Paradigm Academic Press
Frontiers in Management Science
ISSN 2788-8592
JUN. 2023 VOL.2, NO.3

Evaluation & Analysis of Movie Aspects: Based on Sentiment Analysis

Yuan Xu[1]

[1] WMG, University of Warwick

Correspondence: Yuan Xu, WMG, University of Warwick.

doi:10.56397/FMS.2023.06.08

Abstract

The objective of this project is to address the research question "What factor determines the success of an action movie?", where the success factors are quantified as plot, music, plot, and director. Based on the research question, the author conducts a literature review of the development of the UK film industry, the impact of social media and film criticism on the film industry. To evaluate film reviews, sentiment analysis, which can efficiently convert text into structured data, is utilized in the project. In the methodology and the case study, the method of constructing the VADER model and the SVM model with the process including data processing, modelling, result correctness verification, and model evaluation are expounded and shown. Finally, the correctness of the two models is verified and the model output is visualized for analysis. The conclusion of the project is as follows: (1) Responding to the inquiry: In action films, the plot is the aspect that the audience pays the most attention to, but first the actors and music, and finally the director. (2) When compared to the experimental findings of Wang *et al.* (2020a), the project came to different conclusions, presumably as a result of the project's limited access to movies and the change in the data processing approach. (3) Given that the two models employed in this research did well in the evaluation, they can be effectively applied to help production firms raise the caliber of action movies. (4) The models in the research can be easily developed and used to other domains because film review analysis can be compared with film reviews from other industries. However, some restrictions are also brought up, which offers suggestions for further work.

Keywords: social media, Sentiment Analysis, Python, Machine Learning.

1. Introduction

1.1 Project Definition

This project is inspired by public media platforms on the Internet, which have been impacted by diverse cultures and have become an essential part of modern life. Social media is used by people to access not only their lives but also their thoughts and feelings about various goods, services, and companies through comments and sharing, therefore, more and more valuable data on the Internet has become textual form (Xu *et al.*, 2022). However, traditional analytical methods have difficulty addressing textual data, and extracting effective information (features) from massive information has become the key to utilizing network data (Tabinda Kokab *et al.*, 2022; Xu *et al.*, 2022). The essential idea of the project is to examine social media reviews to assess and enhance product quality using an advanced and effective methodology. The development of e-commerce while analyzing data using computer science methods is precisely related to the requirements for my degree.

1.2 General Background and Motivation

1.2.1 Background of Social Media Platforms

With the popularity of the Internet, social media, such as Twitter, Instagram and Facebook, has been widely used by people and has become an important part of the information age (Xu *et al.*, 2022). In addition to sharing everyday life, social media is also being used to discuss the strengths and weaknesses of brands, product quality, services, and current events. For e-commerce, social media has evolved into a medium for businesses and consumers to interact directly (Bilal *et al.*, 2016).

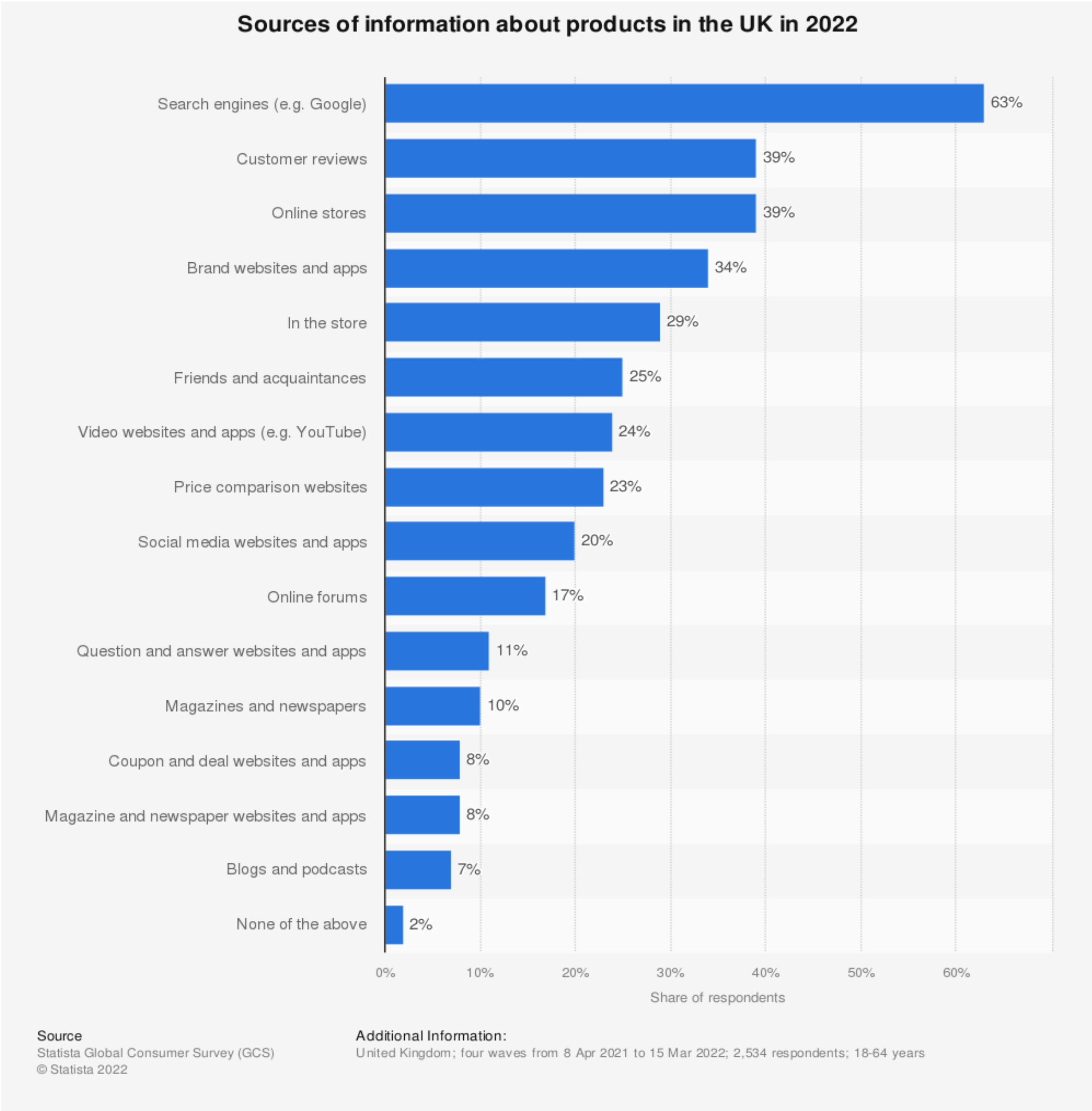


Figure 1. Source of product information in the UK (Statista, 2022)

Alternatively, consumers use social platforms to exchange ideas, recommend products and make relevant recommendations for brands (Abd El-Jawad *et al.*, 2018). Figure 1 shows in the UK, 39% of product information comes from online consumer reviews, compared to 29% from offline sources. Furthermore, the continuous optimization of search engines has given consumers a better online shopping experience, which also means that online user reviews will become more important in products (Erdmann *et al.*, 2022; Statista, 2022).

1.2.2 Motivation

Sentiment analysis is an assignment of analyzing the opinions, sentiments, attitudes, and perceptions of texts about different entities such as topics and products by extracting important information from texts, which is the most common and efficient method for analyzing data in text form (Pang & Lee, 2008). With the rapid development in various fields in recent years, the output of sentiment analysis is not only the different emotions in the text, such as happiness, sadness, and anger, but also the intentions contained in the reviews, which enables companies to understand consumers and ensure customer relationships are maintained (Abbasi *et al.*, 2008; Huber *et al.*, 2000; Sudhir & Suresh, 2021).

1.2.3 Background of Sentiment Analysis

As shown in Figure 2, there are three basic types of sentiment analysis: machine learning, lexicon-based method, and hybrid.

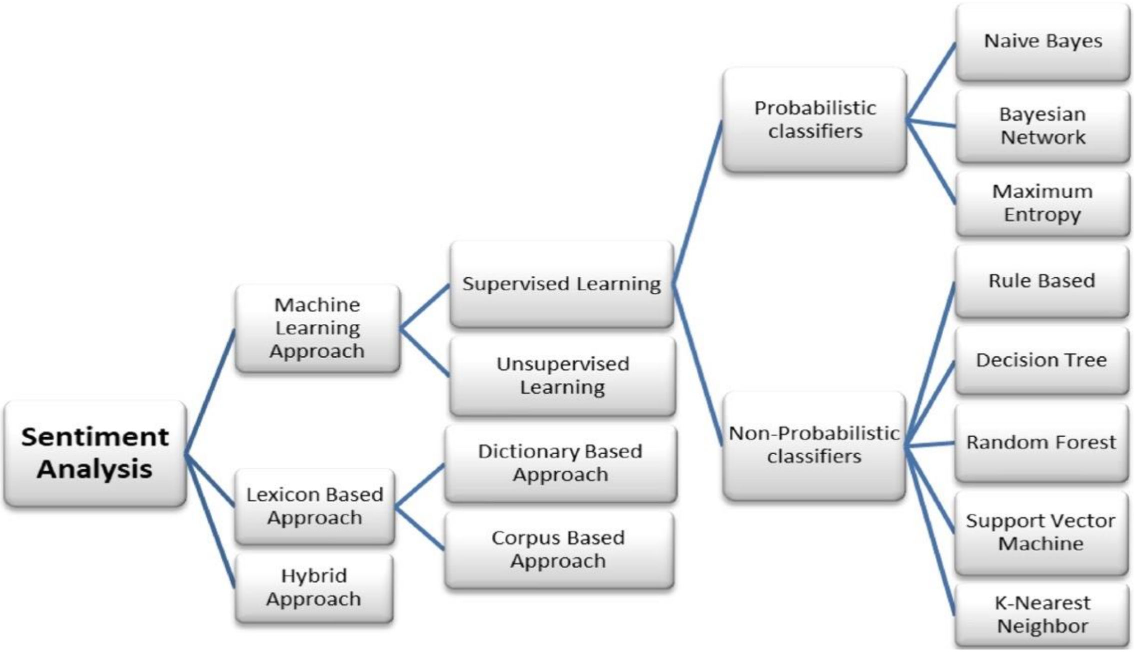


Figure 2. Categories of Sentiment Analysis (Jindal & Aron, 2021)

The first method is to process text data based on various machine learning algorithms, which can be further broken down into supervised and unsupervised learning techniques (Jindal & Aron, 2021). Basically, machine learning methods extract keywords in text which would be regarded as features to learn and analyze (Rodrigues *et al.*, 2018). The lexicon-based method assumes that synonyms have the same sentiment polarity and antonyms have opposite memory, and the text is analyzed by creating a universal dictionary (Birjali *et al.*, 2021).

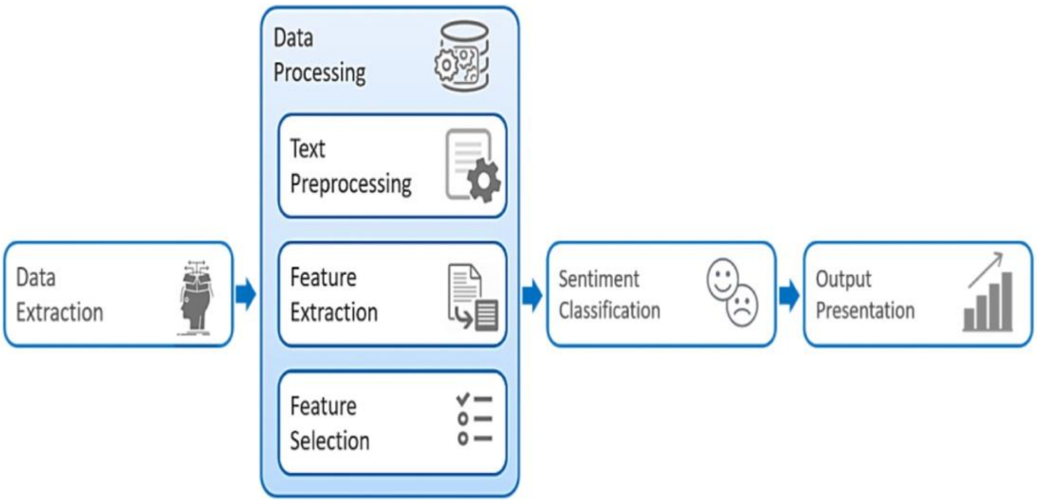


Figure 3. The general workflow of sentiment analysis (Birjali *et al.*, 2021)

Figure 3 depicts the general process of sentiment analysis. Basically, there are four stages in this process, which are data collection, data processing, sentiment classification and model evaluation. Data collection refers to the creation of datasets by extracting from the collected data which frequently social media, review sites, blogs, and forums, which comes from a variety of sources and tools (Birjali *et al.*, 2021; Mei & Frank, 2015). Data processing consists of three parts: text preprocessing, feature extraction and feature selection. The objectives of text reprocessing are for cleaning noises and text data with incorrect grammars and dimensionality reduction (Birjali *et al.*, 2021; Liu, 2012). Feature extraction is the extraction of valuable information (verbs, adjectives, nouns, etc.). The last step of data processing is the feature selection, which is also influenced by feature extraction. Feature selection is to filter out relevant features, so that the data dimension is reduced again, which helps to improve the accuracy and efficiency of the sentiment analysis (Ahmad *et al.*, 2019; Birjali *et al.*, 2021).

