National College of Ireland

SENTIMENT ANALYSIS OVER YELP DATASET

MSc Research Project
Data Analytics

Akhil Bharat Sisal
Student ID: x21214638

School of Computing
National College of Ireland

Supervisor: Dr.Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Akhil Bharat Sisal
Student ID:	x21214638
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	14/12/2023
Project Title:	Sentiment Analysis Over Yelp Dataset
Word Count:	6190
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Akhil Bharat Sisal
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis Over Yelp Dataset

Akhil Bharat Sisal
21214638

Abstract

This study investigate the Sentiment Analysis for a Business Recommendation System for businesses that use Sentiment Analysis to make the buying experience better for customers. The system uses a set of around 100,000 Yelp reviews to look at customer ratings, comments, and reviews in order to make personalised suggestions. By using information from customer data, this method goes beyond traditional recommendation algorithms. The study uses advanced feature extraction methods like TF-IDF vectorization and BERT tokenization, along with machine learning techniques like DBScan, K-means, and K-Nearest Neighbor (KNN), to rate these models on their accuracy, precision, recall, and F1-score. The results show that the KNN model works best, especially when BERT tokenization is used. The model developed shows great potential. The project's successfully opens the door for more research into personalised retail solutions that meet customers' changing needs.

1 Introduction

In today's digital age, retail businesses are increasingly seeking innovative ways to enhance customer experience and boost sales. One of the most promising avenues is the use of advanced data analytics to understand customer preferences and behaviours. This project aims to develop a cutting-edge Recommendation System for the retail sector, leveraging the power of Sentiment Analysis.

The core idea is to analyze customer reviews, ratings, and feedback across various platforms to gauge the sentiments and preferences of consumers. By interpreting this data, the system will not only understand what customers are saying about the business but also how they feel about them. This emotional insight is crucial in tailoring the recommendations that are not just based on traditional metrics like purchase history or browsing patterns, but also on the emotional responses of customers.

The motivation behind this project stems from the evolving landscape of the retail industry, where customer expectations are continually rising, and personalized experiences are becoming the norm. Traditional recommendation systems, while effective, often overlook the nuanced preferences and emotional responses of customers. In an era where consumer feedback is abundant and increasingly accessible, tapping into this resource to enhance recommendation systems is not just innovative but necessary. By integrating sentiment analysis into these systems, we can address the gap between what customers say and what they actually feel, leading to more accurate and satisfying shopping experiences. The significance of this project lies in its potential to revolutionize how retail

businesses interact with and understand their customers. By leveraging sentiment analysis, retailers can gain deeper insights into customer emotions and preferences, allowing for more nuanced and effective recommendations. This approach is particularly relevant in an increasingly competitive market, where customer loyalty and satisfaction are paramount. Furthermore, the insights gained from this project can help retailers in inventory management, marketing strategies, and overall business decision-making, making it a comprehensive tool for retail business enhancement.

The primary aim of this project is to develop a machine learning model supporting a sophisticated recommendation system for the retail sector, which utilizes sentiment analysis to provide personalized and emotionally resonant suggestions to customers. This system aims to not only improve the accuracy of the recommendations but also to enhance the overall customer buying experience. By achieving this, the project seeks to contribute to the broader field of retail analytics and customer relationship management, demonstrating the practical applications and benefits of combining sentiment analysis with recommendation systems in a real-world retail context.

This project's contribution to the field of retail business and data analytics is multifaceted. Firstly, it bridges the gap between quantitative data and qualitative insights, providing a more holistic view of consumer behavior. By integrating sentiment analysis into the recommendations, the system offers a unique approach that goes beyond traditional algorithms. The proposed model not only enhances the accuracy of recommendations but also adds a layer of emotional intelligence to the process. Secondly, the project contributes to the growing body of knowledge in sentiment analysis and its practical applications, showcasing its effectiveness in a retail context. Finally, by implementing this system, retailers can benefit from increased customer engagement and satisfaction, potentially leading to higher sales and customer loyalty. The focus of the study is to develop machine learning model supporting recommendations.

1.1 Research Question

To what extend does non-supervised machine learning algorithms of DBScan and Kmeans compare against the supervised algorithm of K-Nearest Neighbors (KNN) in providing retail recommendations based on sentiments from the Yelp Restaurant Dataset?

In many ways, this project adds to the fields of retail business and data analytics. In the first place, it connects quantitative data with qualitative insights, giving a full picture of how people act. The system takes a different approach than usual algorithms because it uses sentiment analysis to make recommendation. This model not only makes suggestions more accurate, but it also adds an element of emotional intelligence to the process. Second, the project adds to what is known about sentiment analysis and how it can be used in real life by showing how well it works in a retail setting. Lastly, retailers can get more engaged and happy customers by using this system, which could lead to more sales and customer loyalty.

The rest of the report is structured in 5 chapters. The subsequent chapter of Related Work talks about the state-of-the-art of the system followed by the Research Methodology chapter discussing the methods incorporated in the study based on the review conducted. The architecture of the implemented system is discussed in chapter Design Specifications that is followed by the Implementation chapter. The results of the implementation are then discussed in the Evaluation chapter, whereas the study is concluded with the chapter, Conclusion and Future Scope.

