F1 score
Naïve Bayes	57.14%	Negative(-1)	0.00	0.00	0.00
		0(Neutral)	0.00	0.00	0.00
		Positive(1)	0.57	1.00	0.73
SVM	73.80%	Negative(-1)	0.67	0.75	71.00
		0(Neutral)	0.71	0.50	0.59
		Positive(1)	0.77	0.83	0.80
Logistic Regression	73.81%	Negative(-1)	0.67	0.75	0.71
		0(Neutral)	0.71	0.50	0.59
		Positive(1)	0.77	0.83	0.80
Random Forest	76.19%	Negative(-1)	0.88	0.88	0.88
		0(Neutral)	0.67	0.40	0.50
		Positive(1)	0.75	0.88	0.81
KNN	64.29%	Negative(-1)	0.50	0.50	0.50
		0(Neutral)	0.56	0.50	0.53
		Positive(1)	0.72	0.75	0.73
Decision Tree	69.05%	Negative(-1)	0.71	0.62	0.67
		0(Neutral)	0.42	0.50	0.45
		Positive(1)	0.83	0.79	0.81

Accuracy: 76.19%

	precision	recall	f1-score	support
-1	0.88	0.88	0.88	8
0	0.67	0.40	0.50	10
1	0.75	0.88	0.81	24
accuracy			0.76	42
macro avg	0.76	0.72	0.73	42
weighted avg	0.75	0.76	0.75	42

Figure 20: Classification Report of Random Forest

- The Support Vector Machine (SVM) provided competitive performance with 73.80% accuracy, making it a reliable model for classifying sentiment with high dimensionality and sparse features such as text.
- Naive Bayes showed lower performance, only 57.14% compared to the other models, likely due to their simplicity and reliance on assumptions that they don't always hold in complex, imbalanced text data.

4.2 Implications of Finding

The findings of this sentiment analysis project have several important implications for understanding user opinions and improving services based on review feedback across platforms such as Google, Glassdoor, and Indeed. The high accuracy and balanced performance of models like Random Forest and Logistic Regression in classifying sentiment provide valuable insights into user satisfac-

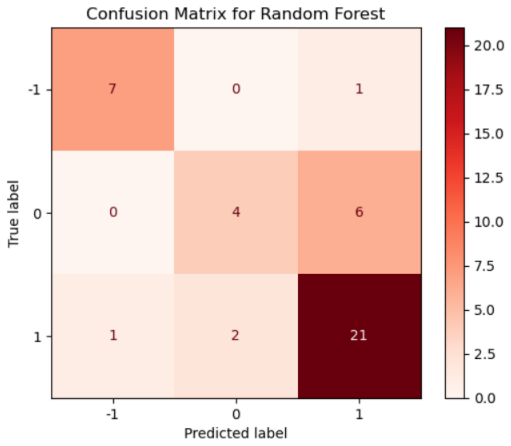

Figure 21: Confusion Matrix of Random Forest

tion, complaints, and neutral feedback.

- Improving University Services: The ability to automatically categorize reviews as positive, neutral, or negative allows the university to take targeted action based on the specific aspects users are discussing.
- Application in Strategic Decision-Making: The insights gained from sentiment analysis can be used for strategic decision-making at the university level. For instance, analyzing reviews over time can help identify trends in student satisfaction or dissatisfaction.
- Resource Allocation and Priority Setting: The findings help in identifying which areas require immediate attention versus those that are functioning well.
- Enhancing Student and Staff Engagement: Understanding the sentiment behind feedback from students, faculty, and staff allows the university to tailor communication and services more effectively. Positive reviews can be celebrated and used for promotional purposes, while negative feedback can help inform support services, academic advising, or campus outreach efforts to ensure better engagement.

4.3 Prediction on New Reviews

Once the best-performing model (in our case, Random Forest) has been selected and trained, it can be used to predict the sentiment of new reviews. For that we need to follow all the steps mentioned in data preprocessing and feature extraction.

