Advancing Sentiment Analysis on Review Data through Sentiment Classification

Sungyeon Lim

1. Research hypothesis & objectives

Online shopping behaviour in the United Kingdom (UK) has undergone significant changes in recent years. According to the International Trade Administration, online sales accounted for 36.3% of the total retail market in the UK as of January 2021, indicating a growing trend. This positioned the UK as the third largest eCommerce market globally [1]. With an increasing number of consumers shopping online, reviews have become increasingly influential in driving eCommerce sales. Consumers often rely on reviews when purchasing decisions, and positive reviews motivate potential consumers to buy products. Moreover, in today's highly competitive business landscape, reviews play a crucial role in shaping business strategies and decisions, providing valuable insights into consumer sentiments towards products and services.

This research focuses on sentiment analysis of reviews to improve sales and enhance business strategies. Sentiment analysis offers a way to analyse and extract insights from the enormous amount of textual data. The research builds on past studies in sentiment analysis and specifically applies them to the context of reviews. This can provide valuable insights into consumer preferences, enabling businesses to make data-driven decisions for product development, marketing strategies, and consumer engagement. Implementing advanced sentiment analysis techniques using Python and SketchEngine on review data will improve the accuracy and interpretability of sentiment analysis results.

The research aims to achieve two main objectives. First, a model will be developed to analyse sentiment in reviews accurately. Different algorithms will be applied to enhance the model's accuracy by utilising various NLP techniques and fine-tuning. The model's performance will be assessed using appropriate metrics. Second, the sentiment analysis results will be used to derive insights and strategies to make data-based decisions. To achieve this, SketchEngine will be used to extract keywords to analyse factors contributing to positive and negative sentiments in reviews.

2. Background

Sentiment analysis is an approach to identifying emotions in text using Natural Language Processing (NLP) techniques. The process involves five steps. First, data is collected from sources such as social media, consumer reviews, or surveys. Second, the collected data is preprocessed to remove noise and irrelevant information. This is achieved using NLP techniques such as converting text to lowercase, removing punctuation, removing stop words, stemming and lemmatization, etc. Third, relevant features are extracted from the preprocessed text data to convert it into a suitable data format for sentiment analysis. This can be done using methods such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings. Fourth, sentiment analysis is conducted by selecting an appropriate approach. There are two common approaches used for sentiment analysis: machine learning and lexicon-based approaches. Finally, the sentiment analysis results are evaluated using various metrics such as accuracy, precision, recall, F1-score, or ROC-AUC. Additionally, the performance can be validated using a separate test dataset to assess its generalisation ability.

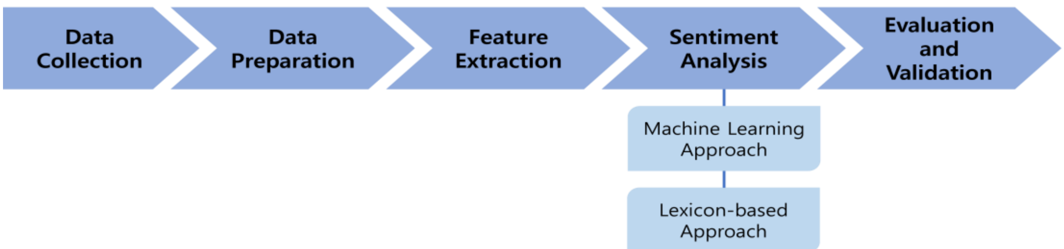

Figure 1 Sentiment analysis process

As previously mentioned, there are two commonly used approaches for sentiment analysis. They differ in the method used and offer their strengths. The first approach is the machine learning approach, which involves training a classification model to identify emotions using machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes. This approach offers high predictive accuracy and has the ability to learn complex patterns and relationships in text data. It is also adaptable to various domains. However, it requires a large amount of data and may suffer from overfitting to the training data. Additionally, deep learning algorithms such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) can be used, which have shown remarkable performance in sentiment analysis, but require more data to train.

The second approach is the lexicon-based approach, which relies on sentiment lexicons or dictionaries containing words or phrases annotated with sentiment labels, such as positive, negative, or neutral. This approach determines sentiment by matching words or phrases in the text with entries in the lexicon and calculating sentiment scores. The lexicon-based approach is transparent and interpretable, allowing quick and straightforward sentiment analysis. Also, as it relies on lexicons, it does not require labelled training data. However, lexicons may not be able to capture variations in sentiment expressions. Moreover, this approach may struggle to accurately interpret sentiment in context-dependent language or ambiguous expressions.