2 Related Work

In this section, several studies related to the application of machine learning algorithms in the field of Recommendation system have been analysed in order to find out the potential gap in the literature.

Sun et al. (2019) developed a semi-supervised fuzzy product ontology mining algorithm to perform fine-grained sentiment analysis on electronic word-of-mouth (eWOM) in online customer reviews. This method uses social media data to improve market intelligence and aid in strategy development, showing significant performance improvements in eWOM analysis.

Mohamed et al. (2019)presented a comprehensive survey on recommender systems, exploring different techniques such as content-based, collaborative, demographic, and hybrid filtering. They discussed important challenges, such as the cold start problem, and emphasized the importance of hybrid approaches in enhancing system effectiveness and efficiency.

Ahmed et al. (2022) discussed the application of AI in sentiment analysis for business performance enhancement. They explored machine learning and deep learning applications in various business sectors and showed that AI-driven sentiment analysis can significantly boost business processes.

Osman et al. (2019) developed a novel recommender system for electronic products that integrates contextual sentiment analysis with traditional recommender systems. Their model, which incorporates user comments and preferences, outperformed traditional collaborative filtering approaches in terms of RMSE and MAE measurements, indicating its effectiveness in the electronic product domain.

Abdulsalam et al. (2023) focused on developing methods to detect suicidal thoughts in Arabic tweets. Their study involved creating a novel dataset and examining several machine learning models, including pre-trained deep learning models like AraBert, AraELECTRA, and AraGPT2. The results showed high accuracy and F1 scores, with deep learning models, particularly AraBert, outperforming others in detecting suicidal ideation.

Loukili et al. (2023) proposed an advanced recommender system for e-commerce platforms using the Frequent-Pattern-Growth algorithm. Their system provides personalized recommendations based on customer purchase history and shows a high probability of customers purchasing the next suggested product, demonstrating its potential in improving e-commerce strategies.

Mehta et al. (2021) conducted a comprehensive survey of hybrid recommendation systems, discussing the integration of various recommendation techniques. They highlighted the advantages of hybrid systems in overcoming challenges like the cold-start problem and enhancing the accuracy and relevance of recommendations.

In their paper, Almahmood and Tekerek (2022) conducted an extensive review of deep learning techniques used in e-commerce recommendation systems. They analyzed various deep learning approaches, such as CNNs, RNNs, and sentiment analysis, highlighting their applications in addressing challenges such as the cold start problem and data sparsity.

Similarly, Birjali et al. (2021) provided an in-depth review of sentiment analysis methodologies and applications, covering machine learning, lexicon-based approaches, and hybrid models. They discussed the challenges and trends in the field and emphasized the significance of sentiment analysis in various domains.

Moreover, Sheng et al. (2021) delved into the role of big data analytics during the

COVID-19 pandemic. They provided a comprehensive review of various analytics methods used in the management field, including descriptive, predictive, and prescriptive analytics, and discussed their impact on decision-making for businesses and governments.

In one study, Al-Abbadi et al. (2022) found that positive online reviews significantly increase the likelihood of purchase. Another study by Bag et al. (2022) revealed that AI technologies positively influence user engagement, leading to satisfying user experiences and increased repurchase intentions. Kauffmann et al. (2020) developed the Fake Review Detection Framework (FRDF) for sentiment analysis and fake review detection, which was tested on Amazon reviews for improved brand management and consumer decision-making. Additionally, Balush et al. (2021) discussed the development of a recommendation system that integrates intelligent search, NLP, and machine learning methods, enhancing e-commerce functionalities and user experience.

Hutto and Gilbert (2014) built a VADER, a simple model which was rule-based for broad sentiment analysis. The model's effectiveness was then compared to 11 commonly used as benchmarks, like Affective Norms for English Words (ANEW). The Linguistic Inquiry and Word Count (LIWC) is a tool used for analysing language, specifically focusing on word usage and linguistic patterns. The tools utilised are Senti WordNet, Inquirer, and machine learning focused techniques. These methods depend on the Maximum Entropy and Naive Bayes algorithms, and techniques based on Support Vector Machines (SVM). This research Outlined the process of creating, verifying, and assessing the character known as "VADER". Researchers employed a blend of qualitative employing both qualitative and quantitative methodologies to generate and authenticate a Lexicon of sentiments utilised in the realm of social media. VADER is employing an economical rule-based methodology to analyse the sentiment expressed in tweets. The research demonstrated that VADER enhanced the advantages of conventional sentiment analysis. Lexicons, such as LIWC, were distinguished from VADER.

Dwivedi et al. (2023) conducted a study on the potential impact of the metaverse on marketing. The research paper, which included contributions from multiple experts, explored how the metaverse can revolutionize consumer engagement, brand interaction, and marketing strategies. The study also identified the challenges and opportunities that come with virtual environments.

Dang et al. (2021) proposed an adaptive recommender system that integrates sentiment analysis with collaborative filtering techniques. By using LSTM networks and CNN, they analyzed user reviews and combined sentiment analysis with user-based collaborative filtering for rating prediction. The system demonstrated significant improvements in performance, with high accuracy and F-score, and AUC values above 84% on Amazon Fine Food and Movie Reviews datasets.