1.3 Research Question and Objectives

Based on the Section 1.2.2, the general research question of the dissertation can be defined as "What factor determines the success of an action movie?". To address the research question, the dissertation consists of the following objectives:

1) To review related information of the dissertation, the contents of literature should be evaluated, including but is not limited to the background of the film industry, the detailed introduction, process, and evaluation of sentiment analysis.

2) To construct the process for the project systematically, a methodology should be created.

Based on Figure 3, there are four sections which must contain in the methodology, which are data collection, data processing, sentiment classification, and evaluation, where:

- Data collection aims for extracting high-quality reviews of action films on the Internet.
- Data processing refers to make the model better analyse the data and the data fit the model efficiently.
- The efficiency and correctness should be evaluated after completing the model of sentiment analysis. If necessary, testing and debugging should be inserted to improve performance of the model.
- The result/output of sentiment analysis should answer the research question directly.

3) To accomplish the dissertation, the conclusion of the research questions should be conducted by evaluating and analyzing the results derived from the accomplished model.

2. Literature Review

2.1 The Background of the Film Industry

To expressly describe the background of the film industry, Section 2.2.1 illustrates the continuous profitability of the film industry through the development and evaluation of the industry. Furthermore, with the advent of the information age, massive amounts of data appear on social media in text form, serving as a medium for dialogue between studios and consumers (Section 2.2.2). How to efficiently analyze film reviews has become a crucial problem for the film industry.

2.1.1 The Development and Evaluation of the Film Industry in the UK

'Film' came into existence in the early 20th century, which was regarded as a medium for the study of movement initially (Simonton, 2011). Significantly, due to the cultural significance and circulation of films, the film industry, which has become a transnational industrial chain, spreads widely around the world and has a large impact on the global economy (Eliashberg *et al.*, 2006; Gomery, 2015).

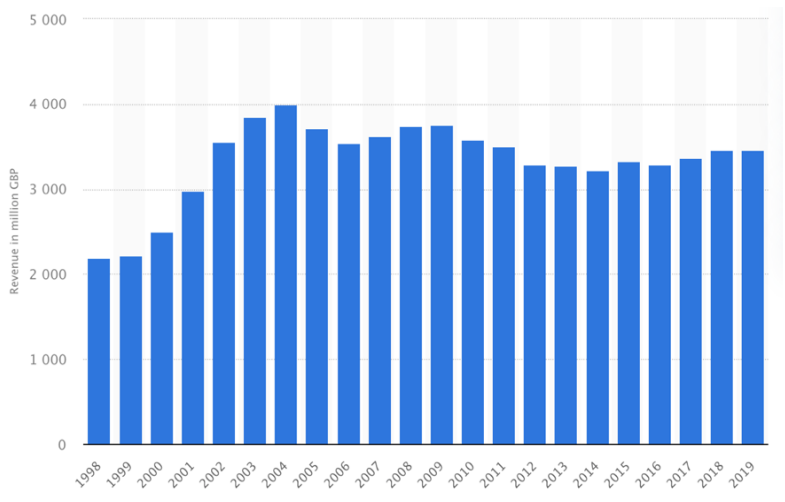


Figure 4. The revenue of the film industry in the UK (British Film Institute, 2021a)

With the advent of the digital age, information technology has brought about tremendous changes to the development of the film industry. Especially, video-on-demand (VoD), which is originated in media industry, has become the mainstream online viewing method of films (Kim & Kim, 2017; Pereira & Tam, 2021). Using VoD technology, the delivery method of movies has become convenient, consumers no longer need to go to the cinema to watch movies, instead, it can be done freely at home using devices that can connect to the Internet (such as iPad, PC, and TV).

Figure 5 shows the revenues for on-demand audiovisual services in the UK from 2011 to 2018. Compared to Figure 4, revenue from on-demand audiovisual services in 2018 was already more than half of the UK film industry's revenue. Based on traditional sales, online viewing mode is more favored by consumers, and it also reduces the cost of film companies to release movies.

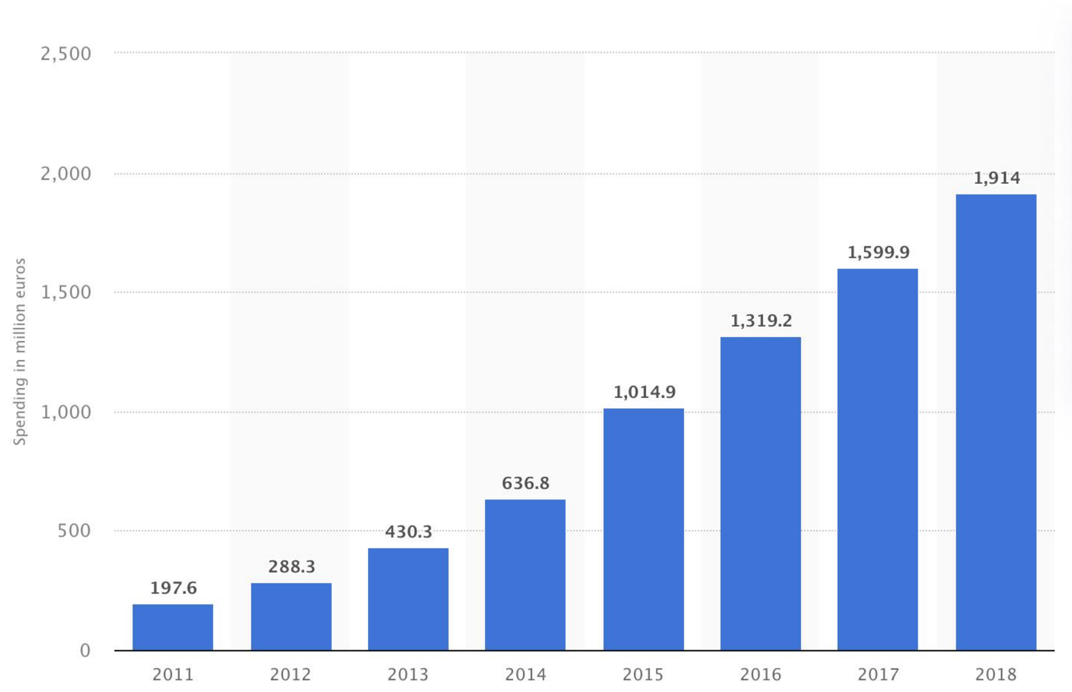


Figure 5. Consumer revenues for on-demand audiovisual services in the United Kingdom (UK) from 2011 to 2018 (European Audiovisual Observatory, 2019)

Furthermore, with the outbreak of COVID-19, the structure of the film industry has changed, while also promoting the development of online movie viewing. Figure 6 presents the changes in the way consumers watch movies from 2009 to 2019 (the outbreak of the epidemic).

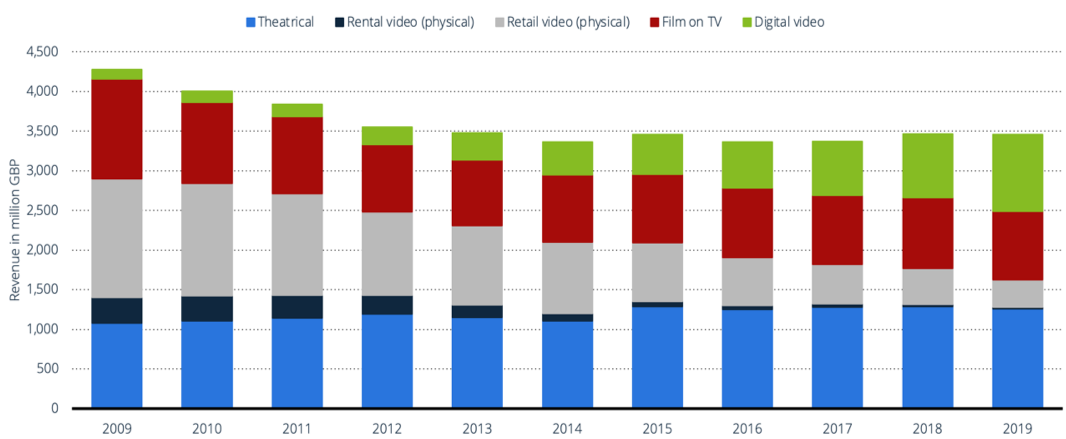


Figure 6. Film industry revenue in the United Kingdom (UK) from 2009 to 2019, by platform (British Film Institute, 2021b)

During the outbreak of COVID-19, the UK has severely restricted the movement of its citizens and even blocked streets, which has dealt a huge blow to the film industry (Edington, 2022). While the way to rent or buy movies from retailers has been abandoned, theater or cinema viewings are the first choices for moviegoers, and more and more people prefer to watch films on digital video and TV. Meanwhile, as shown in Figure 4 and Figure 7, the epidemic has affected the revenues of the film industry. Although the number of film productions has continued to increase, the profits have stalled in recent years.

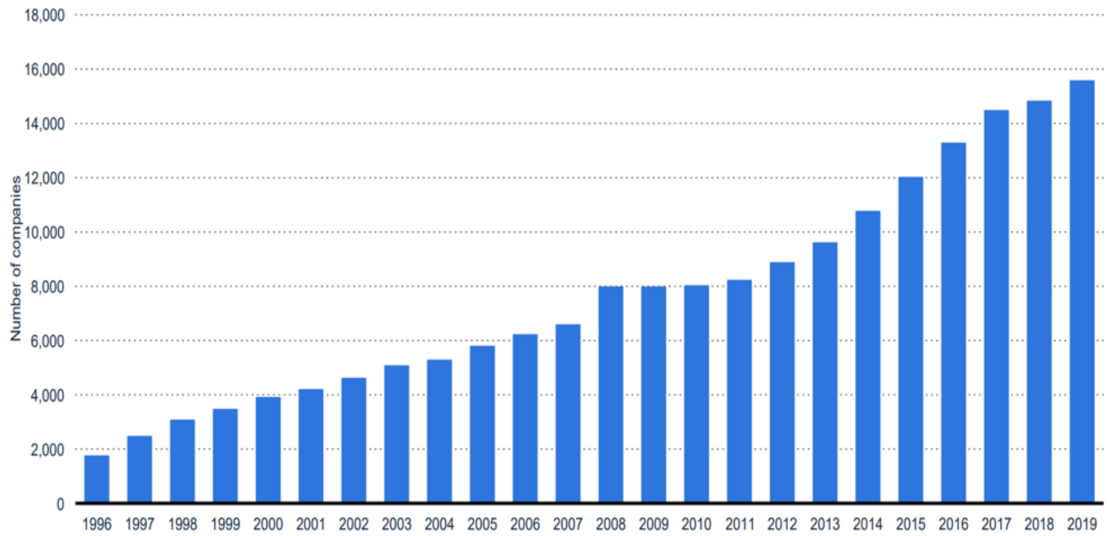


Figure 7. The number of film and video production companies in the United Kingdom from 1996 to 2019 (British Film Institute, 2021c)

Although films can be distributed all over the world, since the beginning of the 1920s, the global film industry has been monopolized by major studios such as Hollywood (Gomery, 2015). Coincidentally, the British film industry faces the same situation. According to British Film Institute (2021c), the number of film production companies in the UK has been increasing since 1996, reaching nearly 16,000 in 2019.