Past research has investigated various methodologies for sentiment analysis. This research builds upon existing research in sentiment analysis and aims to contribute to the advancement of sentiment analysis in the context of reviews.

3. Importance and contribution to knowledge

Advancing sentiment analysis on reviews is highly significant and can offer several contributions. One of the primary benefits of using consumer sentiment analysis is that it allows businesses to gain a deeper understanding of consumer preferences and opinions. With this knowledge, businesses can make optimised decisions regarding product development, marketing strategies, and pricing decisions. This, in turn, leads to more targeted and effective business strategies. Moreover, by identifying areas for improvement and catering to the needs of their consumers, businesses can increase consumer satisfaction and loyalty. This can translate into repeat purchases, revenue growth, and long-term success. It is also important to analyse reviews of competitors to understand the market. By continuously monitoring and analysing sentiment trends, businesses can stay ahead of competitors, identifying market opportunities. In summary, advancing sentiment analysis on reviews empowers businesses to make data-driven decisions and enhance consumer satisfaction, ultimately contributing to increased profitability and long-term economic success.

Secondly, advancing sentiment analysis on reviews can have a significant impact on national and international research initiatives by enabling interdisciplinary collaboration. By analysing reviews across various sectors, such as hospitality, healthcare, and entertainment, sentiment patterns and trends can be uncovered. These insights can provide essential inputs for national and international research to further analyse consumer behaviour, market dynamics, and societal preferences across different industries. Moreover, analysing sentiments expressed in reviews in other linguistic contexts can offer crucial insights into cultural nuances, preferences, and sentiments across various regions and countries. This cross-cultural understanding can contribute to diplomatic strategies and global business expansions. To sum up, research on sentiment analysis of reviews has the potential to generate meaningful insights, facilitate interdisciplinary collaboration, and contribute to national and international research.

4. Pilot study

A pilot study is conducted by selecting a sample of product reviews to examine the feasibility of the research. The data is preprocessed to perform sentiment analysis using NLP techniques. Appropriate methodologies and techniques are applied based on past research on sentiment analysis and the characteristics of review data. The main objective is to evaluate the performance of sentiment classification and analyse reviews based on the classification results.

The dataset used for the study is the Women's E-Commerce Clothing Reviews, which is available on Kaggle. This dataset contains 23,000 reviews with ten variables, including Clothing ID, Age, Title, Review Text, Rating, Recommended IND, and more. However, only the Title, Review Text and Recommended IND variables are used for this pilot study. A new dataset is created with only two variables, combining the Title and Review Text variables, as both contain textual data relevant to sentiment analysis. The Recommended IND variable is a binary variable that indicates whether the consumer recommends the product, where one means recommended, and 0 means not recommended.

The first step in performing sentiment analysis is to preprocess the data. This involves removing missing data. In this dataset, there are 844 instances where both the Title and Review Text are missing, which need to be removed. Then, several NLP techniques are applied, including converting text to lowercase, removing punctuation, removing stop words, removing numbers, removing rare words, and lemmatization.

As previously mentioned, there are two approaches to sentiment analysis. Since product reviews tend to include context-dependent language and various expressions, it can be concluded that the machine learning approach is more suitable. Therefore, several machine learning algorithms are applied to find the best algorithm for this particular dataset. In this pilot study, two different vectorization methods, BoW and TF-IDF, are used to prepare the data for modelling. In this case, CountVectorizer and TfidfVectorizer in Python are applied. Moreover, product reviews often have imbalanced classes, as consumers tend to recommend products unless they are significantly worse than expected. This dataset has 18,535 recommended instances and 4,101 not recommended instances. This imbalance must be considered when modelling.

Various algorithms are utilised to find the most suitable model for the given data. First, the dataset is split into training and test datasets, with an option of 'stratify' to ensure that the dataset is divided in a way that considers imbalanced classes. Two types of vectorizations are then performed: CountVectorizer and TfidfVectorizer. With the vectorised text data, multiple algorithms are applied, including Logistic Regression, Multinomial Naïve Bayes, SVM, Random Forest and Ada Boosting.