Huang et al. (2023) provided a comprehensive review of sentiment analysis techniques in e-commerce platforms. The focus was on machine learning and deep learning methodologies, identifying trends towards deep learning methods, and suggesting future research directions such as universal language models and aspect-based sentiment analysis. Their systematic examination of various approaches offers crucial insights for the field. Dadhich and Thankachan (2022) conducted a study on the sentiment analysis of Amazon product reviews using a hybrid rule-based approach. They compared various sentiment analysis algorithms, evaluating their performance in accuracy, precision, recall, and F1-score. This study contributes significantly to understanding the efficiency and challenges of current sentiment analysis techniques.

Abbasi and Khadivar (2021) improved recommender systems by incorporating senti-

ment analysis of Amazon book reviews. They employed various machine learning classifiers and an ensemble method, leading to improved precision and performance in sentiment analysis. This integration significantly enhances the accuracy and relevance of product recommendations.

Stalidis et al. (2023) conducted a review of recommendation system techniques in e-commerce, with a focus on sustainable marketing strategies. The study examined collaborative filtering, content-based, and hybrid techniques, addressing challenges, and highlighting the role of RS in sustainable e-shopping practices.

Sadhasivam and Kalivaradhan (2019) introduced an ensemble machine learning approach for sentiment analysis of Amazon product reviews. The method combined Naive Bayes and SVM algorithms, enhancing the accuracy of sentiment classification. This approach effectively addresses the complexity of sentiment analysis in large datasets.

Yi and Liu (2020) investigated customer sentiment analysis for shopping recommendations using machine learning. They employed multiclass support vector machines for analyzing customer reviews on social media, demonstrating high accuracy rates in their sentiment analysis model. This approach enhances the shopping experience through personalized recommendations.

3 Methodology

This chapter provides a detailed account of the research methodology, including the research design, data collection methods, and analysis techniques used to ensure the study's validity and strength. This exploratory study of developing a recommendation system using unsupervised machine learning and supervised machine learning modalities is depicted in Figure 1 below.

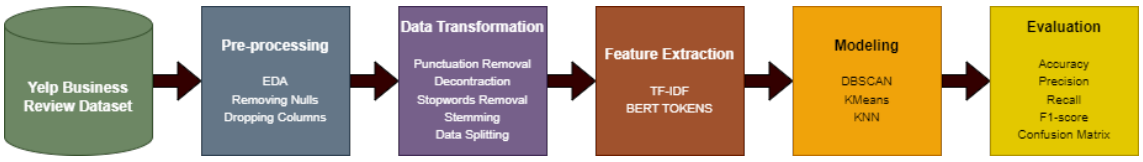

Figure 1: Methodology flow of the study

3.1 Data Collection

This dataset is a rich and detailed collection of 229906 Yelp reviews, capturing diverse aspects of customer experiences and opinions and is available at Data.world[1]. Figure 2 below depicts the contents of the dataset. Each record in the dataset includes several key fields:

Business Information: This includes the name of the business, its categories (for example, "Breakfast and Brunch; Restaurants"), its location (city, full address, latitude, and longitude), and information that is unique to the business, like its ID, opening status, review count, star ratings, and type.

Review Content: Each entry has a unique review ID, the date of the review, and the review itself. This textual data is very important for sentiment analysis because it gives us direct information about what customers think and feel.

[1] `https://data.world/brianray/yelp-reviews`

	Unnamed: 0	business_blank	business_categories	business_city	business_full_address	business_id	business_latitude	business_longitude
0	0	False	Breakfast & Brunch; Restaurants	Phoenix	6106 S 32nd St\nPhoenix, AZ 85042	9yKzy9PApeiPPOUJEtnvkg	33.390792	-112.012504
1	1	False	Italian; Pizza; Restaurants	Phoenix	4848 E Chandler Blvd\nPhoenix, AZ 85044	ZRJwVLyzEJq1VAihDhYiow	33.305607	-111.978758
2	2	False	Middle Eastern; Restaurants	Tempe	1513 E Apache Blvd\nTempe, AZ 85281	6oRAC4uyJCsJl1X0WZpVSA	33.414345	-111.913031
3	3	False	Active Life; Dog Parks; Parks	Scottsdale	5401 N Hayden Rd\nScottsdale, AZ 85250	_1QQZuf4zZOyFCvXc0o6Vg	33.522945	-111.907886
4	4	False	Tires; Automotive	Mesa	1357 S Power Road\nMesa, AZ 85206	6ozycU1RpktNG2-1BroVtw	33.391027	-111.684482

Figure 2: Dataset Samples

Information about reviewers: There is also information about the reviewer, such as their name, the average number of stars they give, the number of reviews they have written, and how active they are in the Yelp community (shown by the "useful," "funny," and "cool" counts).

Review Metrics: Some important review metrics are the number of stars given and the number of votes for "useful," "funny," and "cool" on the review. By analyzing the reviews, one can understand customers' feelings and preferences. The dataset also includes detailed information about businesses, which can be sorted and filtered by type, location, and popularity. Reviewer data can help us understand user profiles better and make personalized suggestions that are more likely to appeal to people with similar interests. Therefore, this multifaceted dataset serves as a great starting point to build a complex and useful recommendation system.