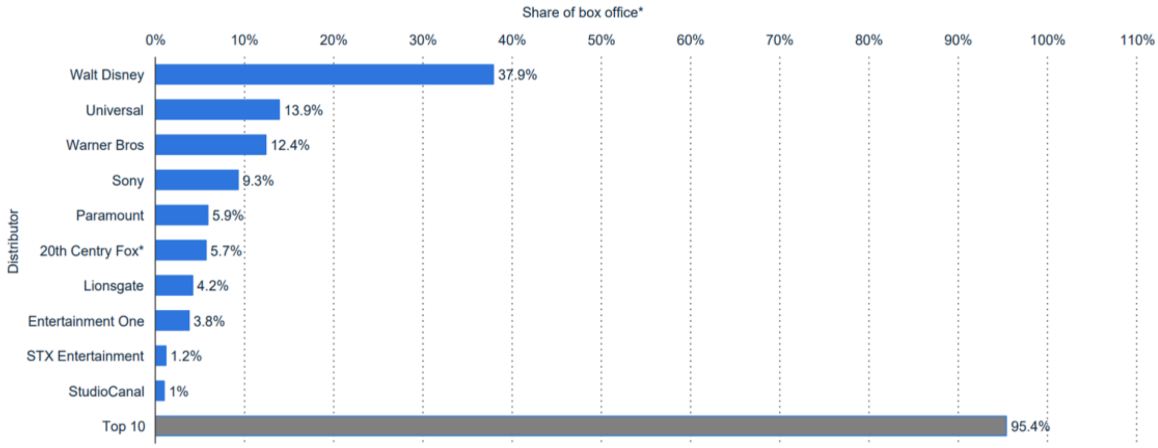


Figure 8. Film distributors ranked by market share in the United Kingdom (UK) and Republic of Ireland in 2019, based on box office gross (British Film Institute, 2020)

2.1.2 The Impact of Social Media and Movie Reviews on the Film Industry

As mentioned in Section 2.1.1, based on the change in the structure of the film industry, social media has become a crucial platform for movie releases. On the other conditions, to attract traffic and consumers' attention to movies, social media is also a significant channel for collecting customer opinions (Chen & Yin, 2017; JustWatch, 2021). In 2021, the number of British social network users has reached 60 million, and netizens express various views, comments, and experiences about products on social media (Duan *et al.*, 2022; Statista, 2021)

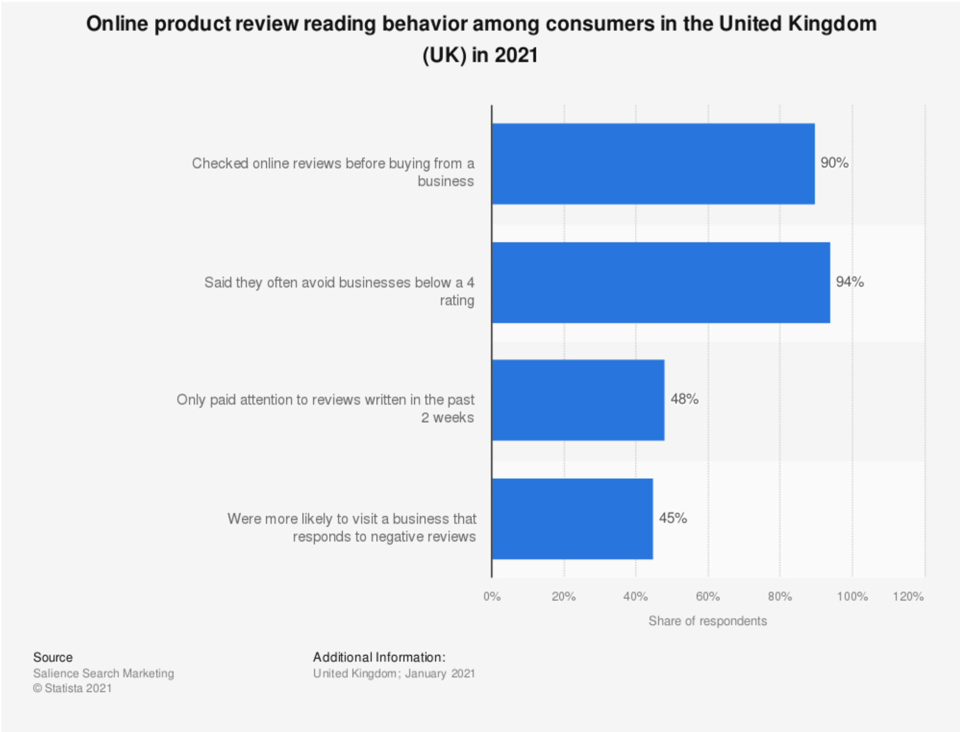


Figure 9. Online product review in the United Kingdom (Salience Search Marketing, 2021)

Basically, social media sites have become the medium through which studios communicate directly with fans (Bilal *et al.*, 2016; Duan *et al.*, 2022). Via reviews, studios can understand the intent of consumers watching movies, and the sales of a film can be predicted based on factors such as the genre, plot, and actors of the movie (Ahmad *et al.*, 2020; Dhar & Chang, 2009). From the perspective of customers, movie reviews can provide them with information before viewing, including plot, cast, and overall layout. According to the research of Pentheny (2015), audiences not only pay attention to film critics' comments on movies, but other consumers' after-viewing can also influence audience's choices, that is, reviews can reduce the uncertainty of movies and make consumers more aware of relevant information, thereby enhancing consumers' willingness to watch (Duan *et al.*, 2022; Park & Kim, 2008).

2.2 Sentiment Analysis

Due to the emergence of large amounts of text data, natural language processing (NLP) has received high attention in various fields, which is the ability of computer programs to understand human language through text analysis (Sudhir & Suresh, 2021). Sentiment analysis is a branch of NLP and a field that is highly integrated with affective computing research (Birjali *et al.*, 2021).

2.2.1 The Development and Functionality of Sentiment Analysis

The most frequent method for assessing data in text form is sentiment analysis (opinion mining), which primarily determines the degree of user attention to an event by examining the emotional inclinations indicated by the text content (Pang & Lee, 2008; Vanaja & Belwal, 2018). The evolution of sentiment analysis may be divided broadly into three stages. Textual sentiment classification was investigated in the late 1990s by Hearst (1992) and Kessler *et al.* (1997), who classified it into three categories: positive, negative, and neutral. It is one of the simplest and most widely used ways currently, and most e-commerce sites use a 1-5 rating system to evaluate the client feedback (Sudhir & Suresh, 2021). At the turn of the twentieth century, sentiment analysis drew the attention of additional academics, the sentiment categorization was expanded, allowing emotions like happiness, sadness, anger, and other properties can be retrieved from texts, which also is called sentiment detection (Boiy & Moens, 2009; Huber *et al.*, 2000; Pang & Lee, 2008). This analysis, based on ratings, gives companies a better understanding of users' specific feelings, but sentiment detection is imprecise due to the variability of emotions and the artistry of discourse. In recent years, with the rapid development of computer science, sentiment analysis techniques have been integrated into various fields. Emerging sentiment analysis can be explained into two categories: aspect-based and intent-based (Sudhir & Suresh, 2021).

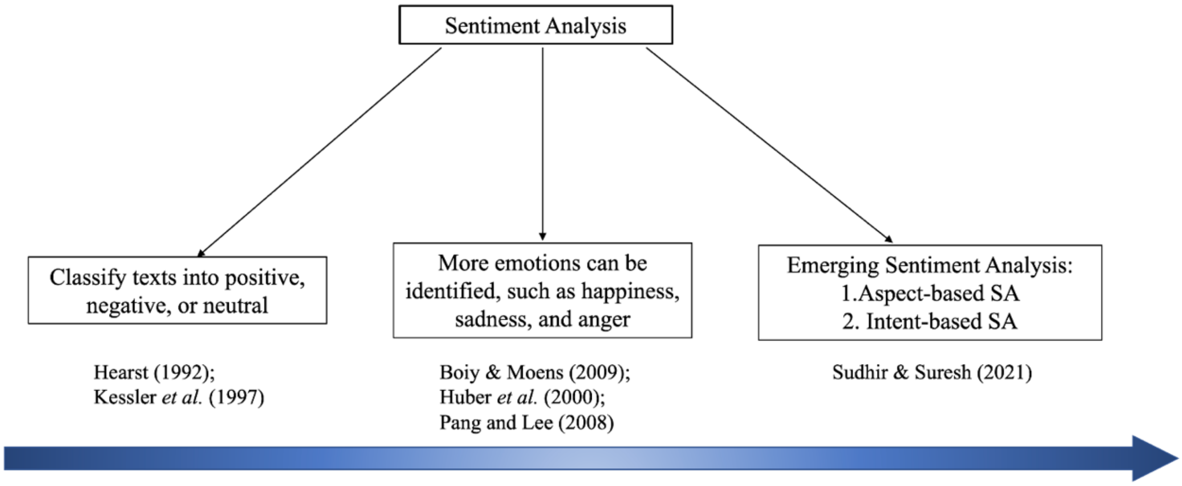


Figure 10. The development of sentiment analysis

Aspect-based sentiment analysis enables businesses to automatically identify and evaluate customer data, synchronously, categorize items based on client attitudes and solve problems, which greatly saves time and cost. Additionally, the objective of intent-based sentiment analysis is to dissect, evaluate, and classify text depending on the customer's intent, which makes companies more personable and keeps the customer relationship intact (Abbasi *et al.*, 2008; Ceron *et al.*, 2015; Sudhir & Suresh, 2021; Wang *et al.*, 2014).

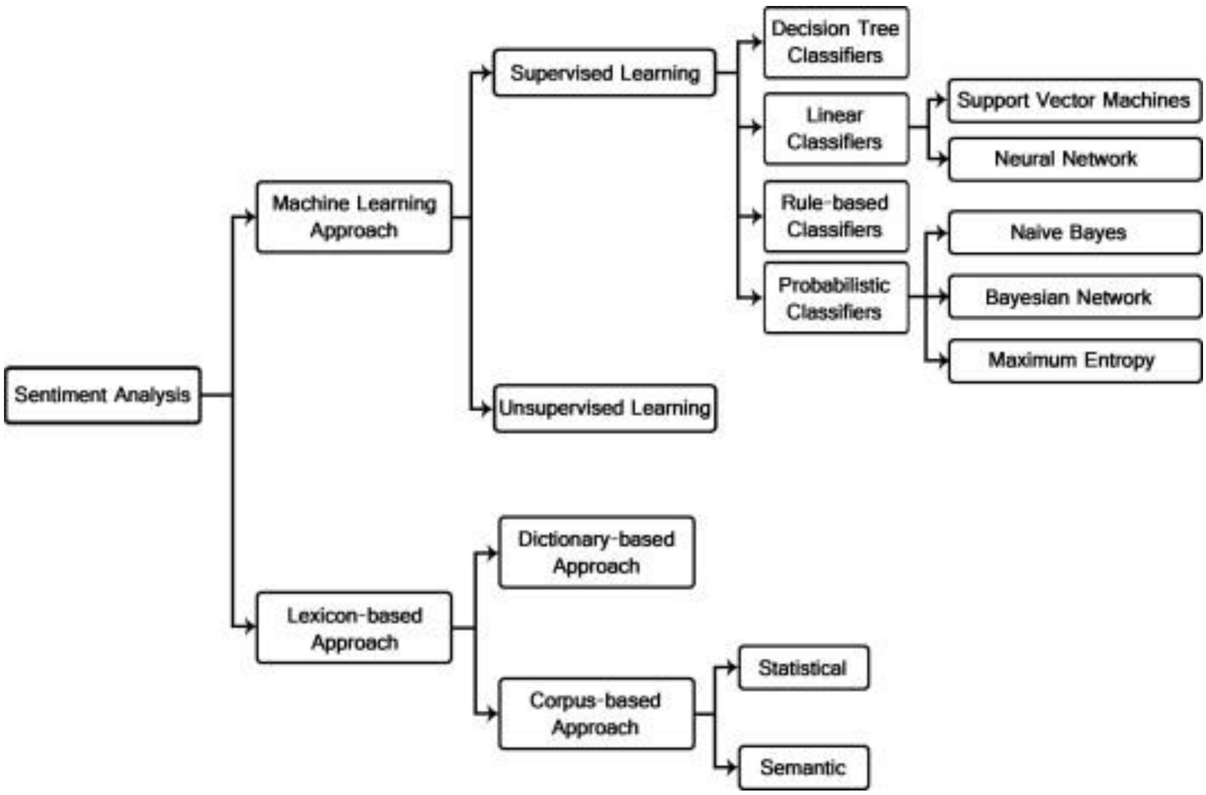


Figure 11. The categories of SA (Medhat *et al.*, 2014)

2.2.2 The Applications of Sentiment Analysis

Sentiment analysis is being applied in numerous fields as the technology improves and expands (Jain *et al.*, 2021). Analyzing textual data in e-commerce can assist to determine customer expectations for products, improve user experience, and boost sales (Vanaja & Belwal, 2018). A classic application of sentiment analysis in e-commerce is that Karthik *et al.* (2018) proposed a feature based product ranking and recommendation algorithm (FBPRRA), and the workflow is shown in Figure 12.