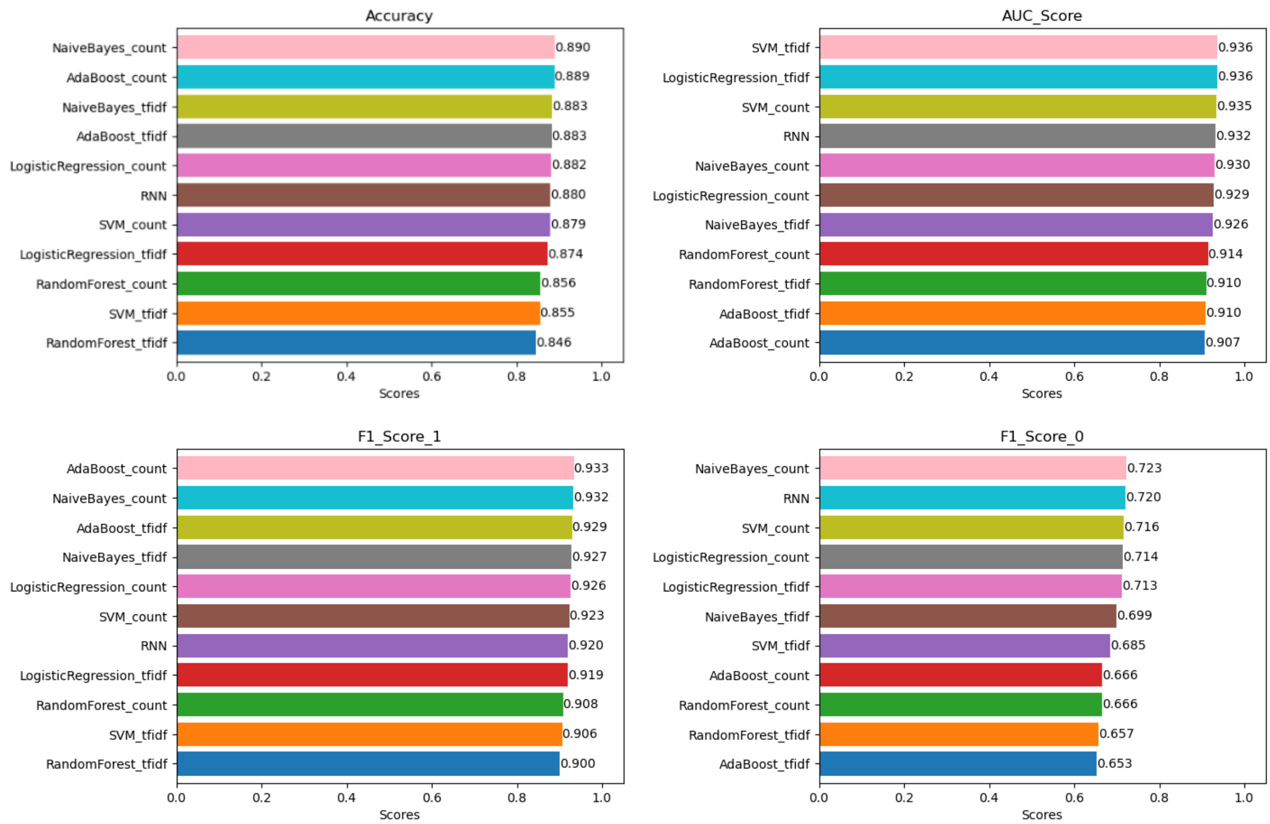

Figure 2 Comparing different models

The models are evaluated using metrics such as accuracy, f1-score and AUC score, which are suitable for imbalanced classification. These metrics are calculated by cross-validation. For further analysis, a deep learning algorithm, RNN, is also applied. For the deep learning model, word embedding is used, and sequences are padded. Fine-tuning is then performed to evaluate each model and find the best parameters.

Figure 2 visualises the accuracy, f1-scores for classes 1 and 0, and AUC score for each model to compare the models with different algorithms. Based on the scores, it can be concluded that the Multinomial Naïve Bayes model with CounterVectorizer classifies the sentiment more accurately. Moreover, Figure 3 provides a detailed classification result of the selected model with the confusion matrix and ROC-AUC curve. It is evident from the results that the model predicts the sentiment with a high level of accuracy, scoring an accuracy of 0.89 and a weighted f1-score of 0.91 with the test dataset.