3.2 Pre-processing

The pre-processing step starts with importing data from a CSV file and then does an exploratory data analysis to figure out how the dataset is structured. In this step, you will look at the data visually and learn basic things about the dataset, such as the types of columns and the number of non-null values. After that, 100,000 records are chosen at random from the dataset, and the index is reset for these records. There is a thorough check for missing values, as well as finding unique values and value counts in columns like "business name," "business neighborhoods," "business categories," "business stars," and "stars." The next step is to clean the data. Columns that aren't needed or are repeated are removed, and rows with null values are taken out. The next step is feature engineering, which makes new features based on how many words, capital letters, and special characters are in the review text.

3.3 Data Transformation

The text data from Yelp reviews goes through a number of steps that prepare it for analysis. To start, the text is cleaned up by removing spaces and special characters. This

makes sure that the dataset is cleaner. The next step is to return English contractions to their full forms. This makes the text more consistent and easier to understand. The next important step is to get rid of usual words that don't add much meaning. These words are called "stopwords." This list of stopwords can be changed by adding more words that aren't necessary. To make the text more meaningful, data from other columns, like "business categories," "business city," and "business name," is joined with the review text.

After this, the transformation involves using a method called stemming to break words down to their base or root form. This process simplifies and harmonises the text, which makes it better for analysis. As the reviews are being read, words are also put into groups based on how they make the reviewer feel, putting them into lists that show whether the feeling is positive or negative. This classification is very important for telling the reviews apart based on their type. The last part of the transformation is putting together all the processed words from all the reviews into a single string that makes sense. This final cleaned text, which is the original text in a format that is easier to process and analyse, is then added to the dataset in a new column. To make sure the text can be read, decoding from UTF-8 format is used to convert raw text to processed text. After these steps, the dataset is ready for more in-depth analysis or machine learning tasks. The noise has been reduced, and the text is more uniform.

In the dataset preparation process, the data is strategically divided into two subsets: a larger portion for training and a smaller portion for testing. This division is essential for machine learning applications, where the model is trained on one set of data and then evaluated on another to assess its performance. Approximately 90% of the data is allocated for training, ensuring that the model has substantial information to learn from. The remaining 10% is reserved for testing, which will later be used to gauge the model's predictive accuracy on new, unseen data. The features for both training and testing are derived from the processed review text, which serves as the input for the machine learning model. The target variable, indicating the sentiment of the reviews, is also separated for both the training and testing sets. This setup allows for a clear understanding of the number of samples and features included in each subset, laying the groundwork for effective model training and evaluation.

3.4 Feature Extraction

Once a divided data is obtained it is then subjected to feature extraction. In the feature extraction step, two different text vectorization techniques are used to extract the features that can be modelled using machine learning models.

3.4.1 Sentiment Score

The VADER Sentiment Analyzer (Hutto and Gilbert (2014)) is used for sentiment analysis as part of the pre-processing. This tool gives each review a sentiment score that is added to the star rating for the business and the star rating for the review to make a new "sentiment" feature. With a cutoff point of 7, this sentiment score is used to decide whether a review is positive or negative. These two-sided numbers are added to the "sentiment" column, and the number counts in that column are used to make a list of all the positive and negative reviews in the dataset. This in-depth pre-processing makes sure that the dataset is clean and full of useful features, which makes it possible to do

more in-depth analysis or modelling.

3.4.2 TF-IDF Features

TF-IDF (Term Frequency-Inverse Document Frequency) vectorization gives numbers to words in a document that show how important they are compared to a corpus, a collection of documents Ramos et al. (2003). By comparing how often a word appears in a document to how often it appears in the whole corpus, TF-IDF can help find words that are specifically important to a selected document. The TF-IDF value for a review is obtained by the formula given below.

$$TF - IDF(w, r, D) = TF(w, r) * IDF(w, D)$$

Where,

$$TF\ (w, r) = \frac{number\ of\ times\ word\ w\ appears\ in\ review\ r}{Total\ number\ of\ words\ review\ r}$$

And

$$IDF(w, D) = \log\left(\frac{total\ number\ of\ reviews\ in\ the\ dataset\ D}{total\ number\ of\ reviews\ with\ word\ w\ in\ them}\right)$$

A typical TF-IDF vector looks like the one shown in Figure 3 below.

```
array([0.        , 0.        , 0.30200157, 0.28487686, 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.3825814 , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.62331254, 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.26798802, 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.30272056, 0.28452665, 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.21986461,
       0.        , 0.        , 0.        , 0.        , 0.        ])
```

Figure 3: TF-IDF vector for a review

3.4.3 BERT Tokens

The BERT (Bidirectional Encoder Representations from Transformers) tokenizer is used in the study to get other sets of features. Traditional tokenization methods divide text into words or sentences. BERT's subword-based tokenization, on the other hand, breaks unknown words into smaller, more recognisable units. This lets the model figure out what even unfamiliar terms mean Devlin et al. (2018).Because it can guess missing words in text and process sentences without favouring one direction over another, BERT is very good at understanding what words mean and how they fit together in a sentence.