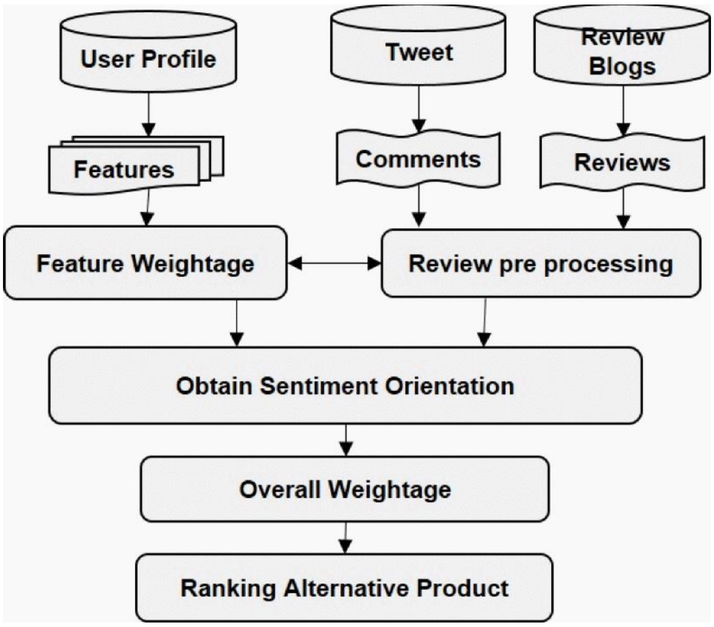


Figure 12. The workflow of FBPRRA (Karthik *et al.*, 2018)

FBPRRA computes the scores of customer satisfaction of each product by analyzing online reviews and applies the scores to the recommendation system. As multiple platforms integrate datasets, FBPRRA can evaluate the quality of a product in various aspects. However, the large amount of data also introduces useless feature information, which leads to the degradation of the accuracy of sentiment analysis and the drawback of long running time.

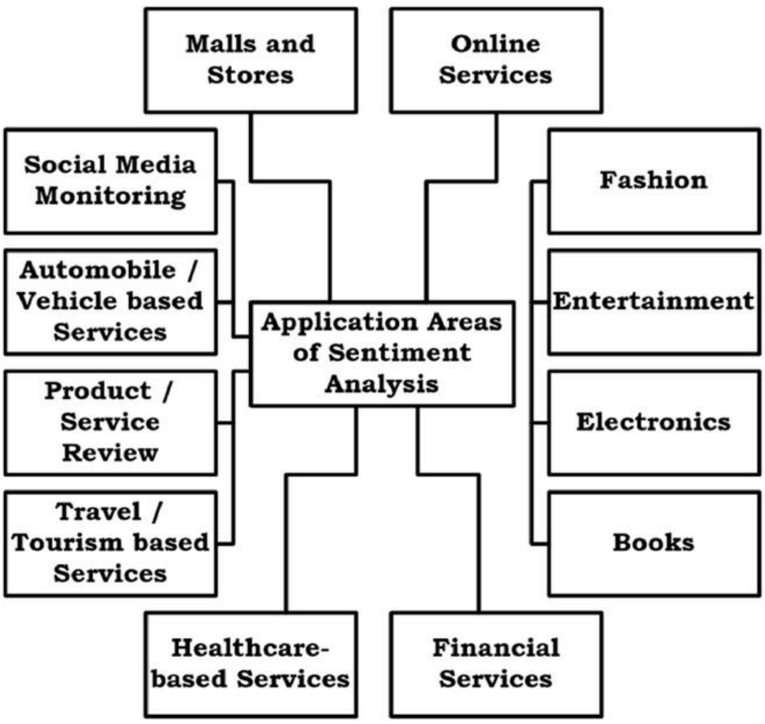


Figure 13. The current applications of sentiment analysis (Nasreen Taj & Girisha, 2021)

As shown in Figure 13, since sentiment analysis is oriented to individuals, events and topics, it can greatly help companies or organizations understand consumers' psychological activities (Medhat *et al.*, 2014). With the in-depth development of natural language processing, sentiment analysis can play an irreplaceable role in more fields

(Sudhir & Suresh, 2021).

2.2.3 General Process of Sentiment Analysis

Figure 14 describes the general process of sentiment analysis, which can be divided into four stages, which include data collection (extraction), data processing, sentiment classification, and output presentation (Birjali *et al.*, 2021).

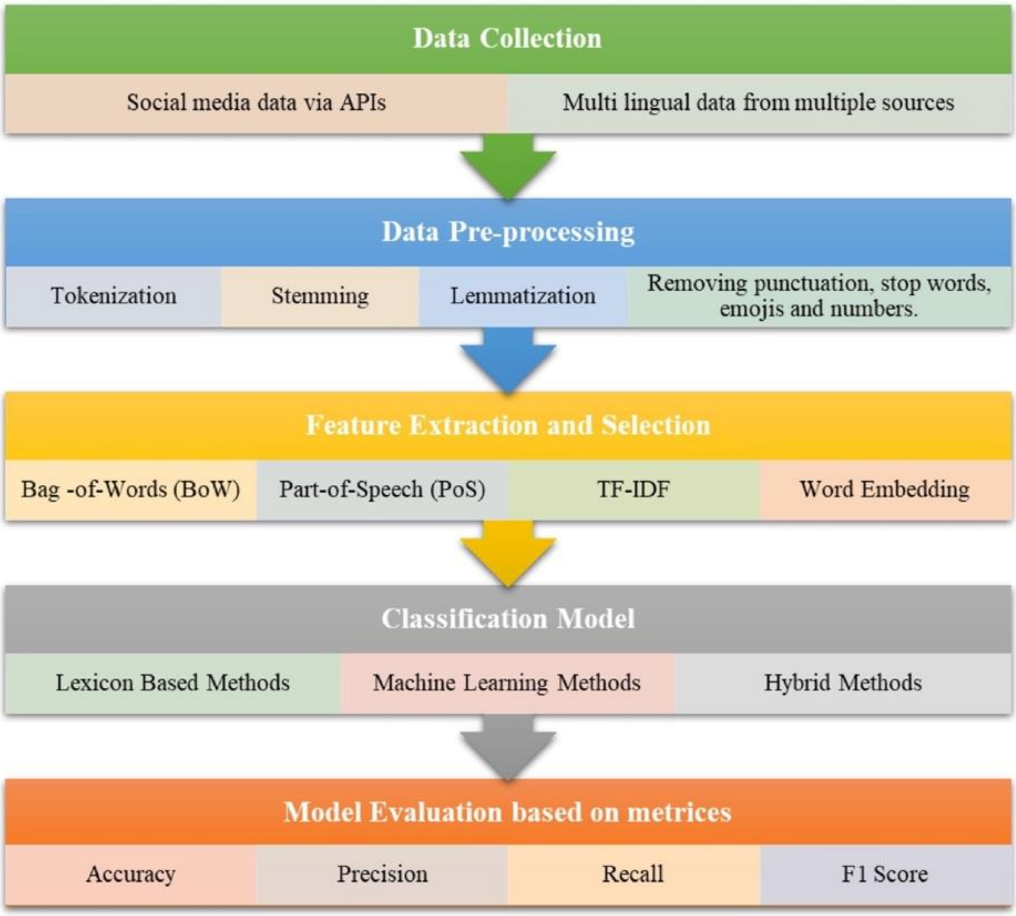


Figure 14. The process of sentiment analysis (Jindal & Aron, 2021)

Before discussing the process of the sentiment analysis, Table 1 expound the common toolkits for NLP tasks in different programming languages. Since the project focus on using Python to solve the research question, *TextBlob* and *NLTK* will be introduced in this part.

Table 1. Common toolkits for text preprocessing and NLP tasks (Bird *et al.*, 2009; Kwartler, 2017; Manning *et al.*, 2014; Pasha *et al.*, 2014)

Toolkit	Language	Description
TextBlob	Python	TextBlob is a Python library that provides a consistent API for diving into common natural language processing (NLP) tasks such as PoS tagging and sentiment analysis.
NLTK	Python	Natural Language Toolkit is a suite of open-source Python modules that help to perform NLP tasks such as tokenization and PoS tagging.
CoreNLP	Java	Stanford CoreNLP is a framework for basic and advanced NLP tasks as well as sentiment analysis
MADAMIRA	Java	MADAMIRA is a tool that performs Arabic NLP tasks like morphological analysis and tokenization.

NLTK is a platform which includes more than 50 corpora and lexical resources such as WordNet (NLTK, 2022).

The platform concentrates on constructing programs in Python by processing data in natural language. Specifically, *NLTK* offers a set of text processing packages for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, all of which are accessible via a Python-friendly interface (NLTK, 2022). Meanwhile, *TextBlob* has similar attributes and properties with *NLTK*, which is an external library to assist to study the natural language process (TextBlob, 2022). A straightforward API is supported in *TextBlob* for doing standard natural language processing (NLP) activities like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, etc.

2.2.3.1 Data Collection

Data collection refers to the creation of datasets by extracting from the collected data which frequently social media, review sites, blogs, and forums (Birjali *et al.*, 2021; Mei & Frank, 2015). As mentioned in Section 2.2.2, professional movie review sites are the best place to collect text data. Nevertheless, it is difficult for public datasets to fully match the needs of the project. Instead, data usually needs to be extracted from web resources.

2.3.3.2 Data Processing

Data processing consists of three parts: text preprocessing, feature extraction and feature selection.

i) Text Preprocessing

The objectives of text reprocessing are for cleaning noises and text data with incorrect grammar/spellings and dimensionality reduction (Birjali *et al.*, 2021; Liu, 2012). Specifically, considering that data obtained from social media may be unstructured, data preprocessing aims to address irrelevant features, noise, and erroneous text in the original data before analysis (Birjali *et al.*, 2021; Liu, 2012). Based on the research of Agrawal (2021); Pinto *et al.* (2016); Weng (2019), common text data preprocessing includes removing punctuations and stop words, computing tokenization, normalization and lemmatization.

ii) Feature Extraction

The term 'feature' in computer science refers to the attributes of the data, and in sentiment analysis, the purpose of feature extraction is to extract important features that describe the text which is unstructured data (Avinash & Sivasankar, 2019). Generally, the basically important features in sentiment analysis include negations, opinion words, part of speech of words, and the frequency of words (Birjali *et al.*, 2021). Negative words, also known as opinion-shifting words, may be keywords in the text that reverse the polarity of sentiment such as never, not, none and so on (Aggarwal, 2015). Opinion words mainly include some adjectives and adverbs, which express emotions subjectively and obviously in the text (Aggarwal, 2015).

There are two significant models which can accomplish the assignment, that are BoW (Bag of Words) and WE (Word Embedding). BoW is one of the most commonly used techniques for converting textual data (unstructured) into numerical representations (structured) (Kasri *et al.*, 2019; Siddharth, 2021). Figure 15 visualizes the inputs and outputs by using the model.

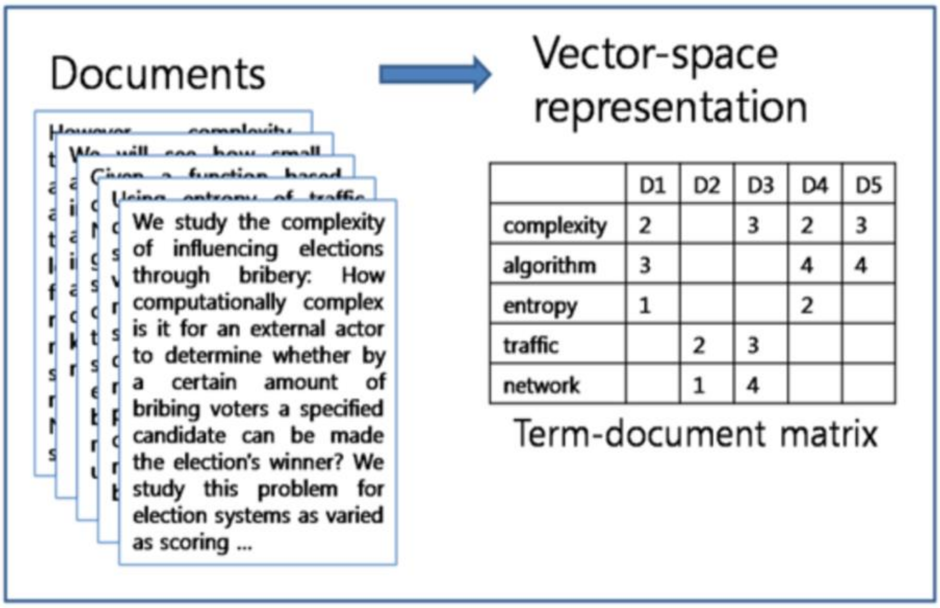


Figure 15. The visualization of BoW (Siddharth, 2021)

Compared with the BoW model, word embeddings convert the unstructured data into vectors and distribute all the converted structured data along the vectors (Birjali *et al.*, 2021). Specifically, word embedding is a method for representing words and documents, and its main goal is to reduce the dimension of data features, predict the meaning of words, and capture inter-word semantics (GeeksforGeeks, 2020).