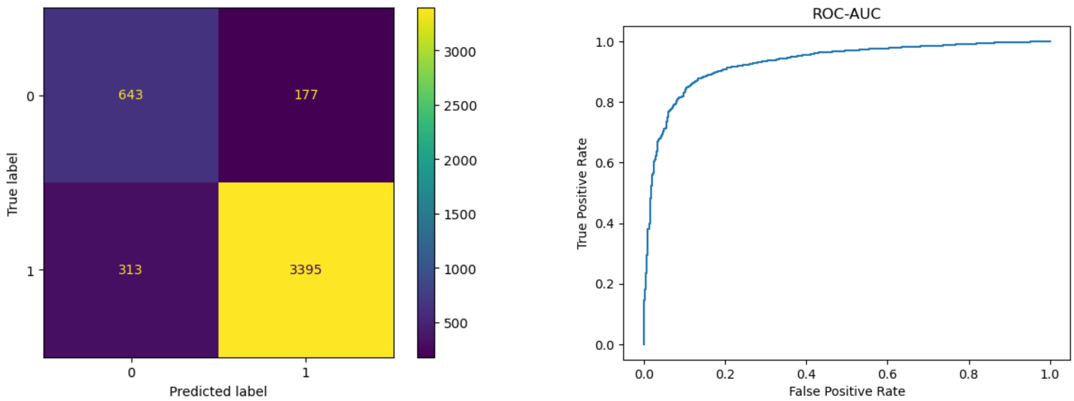

Figure 3 Evaluation of the selected model

Then, the reviews are analysed by dividing them into two groups, classified as positive or negative by the trained model. One way to interpret these collections is by using WordCloud in Python. By visualising word clouds, frequently mentioned words are identified. In Figure 4, words such as ‘comfortable’, ‘fabric’, ‘love’, and ‘perfect’ are shown in the positive reviews, whereas words such as ‘material’, ‘fabric’, ‘size’, and ‘small’ are shown in the negative reviews. Additionally, SketchEngine can be used to analyse these collections further. Not only does SketchEngine extract single words, but it can also extract frequent multi-word expressions. By creating two separate corpora of positive and negative reviews and utilising the tool n-grams, frequent multi-word expressions are identified. In the positive reviews, words such as ‘true size’, ‘fit perfectly’, ‘usual size’, and ‘super soft’ are extracted. In contrast, negative reviews include words such as ‘run small’, ‘didn’t work’, ‘size small’, ‘poor quality’, and ‘arm hole’. This analysis can help identify areas where further improvements are needed.

Figure 4 Word clouds of positive and negative reviews

In conclusion, this pilot study demonstrates the feasibility of further research.

5. Programme and methodology

The research will follow the CRISP-DM methodology. This methodology consists of several phases:

business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

The first step involves understanding the background and objectives of the business to determine a precise business problem. This includes defining business problems, determining business goals, and identifying stakeholders. In this case, the research aims to advance sentiment analysis in the context of reviews. By defining a clear business problem, advanced sentiment analysis will help to enhance sales and business strategies. Then, stakeholders such as marketing and product managers, who can benefit from sentiment analysis insights, are identified.

The second step involves understanding the data. This process includes both collecting and exploring the data. Firstly, data is collected from various sources such as social media, product reviews, surveys, and business-related articles. The collection process can be carried out by web crawling and taking surveys. Then, the collected data is explored and examined to identify its characteristics, as well as its quality, relevance, and suitability for sentiment analysis. This can lead to understanding how data mining can contribute to achieving the set goals.

Then, the data preparation step involves preprocessing and extracting features from the data to make it suitable for sentiment analysis. This includes removing missing data, removing duplicates, and applying NLP techniques in Python. Several NLP techniques can be used, such as converting text into lowercase, removing punctuations, removing stop words, removing numbers, lemmatization, and removing rare words. These techniques help to extract relevant data. After preprocessing the data, feature extraction can be done using different methods, such as WoB, TF-IDF, and word embedding. To evaluate the best model for the data, all three methods are utilised and compared.