The review data is given to the BERT tokenizer, which turns it into tokens of specified lengths. The tokens are then turned into numerical IDs that are given to each unique word in the review. The lengths of these numerical sequences are then standardised by

padding, which makes sure that all data samples are of the same size. Figure 4 below shows BERT IDs for one of the review in the dataset.

```
'best mexican market arizona enorm like costco decor color wooden sign everywher turn unfortun like costco also pack grant loca
t new attract crowd special feel number custom wane much time soon price everi mexican food product could desir fabul fresh foo
d section includ juic stand yogurt stand salsa bar carnita pollo rostizado costilla section guisado plato combinato line taco c
ours fresh made warm tortilla plus outdoor grill pretti much like mexico except cleaner accept credit card anyth heaven groceri
mesa pro ranch market'

array([ 2190,  4916,  3006,  5334,  4372,  2953,  2213,  2066,  3465,
        3597, 25545,  3609,  4799,  3696,  2296,  2860,  5886,  2735,
        4895, 13028,  4609,  2066,  3465,  3597,  2036,  5308,  3946,
        8840, 11266,  2047,  9958,  4306,  2569,  2514,  2193,  7661,
       14071,  2063,  2172,  2051,  2574,  3976,  2412,  2072,  4916,
        2833,  4031,  2071,  4078,  4313])
```

Figure 4: BERT IDs for one of the reviews in the Dataset

4 Design Specification

This chapter presents the design specifications of our project. It delves into the conceptual framework, outlining the technical and functional requirements, system architecture, and the rationale behind design choices.

Figure 5 below shows the system architecture of the implemented study. The architecture consists of two parts viz. business part and the presentation part. The business part of the system is responsible for the processing steps that are performed in order to identify the best model for a recommender system. The presentation part is responsible for visualization of the data as well the results obtained through the comparison of the models. The recommendation provided by the system are also shown in the presentation part.

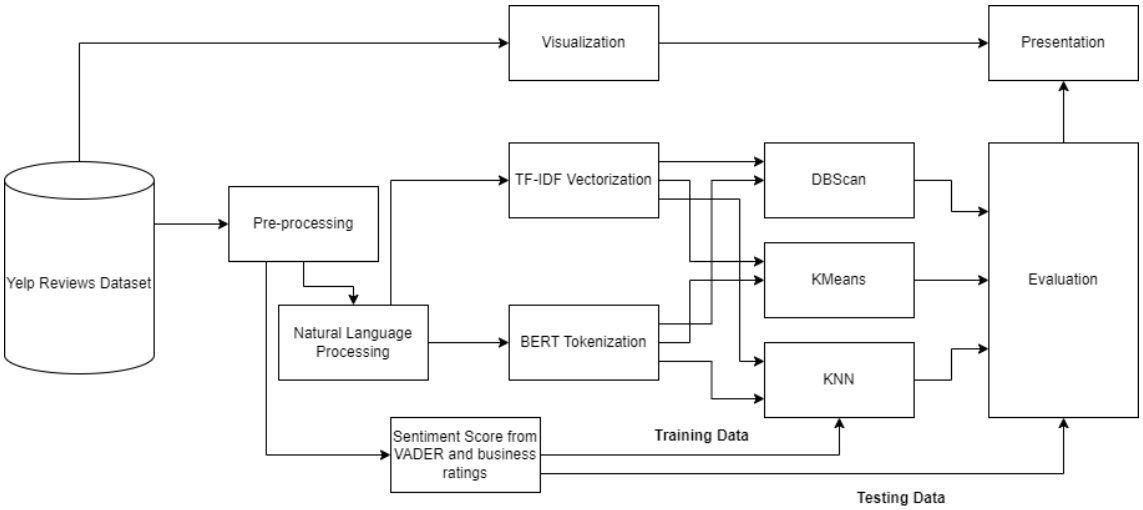

Figure 5: System Architecture

Once a 90:10 split features are obtained from either the TF-IDF vectorization or the BERT tokenization, it is given to the machine learning models. These models learn over the data to identify relationships between them. There are two types of machine learning algorithms that are being studied viz. unsupervised and supervised. The unsupervised algorithms include the DBScan algorithm and K-means clustering algorithm whereas the KNN model is a supervised one.

4.1 Modeling

4.1.1 K-means

K-means is a clustering algorithm that groups data into K separate clusters based on their proximity to the cluster centers. The algorithm assigns each data point to the closest cluster center and updates the center's position based on the average of the points in the cluster. K-means is popular because it is easy to use and works well, especially with large datasets as it does not require the class labels for the samples to be known beforehand. In recommender systems, K-means can be used to group businesses on the features, making it easier to provide recommendations. K-means requires the number of clusters in which the features are to be grouped. The number of clusters depends on the number of classes into which the classification is expected.