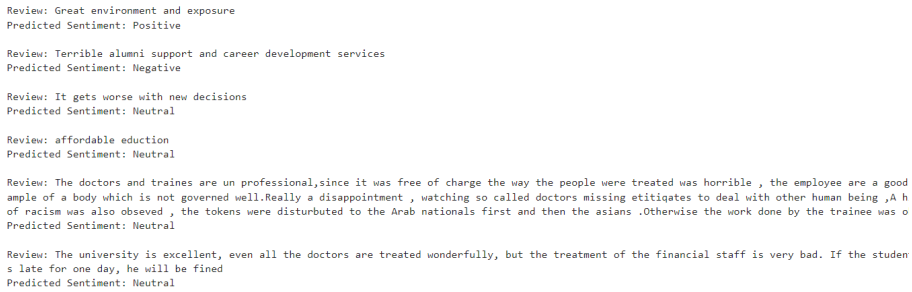

Figure 22: Sample Prediction on New Review

5 Conclusion and Future Works

This project successfully implemented sentiment analysis on university-related reviews collected from platforms such as Google, Glassdoor, and Indeed. The Random Forest model emerged as the best-performing classifier with an accuracy of 76.19%, providing reliable and balanced sentiment predictions across all categories.

The ability to automatically analyze sentiment in reviews has significant implications for improving university services. The results can guide strategic decision-making, allowing the university to prioritize areas requiring immediate attention while celebrating aspects that are positively received by students, staff, and external reviewers.

For future directions, the system can be further improved. Currently, the model focuses on reviews in English. As universities attract diverse students and faculty, reviews in multiple languages are common. Future work could involve multilingual sentiment analysis, assess the intensity of the sentiments, incorporating emotion detection. Sentiment analysis often encounters imbalanced datasets, where certain sentiment categories (e.g., positive reviews) dominate. Future work could explore advanced techniques for handling imbalanced data. In this project, the model provides an overall sentiment classification for each review. In the future, implementing aspect-based sentiment analysis (ABSA) could enable more granular sentiment analysis, identifying specific aspects (e.g., faculty, facilities, administration) of the university that the sentiment pertains to. This would help provide targeted insights on specific areas of concern or praise.This would enable more focused insights on particular areas of concern or commendation

References

[1] Arie Satia Dharma and Yosua Giat Raja Saragih. Comparison of feature extraction methods on sentiment analysis in hotel reviews. *Sinkron: jurnal dan penelitian teknik informatika*, 6(4):2349–2354, 2022.

[2] Huu-Thanh Duong and Tram-Anh Nguyen-Thi. A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, 8(1):1, 2021.

[3] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[4] Monika Kabir, Mir Md Jahangir Kabir, Shuxiang Xu, and Bodrunnessa Badhon. An empirical research on sentiment analysis using machine learning approaches. *International Journal of Computers and Applications*, 43(10):1011–1019, 2021.

[5] Pooja Mehta and Sharnil Pandya. A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*, 9(2):601–609, 2020.

[6] G Revathy, Saleh A Alghamdi, Sultan M Alahmari, Saud R Yonbawi, Anil Kumar, and Mohd Anul Haq. Sentiment analysis using machine learning: Progress in the machine intelligence for data science. *Sustainable Energy Technologies and Assessments*, 53:102557, 2022.

[7] J Fernando Sánchez-Rada, Oscar Araque, and Carlos A Iglesias. Senpy: A framework for semantic sentiment and emotion analysis services. *Knowledge-Based Systems*, 190:105193, 2020.

[8] Harry Surden. Machine learning and law: An overview. *Research Handbook on Big Data Law*, pages 171–184, 2021.

[9] Yili Wang, Jiaxuan Guo, Chengsheng Yuan, and Baozhu Li. Sentiment analysis of twitter data. *Applied Sciences*, 12(22):11775, 2022.