The next step involves developing a sentiment analysis model. This includes training and selecting the best model for the preprocessed data. Based on the understanding of the review data, it can be concluded that data includes context-based language and various expressions. Hence, the machine learning approach can be utilised. Moreover, as the data tends to have imbalanced classes, the imbalance needs to be considered by an option called ‘stratify’ when splitting the dataset for training. Five different machine learning algorithms commonly used for sentiment classification will be applied to find a suitable model: Logistic Regression, Multinomial Naïve Bayes, SVM, Random Forest and Ada Boosting. Both feature extraction methods, WoB and TF-IDF, are used to examine each algorithm. Additionally, a deep learning model, RNN, is trained with word embedding. Lastly, fine-tuning is done to optimise the best hyperparameters for each model.

The fifth step involves evaluating and validating the models. In order to assess the models, appropriate metrics such as accuracy, f1-score and AUC score are used, which are suitable for imbalanced classes. These metrics can be calculated by taking averages through cross-validation. Based on these metrics, the best-performing model is selected. Once the model is selected, it is validated using the test dataset. This step is critical as it ensures the model's reliability for sentiment analysis. To further analyse the sentiment classification results, SketchEngine's n-grams tool is utilised to identify keywords with high frequency from two separate corpora, positive and negative reviews. This provides valuable insights for improvement.

In the final step of the process, the trained model is deployed into the business to enable data-driven decision-making. It is important to ensure the model is easily accessible to all relevant users and stakeholders. User evaluation is then conducted by recruiting participants representing target user groups to evaluate the research based on user feedback. The feedback is then analysed and incorporated into refining the sentiment analysis models and methodologies. By monitoring the model’s performance and collecting feedback from users through user evaluation, the research can be further developed and maintained. Then, documentation and training are provided to stakeholders to enable them to utilise the sentiment analysis results more effectively.

In order to manage the research and ensure it stays on track, several milestones and deliverables are established. The first milestone involves completing the research plan and obtaining approval from stakeholders. Once this is achieved, a research plan outlining objectives and timelines can be provided. The second milestone is the completion of data collection and preprocessing. At this stage, a cleansed and preprocessed review dataset suitable for sentiment analysis will be prepared. The next milestone involves the development and validation of sentiment analysis models. Upon completion, trained sentiment analysis models will be ready to classify the preprocessed review data. The fourth milestone is the assessment of model performance, which includes an evaluation report of the selected model along with model performance metrics. The following milestone involves collecting and analysing user feedback on the deployed model, resulting in a user feedback report on model usability. The milestone is completing the final model adjustments based on user feedback and documenting the entire research process, which includes providing a finalised sentiment model and documentation for stakeholders.

This completes the entire research process. Figure 5 illustrates the research phases along with timelines, denoting the milestones with stars.