4.1.2 DBScan

The DBScan algorithm, short for Density-Based Spatial Clustering of Applications with Noise, is a method that groups data points based on their density. This tool finds points that are close to each other and marks as "outliers" points that are in areas with low density Ester et al. (1996) DBScan doesn't need to know beforehand the number of clusters to create like K-means does. This makes it very useful when working with complicated data structures or noise, since it can effectively tell the difference between noise and clusters. In recommender systems, DBScan can be used to find dense groups of features that are very similar. By focusing on features that are closely linked, it can help make better suggestions

4.1.3 K-Nearest Neighbors

As an instance-based learning algorithm, KNN is both simple and powerful. It is used for classification and regression. It works by finding a query point's K nearest neighbours and making guesses based on the properties of these neighbours Cover and Hart (1967). KNN is very useful in recommender systems because it can tailor recommendations to each user. KNN can make personalised suggestions that are very close to what a user wants by finding the users or items that are "nearest" to that user or item in the feature space.

4.2 Evaluation

The models mentioned are evaluated and compared based on the evaluation metrics of Accuracy, Precision, Recall, F1-score and Confusion Matrix.

4.3 Conclusion

This chapter described the design of the system implemented in the study. It gives the brief about the models used for the implementation of the recommendation system. A step by step working of the system is also described in the chapter. Once the design of the system is finalized, the implementation of the system can be performed. This is described in the upcoming chapter.

5 Implementation

5.1 Environmental Setup

This chapter describes the implementation of the system. This includes a step-by-step breakdown of the development process, challenges encountered, and how they were overcome, alongside a detailed description of the final system configuration.

5.2 Data Handling

The implementation of the system is done through two experiments. In the first experiment, a recommendation system based on the TF-IDF features, whereas in the second experiment, a system based on the BERT tokens is evaluated. The experiments have been implemented using the Python programming language because of its simplicity and the vast number of resources available for the language. The implementation is done in the Jupyter Notebook in the Anaconda environment.

5.3 Experiment 1: TF-IDF-based features for modelling

The TF-IDF features are obtained using Scikit Learn's feature extraction module. To do this, the text is broken up into tokens, and English stopwords are removed. To keep things consistent, it is also changed to lowercase. The length of the TF-IDF vector depends on the length of the text it is processing. As the reviews in the dataset will be of variable sizes, a static size of the features is required as models do not work on data of variable sizes. Hence, the number of features that can be extracted is limited to 45 to avoid the curse of dimensionality. The number is chosen such that it acts as a balance between computational complexity keeping most of the relevant information in the review intact.

5.3.1 DBScan

The DBScan model in the study is implemented using the cluster module of the Scikit-learn library. The model has been set up with eps=1 and min_samples=25. These parameters are chosen through extensive testing with different combinations. 10 different combinations of these values are tested to obtain the best accuracy for the model. The Epsilon (eps) in the implementation is the minimum distance between two points that are considered as part of the same neighbourhood, while min_samples is the minimum number of points required to form the dense region.

5.3.2 K-means Clustering

The K-means clustering algorithm in the study is implemented through the 'cluster' module of the Scikit-learan library. To implement this, certain values are set for it, such as random state=0, n_init=5, and algorithm="auto". The n_clusters parameter is used to determine how many clusters to make. For the study, two clusters are created to classify the features into. Random state in the implementation ensures that the results can be repeated whenever the code is run, n_init tells the algorithm how many times to run with different centroid seeds, and setting the algorithm parameter to "auto" allows it to determine the best way to fit the data.

5.3.3 KNN

Finally, the KNN model in the study is implemented using the 'neighbors' module of the Scikit-Learn library. KNN is a supervised machine learning algorithm hence it requires the class labels associated with the data samples beforehand. These class labels in the study are obtained through the sum of sentiment scores that are obtained through VADER sentiment analysis, the business rating feature from the dataset and the scores feature from the dataset as discussed in section 3.4.1. The KNN model in the study is implemented with parameters: n_neighbors=35, algorithm="auto," and metric="euclidean." The number of neighbours to look at is set to 35 by the n_neighbors parameter, which is selected as a balance between too small and too many features to compare with. The metric parameter is set to "euclidean," which uses the Euclidean distance to calculate the distance between two feature vectors.

5.4 Experiment 2: BERT Tokenization-based features for modelling

The "bert-large-uncased" model that is trained on a large corpus is chosen for the BERT tokenizer implementation, and is selected to provide 956 tokens. The process follows the steps mentioned in section 3.4.3. Feature sequences containing unique IDs are generated for all the cleaned reviews in the training set. These sequences are then processed further by padding (to make sure that sequence lengths are all the same) and truncation (to keep data input consistent), which limits the length of each sequence to 50 IDs. These generated feature vectors are then given to the classification models.

5.4.1 DBScan

The DBScan model is applied with eps=2 and min_samples=5, using the 'cosine' metric. These parameters allow the model to form clusters based on cosine similarity. The model clubs together most similar samples such that they are densely packed. The parameters for the model are selected in a process similar to that discussed in section 5.3.1.

5.4.2 K-means Clustering

A K-means model with n_clusters=2, random_state=0, n_init=5, and 'auto' algorithm is used for application on BERT token IDs.

5.4.3 KNN

The KNN model is implemented with a 'brute' algorithm, 'cosine' metric, n_neighbors=35, and 'distance' weights. The parameters are chosen because the accuracy provided by the model is highest for this combination.