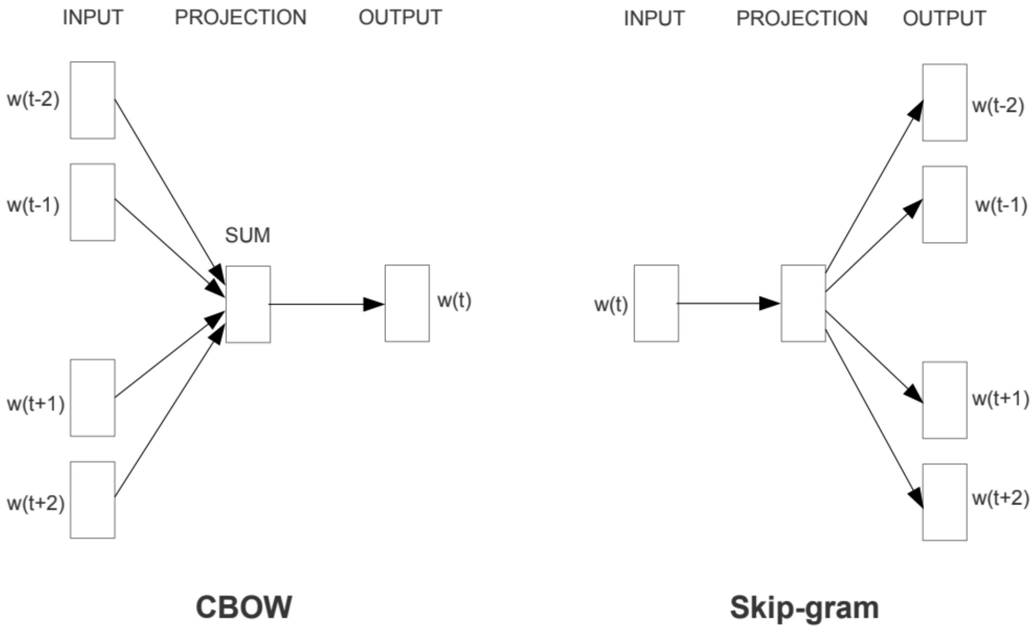


Figure 16. The processes of CBOW and Skip-gram (Mikolov *et al.*, 2013b)

iii) Feature Selection

The last step of data processing is the feature selection, which is also influenced by feature extraction. Feature selection is to filter out relevant features, so that the data dimension is reduced again, which helps to improve the accuracy and efficiency of the sentiment analysis (Ahmad *et al.*, 2019; Birjali *et al.*, 2021).

Fundamentally, there are two broad methods to select features, that is lexicon-based methods and statistical methods (Kumar & Kaur, 2020). Lexicon-based methods refer to the manual creation of feature sets, i.e., starting with strongly and clearly expressed emotional terms to form sets, and then enriching sets with synonyms or online resources (Baccianella et al., 2010). Although the features selected by this method can perform perfect in sentiment analysis models, as the amount of data increases, it is almost difficult to manually create a dictionary set covering all keywords. Regardless, the statistical method automatically selects features with a little lower accuracy, however, it can save much time and has more efficient in feature selection (Duric & Song, 2012).

Figure 17. The workflow of the filter method (sauravkaushik8, 2016)

Figure 17 illustrates the workflow of the filter method, which is the most common in the feature selection (Kou *et al.*, 2020). In this method, feature selection is completely independent of any machine learning algorithm, instead it picks the highest priority features by counting and measuring the ranking of features in the training data to generate the crucial features (Birjali *et al.*, 2017). Based on the research of Sánchez-Marono *et al.* (2007) and Hoque *et al.* (2014), due to the low cost of the filter method, it is suitable for datasets with numerous features, and the priority of features can be ranked using statistical methods such as Pearson coefficient, ANOVA and chi-square test.

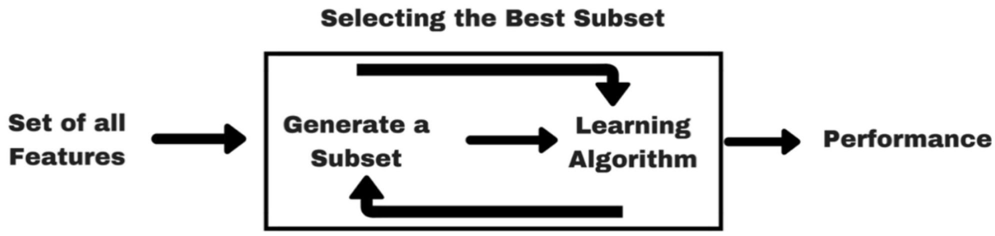


Figure 18. The workflow of the wrapper method (sauravkaushik8, 2016)

As shown in Figure 18, another method is called wrapper, which finds the optimal features based on the training process, therefore, it depends on the machine learning (Birjali *et al.*, 2021).the basic process of wrapper method has two directions: select features forward or remove features backward. The former starts with an empty set and continuously adds features that can improve the model in iterations, while the latter lists all features in a set and repeatedly eliminates features to optimize in iterations (Urbanowicz *et al.*, 2018).

2.2.3.3 Sentiment Classification

Most importantly, sentiment classification refers to establishing an effective model for sentiment analysis, and all algorithms which are shown in Figure 11 are achievable models. In this section, the three classic machine learning models will be introduced.

i) Support Vector Machine (SVM)

As the most classic linear method in supervised learning, SVM is a non-probabilistic classifier for classifying discrete and continuous variables (Birjali *et al.*, 2021). In sentiment analysis, SVM separates text data by finding the optimal hyperplane, and the effective separation is manifested in the largest margin between the hyperplane and the nearest training point of any class (Hastie *et al.*, 2009; Joachims, 1998). Figure 19 depicts the geometric principles of SVM, where (a) describes the linear separation and (b) shows the nonlinear separation.

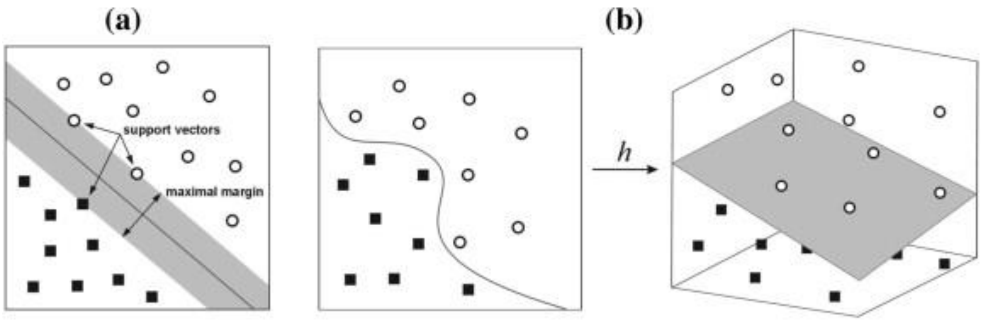


Figure 19. The geometric principles of SVM (Moraes *et al.*, 2013)

Since the SVM model is not only uncomplicated to train, but also efficient and stable in high-dimensional spaces, it is an algorithmic choice worth considering for simple the sentiment analysis (Birjali *et al.*, 2021). Rana and Singh (2016) have used SVM and Naive Bayes to analyze movie reviews (positive or negative) and concluded that linear SVM produces the best accuracy.

ii) Decision Tree

A decision tree is a flow-chart-like tree structure, which can effectively handle non-linear classification datasets. (Sudhir & Suresh, 2021). Figure 20 shows the structure of the decision tree, and the essential elements of the decision tree include a root node, decision nodes, and leaf nodes, where the internal nodes represent the tests for attributes, each branch reflects the results of the tests, and the leaf nodes represent the output (class) (Ashari *et al.*, 2013; Jacobs *et al.*, 2015).

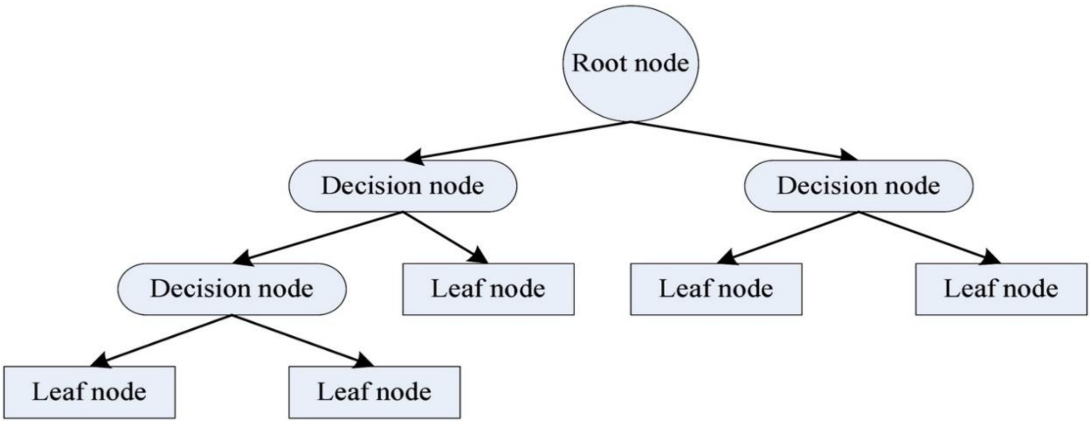


Figure 20. The structure of the decision tree

iii) Recurrent Neural Networks (RNN)

RNN is a model in deep learning, which is implemented based on an artificial neural network (Birjali *et al.*, 2021). RNN forms a continuous memory network by cyclically linking artificial neurons, as shown in Figure 21. Unlike other neural networks, all inputs and outputs in RNN are interconnected, hence, this model is extremely sensitive to time series, and the ability to capture and store long sequences of information is widely used in the sentiment analysis (Sharfuddin *et al.*, 2018).

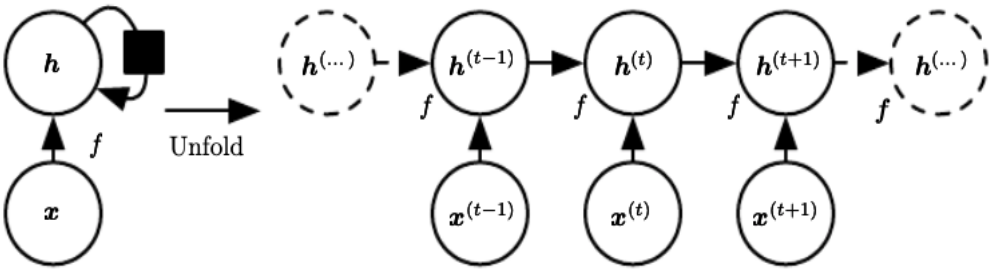


Figure 21. The basic loop of RNN (Sharfuddin *et al.*, 2018)

However, as the size of the input text data increases, the ability of the RNN to link information is greatly weakened (vanishing gradient) (Birjali *et al.*, 2021). Therefore, RNN has produced two variants: long-short term memory (LSTM) and gated recurrent unit (GRU). The basic principles of the three types of RNN are shown in Figure 22.