6. Workplan diagram

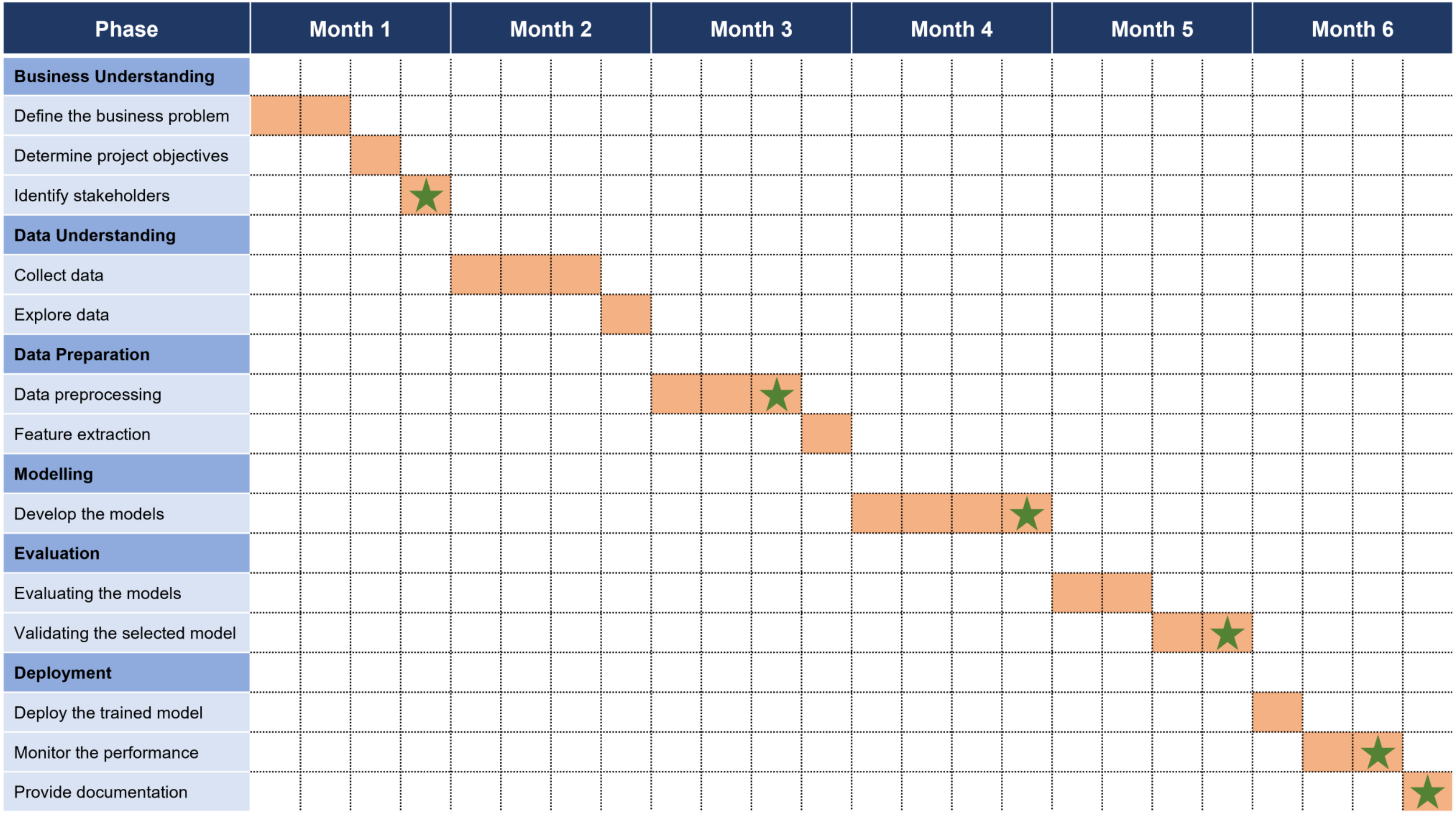

Figure 5 Workplan diagram

7. References

[1] International Trade Administration. 2023. United Kingdom – eCommerce. [Online]. [Accessed 2 April 2024]. Available from: https://www.trade.gov/country-commercial-guides/united-kingdom-ecommerce

[2] Xing, F. and Zhan, J. 2015. Sentiment analysis using product review data. Journal of Big Data. 2(5), pp.1-14.

[3] Shivaprasad, T.K. and Shetty, J. 2017. Sentiment Analysis of Product Reviews: A Review. International Conference on Inventive Communication and Computational Technologies (ICICCT) 2017, pp. 298-301.

[4] Brooks, N. 2018. Women's E-Commerce Clothing Reviews. Kaggle. [Online]. [Accessed 28 March 2024]. Available from: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews/data.

[5] Jurafsky, D. and Martin, J. 2020. Speech and Language Processing. 3rd Edition draft, pp.4-31.

[6] Bird, S., Klein, E., and Loper, E. 2009. Natural Language Processing with Python. O'Reilly Media, pp.221-259.

[7] Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K., and Caro, J. 2013. Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning. IISA 2013, pp.1-6.

8. Appendix

Word

Word is a widely used word-processing software developed by Microsoft that enables users to create, edit, format, and collaborate on documents. It offers a variety of templates tailored to different document types and allows users to customise them according to their needs. Word provides numerous formatting options, including fonts, styles, sizes, colours, and alignment options, that help increase readability and enhance visual appeal. Word also includes tools for spell-checking and autocorrection to ensure that documents are free from errors.

This research paper is written in Word, utilising different document features. It has been customised using a template with 2cm margins, Arial font, size 11, and single spacing.

Grammarly

Grammarly is an advanced writing assistance tool. It offers a range of features to enhance the quality of writing. First, it can perform a spelling and grammar check, automatically detecting and correcting spelling mistakes, punctuation, and grammatical inaccuracies. Furthermore, Grammarly evaluates the consistency of writing style, helping maintain a consistent tone throughout the document. Additionally, it suggests alternative vocabulary choices to enrich the writing and avoid repetition, enhancing clarity.

This research paper uses a tool to correct misspellings and punctuation errors and provides suggestions for improving the clarity and engagement of the document.