5.5 Experiment 3: Using Python's Surprise Library

Finally, the Surprise library for Python used for implementing a recommendation system is implemented in the study for comparative analysis. This is done by trying three popular algorithms for recommendation systems viz. Singular Value Decomposition (SVD), KNN Basic and KNN with means. The cross_validate function of the library is used to cross

validate these algorithms with RMSE is as the validation metric.A 3-fold cross validation is used to achieve this.

6 Evaluation

This chapter is dedicated to the evaluation of our project. It encompasses the strategies used for testing, the success metrics, the results analysis, and a critical assessment of the project's effectiveness against its intended objectives.

Using the given dataset to test different machine learning models shows that they are not all as good as each other, as shown by metrics like accuracy, F1-score, precision, and recall. These measurements give a full picture of the pros and cons of each model. The evaluation results for the models are given in Table 1 below.

Model	Accuracy	F1-Score	Precision	Recall
DBSCan TF-IDF	17.92	17.92	17.92	17.92
KMeans TF-IDF	45.38	50.57	73.52	45.38
KNN TF-IDF	82.52	82.52	78.65	82.52
DBSCan Bert	17.92	15.19	8.96	50.0
KMeans Bert	33.02	35.00	71.40	33.03
KNN Bert	82.09	82.09	82.09	82.09

When we look at the DBScan over TF-IDF model, it gives the same score of 17.92% for all metrics. The fact that the accuracy, precision, recall, and F1-score are all about the same suggests that the model may be too biased toward one class or the dataset is very balanced or this could have arised due because of limiting the feature length to be 45. The KMeans TF-IDF model, on the other hand, is better, with an F1-score of 50.57% and an accuracy of 45.38%. It has a much higher precision rate of 73.52%, which means it is very good at finding positive sentiments from the dataset. However, its lower recall rate of 45.38% means it misses a lot of real positives. The KNN TF-IDF model does better than the first two with higher scores; it has an F1-score of 82.52% and an accuracy of 82.52%. But its high recall of 82.52% suggests that it is good at finding positive sentiments in the dataset. The models that use BERT tokenization perform in different ways. The DBScan model is almost the same as the TF-IDF model in terms of accuracy (17.92%), but it has a lower F1-score (15.19%) and a very lower precision of 8.96%, which means it makes a lot of false positive predictions. Its 50% recall rate means that it is correctly identifying half of the positive cases. The KMeans Bert model, on the other hand, does not do as well, with an F1-score of 35% and an accuracy 33.02%. Even though the model has a high precision of 71.40%, it has a very low recall making it unable to find most of the real positive cases. The KNN Bert model, on the other hand, works the best out of all the configurations. It achieves impressive score of 82.09% across all the metrics

6.1 Discussion

Although both the TF-IDF and BERT versions of the DBScan models work poorly, it's possible that this is because they can't handle the complexity and high dimensionality of text data well. The KMeans models do relatively well with TF-IDF but badly with BERT. This is probably because they have a hard time adapting to the complex feature space

```
Recommendations:

81880       RiteWay Catering & Food Truck
79203                        Crazy Buffet
21671            Jim's Shoe & Boot Repair
8692                    Superstition Farm
63993                  Imax Theater Tempe
45431    Wild Horse Pass Hotel and Casino
83541                               Prado
9001                      The Home Depot
48503                        Smashburger
21366                       Souper Salad
37302            Panchero's Mexican Grill
Name: business_name, dtype: object
```

Figure 6: Results for KNN-based recommendation system

that BERT creates. In contrast, the KNN models, especially those with BERT, perform better, achieving a good balance between accuracy, precision, recall, and F1-score. These results show that the KNN algorithm works well with the rich, context-aware features that BERT tokenization creates. To make text classification tasks go more smoothly, the overall analysis shows how important it is to choose the right tokenization methods and algorithms that are tailored to the dataset's unique features. Because the results were so different, it may be necessary to fine-tune the parameters or look into other methods that can handle the complexity of high-dimensional text data better.

6.2 Recommendation System

A KNN based recommendation system is implemented based on the results obtained. The KNN model is implemented with 'cosine' metric, and 'brute' algorithm. The results for the system to recommend a business is shown in figure below.

7 Conclusion and Future Work

This study successfully combines advanced data analytics, especially sentiment analysis, into a recommendation system. The goal is to make the service experience better for customers. The system looks at reviews, ratings, and feedback from customers to find out not only what they think about businesses but also how they feel about them. By taking both numbers and feelings into account, the system can make more personalised and emotionally relevant suggestions. This is a big improvement over traditional recommendation systems that don't always take into account complex customer preferences.

The information from this system can help retailers better manage their inventory, make targeted marketing plans, and make smart decisions, all of which will improve the overall performance of their businesses. The addition of sentiment analysis is a new idea

in recommendation systems that gives a fuller picture of how people behave and what they like.

Future Work In the future, there are many ways that this system could be improved and made better. In the future, more advanced machine learning algorithms and deep learning techniques could be added to make recommendations even more accurate and tailored to each person. Additionally, the system could be changed and tested in different types of businesses to see how well it works in different market situations. Real-time analysis and integration with e-commerce platforms could also be used to make instant suggestions based on live customer interactions and feedback.