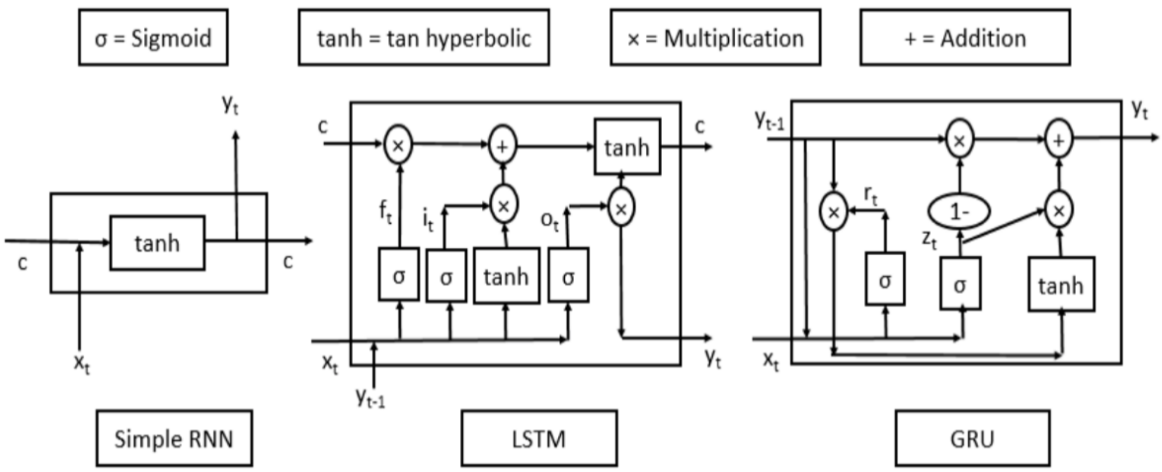


Figure 22. The basic principles of the three types of RNN (Khan *et al.*, 2019)

2.2.4 Evaluation of Sentiment Analysis Algorithms

As mentioned in Section 2.2.1, sentiment analysis can be accomplished by at least two types of techniques. Thus, it is impossible to simply use machine learning or the lexicon-based algorithm to evaluate the performance of an algorithm. However, Barr *et al.* (1995) proposed to evaluate the performance of an algorithm from the macro perspectives of quality of the solution, running time, robustness, flexibility, and simplicity, which is consistent with measuring the pros and cons of sentiment analysis algorithms.

Table 2. The meanings of accuracy, precision, recall, F1-score (Birjali *et al.*, 2021)

Metric	Description	Calculation	Assessment
Accuracy	Accuracy is the most commonly used metric for classification problems, and it represents the ratio between the correctly predicted examples to the total number of examples. The complement of this metric is called error and can be calculated as 1- accuracy.	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy is a good choice for sentiment classification in machine learning when the classes in the dataset are nearly balanced.
Precision	The precision represents the proportion of samples that correctly predicted as positive to the total number of predicted positive samples. In other terms, precision measures the quality of being exact.	$\frac{TP}{TP+FP}$	This metric is suitable for problems where prediction should be confident
F1-score	F1-score is a number between 0 and 1 that helps to measure both precision and recall by calculating the harmonic mean of these two metrics.	$\frac{2*Precision*Recall}{Precision+Recall}$	Because it is not easy to compare two classifiers with high recall and low precision or vice versa, the F1-score can be the key to this problem. F1-score manages the tradeoff for a model that requires a confident prediction and dominant capture of a class.

Specifically, because the words in the text can be deceptive, it is difficult for an algorithm to correctly evaluate all emotional tendencies so that the optimal solution is unavailable (Barr *et al.*, 1995; Chakraborty *et al.*, 2019). Regardless of the sentiment analysis model, it often draws on the evaluation methods of classification algorithms and uses accuracy, recall, precision, and F1 Score to measure the performance of an algorithm, and the meanings are shown in Figure 5 (Bhowmik *et al.*, 2022; Birjali *et al.*, 2021; Joseph *et al.*, 2022; Vashishtha & Susan, 2019). Meanwhile, as mentioned in Figure 14, there are four stages to constructing a sentiment analysis solution, and the different programming in each stage affect the length of the running time, which refers to the time interval from the beginning to the end of the program execution (Barr *et al.*, 1995). A significant algorithm needs to efficiently produce an ideal model in a limited time. Since sentiment analysis is used in different fields to complete a variety of assignments, robustness indicates that the model should not be too sensitive to different textual differences, that is, for any given instance, the specific performance of the algorithm and the quality of the solution should be stable at a standard (Bräysy & Gendreau, 2005; Medhat *et al.*, 2014). In different circumstances, simplicity states that the algorithm should be easy to implement, and the model can be constructed in inefficient time, while flexibility is related to robustness.

3. Methodology

3.1 Research Question and Objectives

As mentioned in Section 1.2, the research question is "What factors determine the success of action films?" The reason for selecting the action movies is that considering the variable measurement criteria should be supplied for different film genres, it is arduous to evaluate all types of films together. To draw forth the research question in detail, the movie genres that are most popular with audiences in the UK are categorized, as shown in Figure 23.

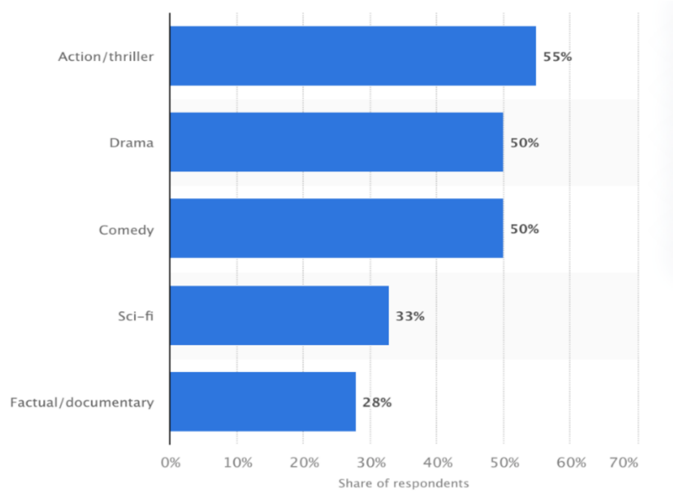


Figure 23. Five most popular film genres in the UK as of October 2013 (Statista, 2014)

The research manifestly shows that in 2013, action and thriller films were the two most popular movie genres for British audiences and the action film is selected as the research object. To address the research question, the four aspects of a film are specified as the factors based on the research of Wang *et al.* (2020a): Actor, Plot, Director, and Music. Therefore, the comprehensive research question can be defined as "Among the four factors of director, plot, actors and music, which factor determines the success of an action movie?"

3.2 System Structure

For discussing the process of methodology in detail, the Section 3.3 describes the general structure of Chapter 3. Following the research question, the dissertation is split into the following 6 research objectives, as shown in Figure 24.

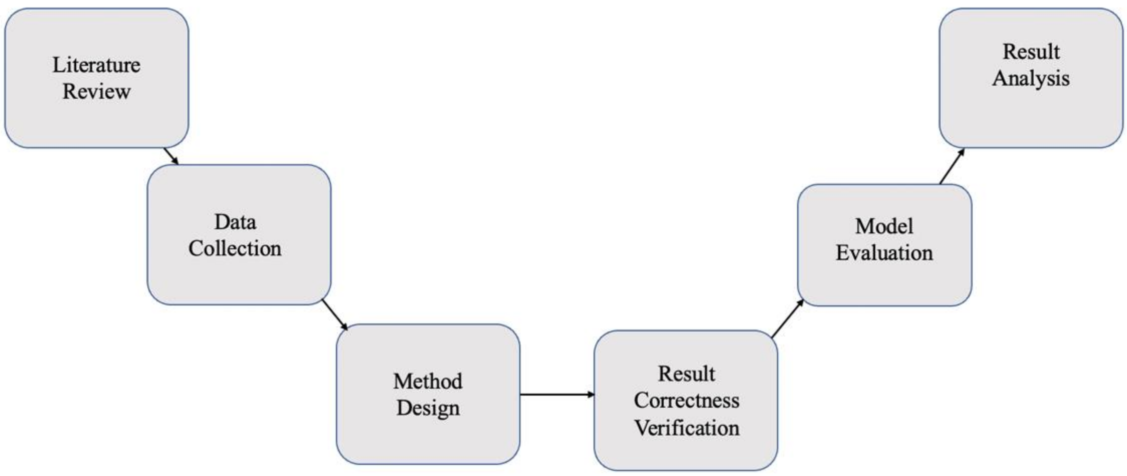


Figure 24. The 6 research objectives

The objective 1 has been done in Chapter 2 and in following sections each objective will be discussed respectively. Specifically, Figure 25 depicts the workflow of the methodology.

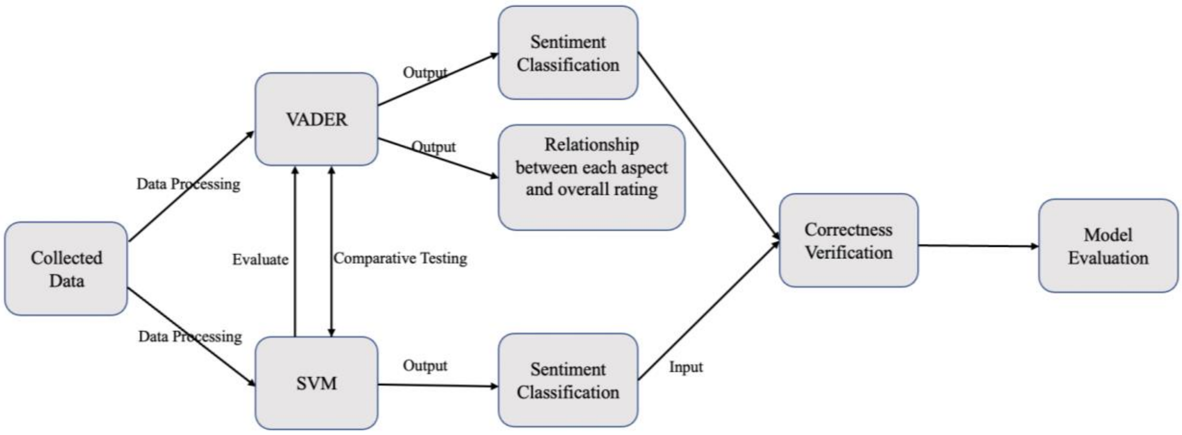


Figure 25. The workflow of the methodology

Section 3.3 will introduce the entire data collection from the research method (qualitative), the type of data required, and the data collection method. Compared with obtaining data through crawlers on the Internet, researchers prefer to collect data from Kaggle, which is a community for data science (Kaggle, 2022). When completing collecting data, the collected data will be subjected to text preprocessing, feature extraction and feature selection, and then transmitted as input to the sentiment analysis model. For addressing the research question significantly, there are two sentiment analysis model constructed. The first one is a lexicon-based model called VADER (Valence Aware Dictionary for Sentiment Reasoning), which evaluates the sentiment by scoring each word in text, in which the relation between the overall ratings and each aspect can be assessed (Hutto & Gilbert, 2014). The second model is SVM, which has introduced in Section 2.1.3 c. and aims for evaluating the performance of VADER and a comparative testing between both models is computed.

3.3 Data Collection

Before collecting data, the types of research methods should be considered. Esteban-Bravo and Vidal-Sanz (2021) conducted that the quantitative research, which are shown in Figure 26, design for collecting numerical or structured data by using statistical inference tools and its objective is to draw the conclusive research using representative samples of the population. On the other hand, the definition of the qualitative research refers to the interpretation of data from preliminarily set theories or hypotheses by collecting weakly structured or unstructured data from a social system, and Figure 27 indicates the main categorized qualitative research methods. Basically, the biggest difference between the two kinds of research is the type and method of data collected.