ChatGPT

ChatGPT is an Artificial Intelligence (AI) based conversational agent developed by OpenAI. It employs the Generated Pre-trained Transformer (GPT) architecture to create human-like responses based on user inputs. ChatGPT can engage in natural language conversations on a wide range of topics and provide answers and recommendations.

In this research, ChatGPT is ustilised for recommendations of topics of the research. Here is an example of a query and its results.

Query: Recommend topics for a 6-month research project on sentiment analysis

Answer: Sure! Here are some potential topics for a 6-month research project on sentiment analysis:

1. Deep Learning Architectures for Sentiment Analysis: Investigate the performance of different deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer-based models like BERT for sentiment analysis tasks.
2. Multimodal Sentiment Analysis: Explore how combining textural, visual, and auditory modalities can improve sentiment analysis accuracy, especially in contexts like social media or product reviews where multiple types of data are available.
3. …

   These topics cover a range of areas within sentiment analysis, from technical aspects like model architectures and multilingual processing to more applied areas like domain-specific analysis and ethical considerations. Depending on your interests and expertise, you can choose a topic that aligns well with your goals for the research project.

Google Scholar

Google Scholar is a search engine that provides access to a vast collection of scholarly literature from various academic fields. It offers a database of academic publications, including journal articles, theses, and dissertations across different disciplines. Additionally, its citation tracking feature helps researchers identify other relevant papers within their field of study.

Through the research on sentiment analysis, Google Scholar is used to access a diverse range of publications related to the field as well as relevant disciplines. It helps to identify existing research studies and methodologies related to sentiment analysis. By entering relevant keywords and phrases, Google Scholar can search through a vast collection of scholarly articles.

Kaggle

Kaggle is an online platform that provides access to a vast collection of datasets. It allows users to participate in competitions and collaborate with others from all over the world. Kaggle helps researchers search for and download relevant datasets for their research. It also provides a platform for sharing and exploring data analysis code.

In this research, the dataset, the Women's E-Commerce Clothing Reviews, is downloaded from Kaggle.

Python

Python plays a critical role in text preprocessing and sentiment analysis. Python libraries offer various tools for preprocessing, feature extraction, modelling, evaluation, and visualisation.

In this research, several libraries are utilised for sentiment analysis. Initially, NLTK is used for text preprocessing, which involves performing tasks such as tokenization, stopword removal, lemmatization, etc. Another library called scikit-learn provides tools for modelling and evaluation, such as train_test_split, confusion matrix, CountVectorizer, TfidfVectorizer, Logistic Regression, Naïve Bayes and so on. Additionally, the results are visualised using matplotlib and WordCloud libraries.

Below are some examples of codes used for the pilot study.

```
# stopword removal
import nltk
from nltk.corpus import stopwords
STOPWORDS = set(stopwords.words('english'))
def remove_stopwords(text):
    return " ". Join([word for word in str(text).split() if word not in STOPWORDS])
```

```
# train_test_split
import sklearn.model_selection import train_test_split
X = df_text['Combined']
y = df_text['Recommended IND']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y,
                                                     random_state=101)
```

```
# CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
X_train_c = vectorizer.fit_transform(X_train)
X_test_c = vectorizer.transform(X_test)
```

```
# Naïve Bayes
from sklearn.naive_bayes import MultinomialNB
model1 = MultinomialNB()
model1.fit(X_train_c, y_train)
```

```
# WordCloud
from wordcloud import WordCloud
wordcloud = WordCloud(background_color='white', max_words=len(pos_words),
                max_front_size=40, relative_scaling=0.5, colormap='summer).generate(pos_words)
plt.figure(figsize=(13, 13))
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

SketchEngine

SketchEngine is a web-based platform that offers a wide range of functions for corpus linguistics research. It provides access to a vast range of corpora in multiple languages, covering various domains of text. It also offers several tools for querying corpora, analysing collocations and word combinations, generating concordances, extracting n-grams and keywords, and conducting statistical corpus analysis.

In this research, SketchEngine is used to create two separate corpora for positive and negative reviews. This is done by uploading CSV files containing text data of positive and negative reviews. The keywords with high frequency are extracted from the corpora using the tools of wordlist and n-grams. Not only single words are extracted, but multiword expressions are also extracted by the tool n-grams.