The architecture and method of the system, which use both unsupervised and supervised machine learning along with advanced feature extraction techniques such as TF-IDF and BERT tokenization, makes these future improvements possible. As shown by the evaluation results, these techniques were successfully used, which shows that the system can handle complex data and draw useful conclusions from it. In conclusion, this project makes a big difference in the fields of retail business and data analytics by combining quantitative and qualitative insights. It also creates exciting new research and development opportunities in the areas of personalised service experiences and advanced recommendation systems.

References

Abbasi, F. and Khadivar, A. (2021). Collaborative filtering recommendation system through sentiment analysis, *Turkish Journal of Computer and Mathematics Education* **12**(14): 1843–1853.

Abdulsalam, A., Alhothali, A. and Al-Ghamdi, S. (2023). Detecting suicidality in arabic tweets using machine learning and deep learning techniques, *arXiv preprint arXiv:2309.00246* .

Ahmed, A. A. A., Agarwal, S., Kurniawan, I. G. A., Anantadjaya, S. P. and Krishnan, C. (2022). Business boosting through sentiment analysis using artificial intelligence approach, *International Journal of System Assurance Engineering and Management* **13**(Suppl 1): 699–709.

Al-Abbadi, L., Bader, D., Mohammad, A., Al-Quran, A., Aldaihani, F., Al-Hawary, S. and Alathamneh, F. (2022). The effect of online consumer reviews on purchasing intention through product mental image, *International Journal of Data and Network Science* **6**(4): 1519–1530.

Almahmood, R. J. K. and Tekerek, A. (2022). Issues and solutions in deep learning-enabled recommendation systems within the e-commerce field, *Applied Sciences* **12**(21): 11256.

Bag, S., Srivastava, G., Bashir, M. M. A., Kumari, S., Giannakis, M. and Chowdhury, A. H. (2022). Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion, *Benchmarking: An International Journal* **29**(7): 2074–2098.

Balush, I., Vysotska, V. and Albota, S. (2021). Recommendation system development based on intelligent search, nlp and machine learning methods., *MoMLeT+ DS*, pp. 584–617.

Birjali, M., Kasri, M. and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends, *Knowledge-Based Systems* **226**: 107134.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE transactions on information theory* **13**(1): 21–27.

Dadhich, A. and Thankachan, B. (2022). Sentiment analysis of amazon product reviews using hybrid rule-based approach, *Smart Systems: Innovations in Computing: Proceedings of SSIC 2021*, Springer, pp. 173–193.

Dang, C. N., Moreno-García, M. N. and Prieta, F. D. l. (2021). An approach to integrating sentiment analysis into recommender systems, *Sensors* **21**(16): 5666.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Dwivedi, Y. K., Hughes, L., Wang, Y., Alalwan, A. A., Ahn, S. J., Balakrishnan, J., Barta, S., Belk, R., Buhalis, D., Dutot, V. et al. (2023). Metaverse marketing: How the metaverse will shape the future of consumer research and practice, *Psychology & Marketing* **40**(4): 750–776.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, *kdd*, Vol. 96, pp. 226–231.

Huang, H., Asemi, A. and Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions, *IEEE Access* .

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the international AAAI conference on web and social media*, Vol. 8, pp. 216–225.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R. and Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making, *Industrial Marketing Management* **90**: 523–537.

Loukili, M., Messaoudi, F. and El Ghazi, M. (2023). Machine learning based recommender system for e-commerce, *IAES International Journal of Artificial Intelligence* **12**(4): 1803–1811.

Mehta, P. P., Dongare, O., Tekale, R., Umare, H. and Wanve, R. (2021). A survey on hybrid recommendation systems, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* .

Mohamed, M. H., Khafagy, M. H. and Ibrahim, M. H. (2019). Recommender systems challenges and solutions survey, *2019 international conference on innovative trends in computer engineering (ITCE)*, IEEE, pp. 149–155.

Osman, N., Noah, S. M. and Darwich, M. (2019). Contextual sentiment based recommender system to provide recommendation in the electronic products domain, *International Journal of Machine Learning and Computing* **9**(4): 425–431.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning*, Vol. 242, Citeseer, pp. 29–48.

Sadhasivam, J. and Kalivaradhan, R. B. (2019). Sentiment analysis of amazon products using ensemble machine learning algorithm, *International Journal of Mathematical, Engineering and Management Sciences* **4**(2): 508.

Sheng, J., Amankwah-Amoah, J., Khan, Z. and Wang, X. (2021). Covid-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions, *British Journal of Management* **32**(4): 1164–1183.

Stalidis, G., Karaveli, I., Diamantaras, K., Delianidi, M., Christantonis, K., Tektonidis, D., Katsalis, A. and Salampasis, M. (2023). Recommendation systems for e-shopping: Review of techniques for retail and sustainable marketing, *Sustainability* **15**(23): 16151.

Sun, Q., Niu, J., Yao, Z. and Yan, H. (2019). Exploring ewom in online customer reviews: Sentiment analysis at a fine-grained level, *Engineering Applications of Artificial Intelligence* **81**: 68–78.

Yi, S. and Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review, *Complex & Intelligent Systems* **6**(3): 621–634.