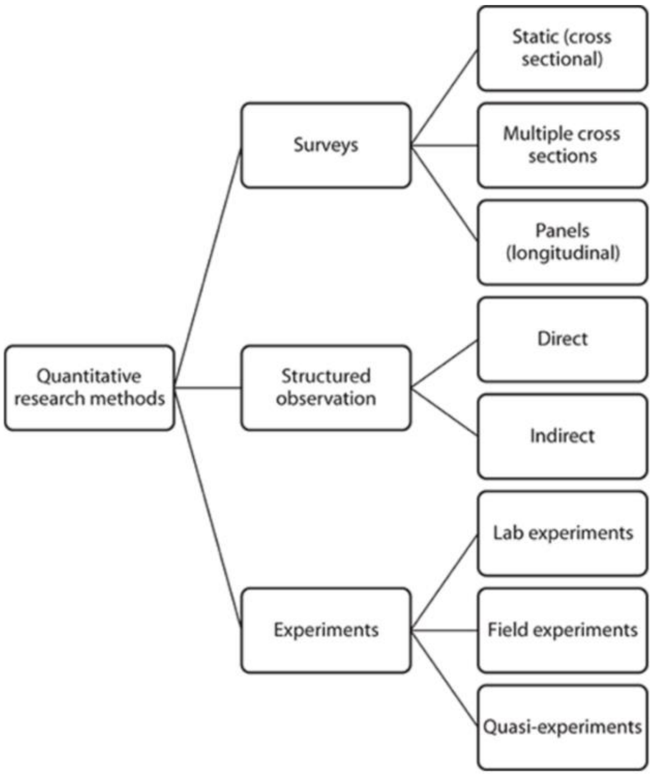


Figure 26. The main quantitative research methods (Esteban-Bravo & Vidal-Sanz, 2021)

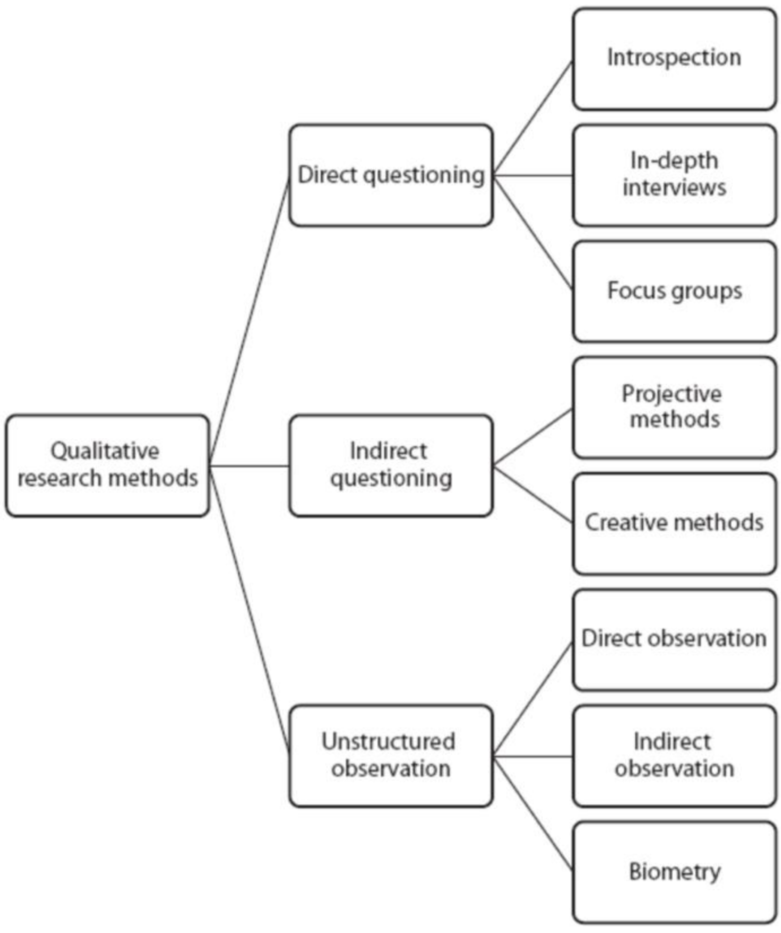


Figure 27. The main categorized qualitative research methods (Esteban-Bravo & Vidal-Sanz, 2021)

In this project, the qualitative research method is used for collecting data, and the reasons are as following.

First of all, Chapter 2 has concluded that the movie reviews on social media are the best medium for evaluating the performance of a movie. Hence, the structure of the dataset is the text data instead of the numerical data, which indicate the quantitative search method is not suitable for this project. The second point is that the evaluating method of the project is the sentiment analysis, which analyzes the text data typically through non-statistical methods, and the process of the method is interpreting text, which match the definition of the qualitative research (Esteban-Bravo & Vidal-Sanz, 2021). Additionally, referring to Leavy (2014), the study of published reports, web pages, books, and other forms of texts falls under content analysis, which is a method of qualitative analysis (Berelson, 1952; Goode & Hatt, 1952). Obviously, the content analysis under this definition includes sentiment analysis by using data in text format. Moreover, many scholars have explicitly mentioned the sentiment analysis as a qualitative research method in their research. For instance, in the research of Gascón *et al.* (2016), the sentiment analysis was regarded as a qualitative method. Based on the three reasons above, it can evidently conduct that the search method of the project is qualitative search and analysis.

ParseHub is selected to web scraping for the project, which is a free web scraping tool with an easy-to-use and intuitive user experience. For any website, ParseHub can scrape structured data on the website without coding (Demchenko, 2020). Following the tutorial of ParseHub, there are several basic steps to scrape a website as shown in Figure 28, and Figure 29 depicts the interactive page of scraping data using ParseHub.

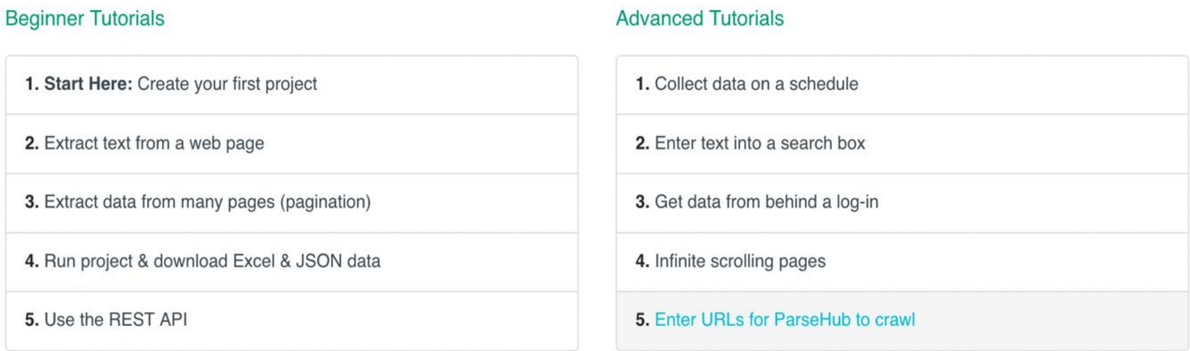


Figure 28. The steps of scraping website (Perez, 2019)

1) Create a new project and input URL of the website which will be scrapped.
2) Browse the web and manually select the data that needs to be scrapped.
3) Save selected data in variables and correlate the reciprocal data.
4) If the needed data exists on several pages, it can be added associated pages.

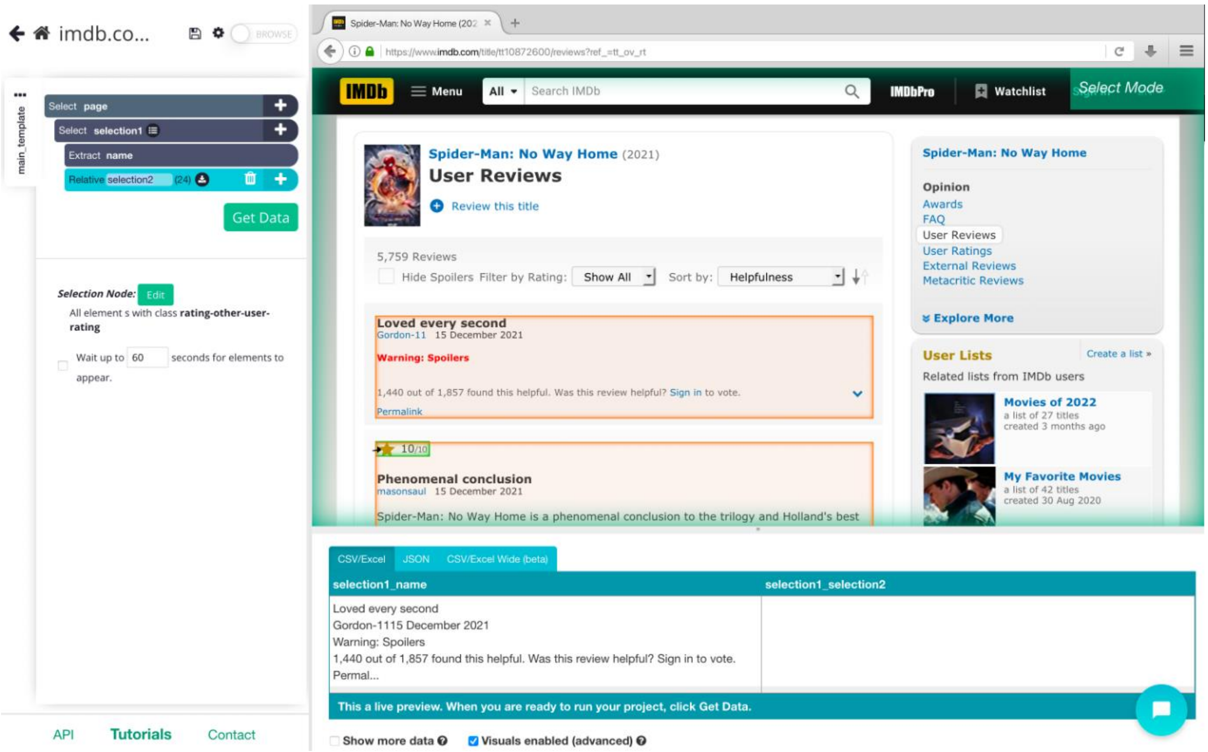


Figure 29. Web scraping for reviews of Spider Man by using ParseHub

When the data is collected, the features that should be included are shown in Table 3.

Table 3. The data template

Data Template											
Review ID	Movie ID	Title	Category	Director	Published Data	Movie Length	Review Title	Content	Rating	Sentiment Classifier	URL
1											
2											
3											
4											
5											
6											

Data collection will be subject to the ethical permission of the university of Warwick, and the data sources do not involve participants, instead, the data should be from public social medias (IMDB) or datasets. Basically, a total of 1,000 data is screened in the project for the following reasons: (1) Too many movie review texts will extremely diminish the analysis rate. (2) If the sample size is small, the training accuracy of the SVM model used in the project will be reduced, thereby affecting the quality of the model. (3) Based on Table 3, the data set used in the project contains many columns, so each piece of data has high dimension features. Therefore, compared with ordinary datasets, the analysis rate of this data set will be relatively lower. For the SVM model, 1000 data can improve the efficiency of the model as much as possible on the premise that an excellent model can be trained. For the VADER model, 1000 data can analyze enough information to support discussion throughout the project.

3.4 Method Design

3.4.1 Text Preprocessing

Based on Section 2.1.3 (i), the contents of text preprocessing include removing punctuations and stop words, computing tokenization, normalization and lemmatization.

3.4.1.1 Removing Punctuations and Stop Words

To remove punctuations, the *string* library in Python has defined a list of punctuations, which includes the basic symbols and can be called directly (Deepanshi, 2021). Figure 30 shows the process of punctuations removal.

v2	clean_msg
Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...	Go until jurong point crazy Available only in bugis n great world la e buffet Cine there got amore wat
Ok lar... Joking wif u oni...	Ok lar Joking wif u oni
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005 Text FA to 87121 to receive entry questionstd txt rateTCs apply 08452810075over18s
U dun say so early hor... U c already then say...	U dun say so early hor U c already then say
Nah I don't think he goes to usf, he lives around here though	Nah I dont think he goes to usf he lives around here though

Figure 30. The punctuations removal example (Deepanshi, 2021)

For the removal of stop words, *NLTK* (Natural Language Toolkit) should be run (Bird *et al.*, 2009). *NLTK* contains about 180 stop words and can add words as needed. Figure 31 depicts an example of *NLTK* removing stop words (Agrawal, 2021).

msg_tokenied	no_stopwords
[go, until, jurong, point, crazy, vailable, only, in, bugis, n, great, world, la, e, buffet, cine, there, got, amore, wat]	[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]
[ok, lar, joking, wif, u, oni]	[ok, lar, joking, wif, u, oni]
[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, nay, 2005, text, fa, to, 87121, to, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]
[u, dun, say, so, early, hor, u, c, already, then, say]	[u, dun, say, early, hor, u, c, already, say]
[nah, i, dont, think, he, goes, to, usf, he, lives, around, here, though]	[nah, dont, think, goes, usf, lives, around, though]

Figure 31. Using *NLTK* to delete the stop words (Deepanshi, 2021)

3.4.1.2 Tokenization

To accomplish the assignment of tokenization, the function *word_tokenize* in *NLTK* toolkit should be invoked. Figure 32 states the process of tokenization by using *NLTK*. Each sentence is split into words (smaller units), can all words of a sentence store as the array form, which can be indexed more efficiently (Johnson, 2022).

msg_lower	msg_tokenied
go until jurong point crazy available only in bugis n great world la e buffet cine there got amore wat	[go, until, jurong, point, crazy, available, only, in, bugis, n, great, world, la, e, buffet, cine, there, got, amore, wat]
ok lar joking wif u oni	[ok, lar, joking, wif, u, oni]
free entry in 2 a wkly comp to win fa cup final tkts 21st may 2005 text fa to 87121 to receive entry questionstd txt ratetcs apply 08452810075over18s	[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, to, 87121, to, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]
u dun say so early hor u c already then say	[u, dun, say, so, early, hor, u, c, already, then, say]
nah i dont think he goes to usf he lives around here though	[nah, i, dont, think, he, goes, to, usf, he, lives, around, here, though]

Figure 32. Computing tokenization by *NLTK* (Deepanshi, 2021)

3.4.1.3 Normalization

Figure 33 computes an example for normalizing the text data in Python.

clean_msg	msg_lower
Go until jurong point crazy Available only in bugis n great world la e buffet Cine there got amore wat	go until jurong point crazy available only in bugis n great world la e buffet cine there got amore wat
Ok lar Joking wif u oni	ok lar joking wif u oni
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005 Text FA to 87121 to receive entry questionstd txt rateTCs apply 08452810075over18s	free entry in 2 a wkly comp to win fa cup final tkts 21st may 2005 text fa to 87121 to receive entry questionstd txt ratetcs apply 08452810075over18s
U dun say so early hor U c already then say	u dun say so early hor u c already then say
Nah I dont think he goes to usf he lives around here though	nah i dont think he goes to usf he lives around here though

Figure 33. Lowering the text (Deepanshi, 2021)

Generally, all letters in the data should be converted into lowercase, which can be simply achieved by *string.lower*() function. To convert the abbreviations into standard format, *contractions* library should be installed.

3.4.1.4 Lemmatization

lemmatization (stemming) is the last step for text preprocessing, which restoring synonyms of different structures in English to their basic forms (roots). Figure 34. shows using *NLTK* in Python to address lemmatization.

no_stopwords	msg_stemmed
[go, jurong, point, crazy, available, bugis, n, great, world, la, e, buffet, cine, got, amore, wat]	[go, jurong, point, crazi, avail, bugi, n, great, world, la, e, buffet, cine, got, amor, wat]
[ok, lar, joking, wif, u, oni]	[ok, lar, joke, wif, u, oni]
[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receive, entry, questionstd, txt, ratetcs, apply, 08452810075over18s]	[free, entri, 2, wkli, comp, win, fa, cup, final, tkt, 21st, may, 2005, text, fa, 87121, receiv, entri, questionstd, txt, ratetc, appli, 08452810075over18]
[u, dun, say, early, hor, u, c, already, say]	[u, dun, say, earli, hor, u, c, alreadi, say]
[nah, dont, think, goes, usf, lives, around, though]	[nah, dont, think, goe, usf, live, around, though]

Figure 34. Lemmatization in Python by using *NLTK* (Deepanshi, 2021)

3.4.2 Feature Extraction

Following Section 2.2.3 (ii), in the project the model of Bag of Words (BoW) will be constructed to extract and calculate the important features in each review. Following the tutorial of GeeksforGeeks (2019), there are three steps constructed the BoW model without using library in Python.

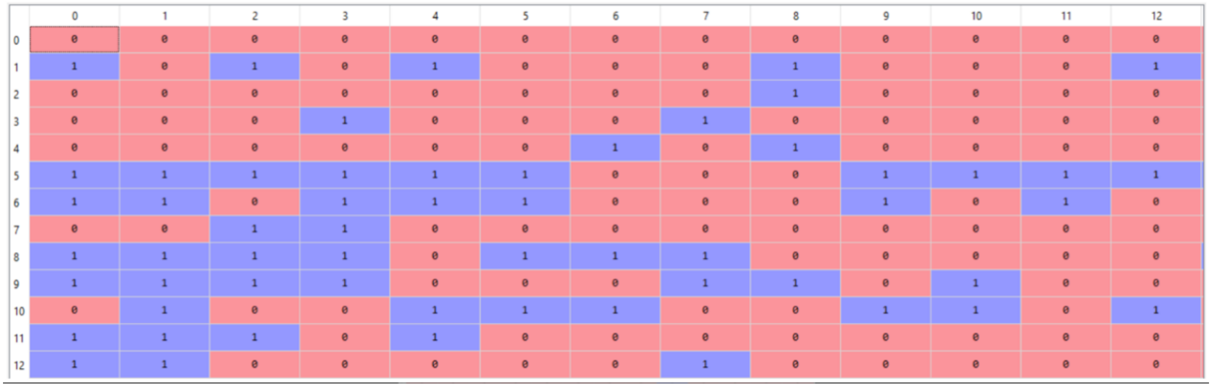
1) The first crucial progress is text reprocessing including lowering each letter, removing punctuation and special symbols such as ‘#’ and ‘&’, which has done in Section 3.4.1.

2) Capture the most frequent words in the text. Since all words have been normalized, which can be traversed to calculate the frequency of occurrence. A dictionary can be declared in the code to store the most frequent words in the dictionary. Figure 35 shows a result example after counting each word in the sentences, in which the term ‘value’ represents the number of occurrences.

Key	Type	Size	Value
a	int	1	2
addressing	int	1	1
again	int	1	1
aged	int	1	1
agricultural	int	1	1
also	int	1	1
am	int	1	1
amaury	int	1	1
and	int	1	7
are	int	1	1
as	int	1	1

Figure 35. The dictionary of BoW (GeeksforGeeks, 2019)

3) Compute the BoW model.



	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	0	0	1	0	0	0	0
3	0	0	0	1	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	1	0	1	0	0	0	0
5	1	1	1	1	1	1	0	0	0	1	1	1	1
6	1	1	0	1	1	1	0	0	0	1	0	1	0
7	0	0	1	1	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	1	1	1	0	0	0	0	0
9	1	1	1	1	0	0	0	1	1	0	1	0	0
10	0	1	0	0	1	1	1	0	0	1	1	0	1
11	1	1	1	0	1	0	0	0	0	0	0	0	0
12	1	1	0	0	0	0	0	1	0	0	0	0	0

Figure 36. An example of output of Bow model (GeeksforGeeks, 2019)

To generate more flexible models and improve run-time memory utilization efficiency, a vector is utilized to evaluate whether a word is crucial. Figure 36 states the result of BoW model, where '0' represents irrelevant features and '1' represents important features.

3.4.3 Feature Selection

In feature extraction, text data has been transformed into structured data which represents by features. Feature selection is to further retain the features with strong correlation, while removing irrelevant and redundant feature variables, hence the feature dimension space will be compressed, and the accuracy of sentiment classification will also be improved (Ahmad *et al.*, 2019; Birjali *et al.*, 2021).

Table 4. Lexicons of movie aspect and element-related terms (Wang *et al.*, 2020a)

Aspects	Movie aspects related words	Specific terms
Actors	Actors, actress, acting, cast, casting, role, performance	Actor names
Directors	Director, directed, filmed, edition	Director name
Plot	Plot, story, character, romance, revenge, ending, script	Script Writer name
Music	Lyrics, songs, ending song, background music, sound	Singer name

Table 5 depicts an example of the processed data which has been extracted the four-aspects crucial features.

Table 5. The processed data example (Wang *et al.*, 2020a)

Review	Movie	Genre	Director	Actors
I was really excited for this movie as Johnny Depp and Robert Pattinson take part in it, but it was trash. The story is so poor, the characters are so boring and I do not know why they made this film	Waiting for the Barbarians	Drama	Ciro Guerra	Mark Rylance, Jonny Depp, Robert Pattinso

The above three stages are the central steps of data processing, in which the text data has been converted into structured data, and all rest data can be directly input into the constructed model to output and address the research question.

3.4.4 Modelling

Modeling is the most important part of analyzing data. The processed data will be fed into the model, and the research problem will be solved through the training, evaluation or integration of the data by the models. In the project, the input is crucial features in the collected reviews. Based on the research question and the objectives, the output should be the intensity and orientation towards factors that is actors, directors, plot and music. For sentiment classification, two models will be construct.

3.4.4.1 VADER

Sentiment Intensity Analyzer in VADER , which is validated and based gold-standard lexicons (Hutto & Gilbert, 2014). Compared with traditional machine learning methods, VADER not only saves running time (no data is needed to train to construct the model), but also the model is efficient, as shown in Figures 37 and 38 (Naresh *et al.*, 2019; Swarnkar, 2020; Wang *et al.*, 2020a).

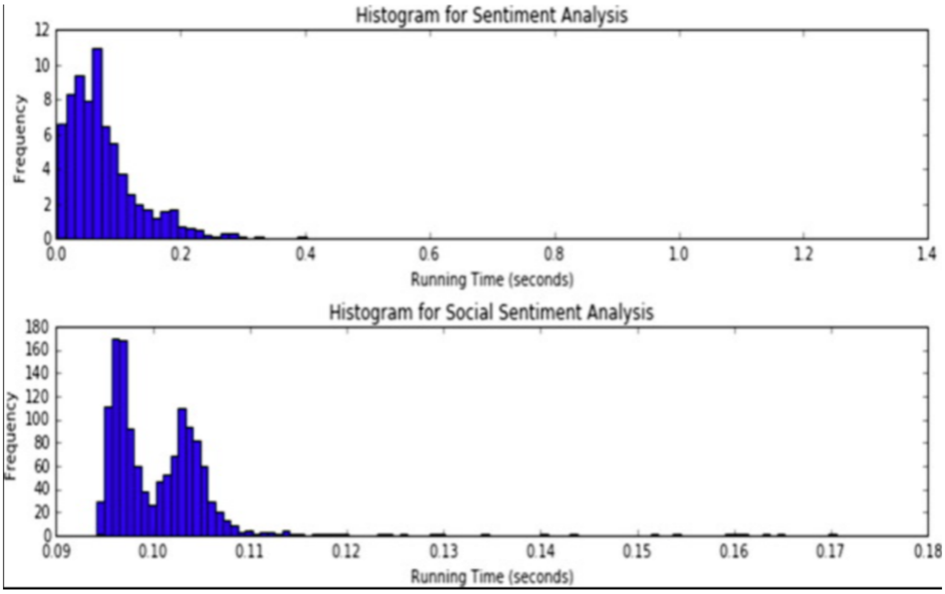


Figure 37. The running time of NLP vs VADER (Raj P M & Sai D, 2022)

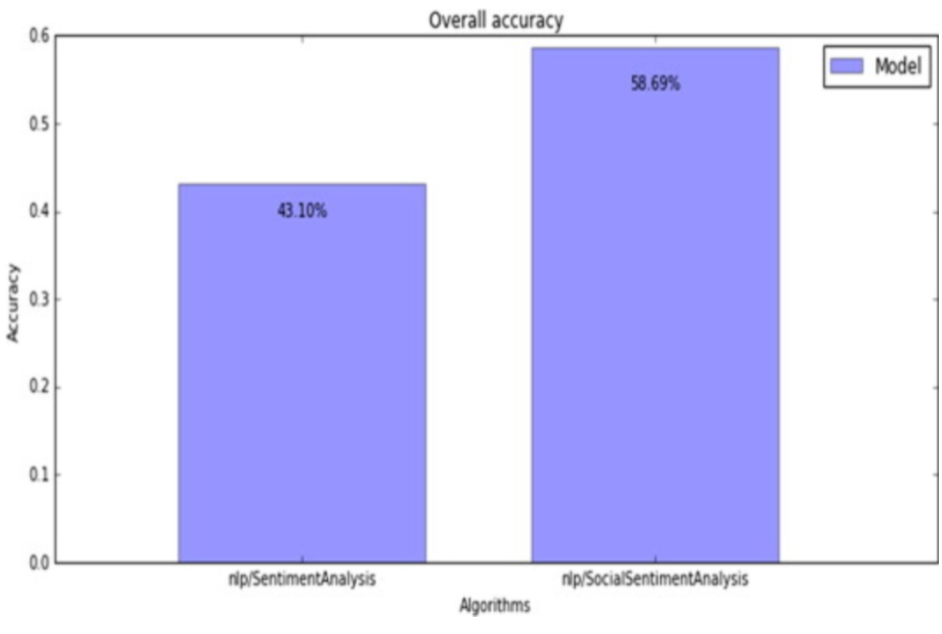


Figure 38. The accuracy comparison of NLP vs VADER (Raj P M & Sai D, 2022)

More influentially, based on the research of Raj P M and Sai D (2022), the basic principle of VADER is to score each word in the dictionary to synthesize scores for a sentence and a document. Therefore, the data characteristics of the four aspects separated from the film reviews can be directly scored by the VADER, so that the relationship between aspects and the overall score in each review can be found and the research problem can be solved. In Python, the VADER model also includes in the package of *NLTK*, which indicates attributes for the ease of use of the VADER model (Pipis, 2020).

According to Wang *et al.* (2020a), the input of modelling is features of reviews and each aspect data, and the output is the scores for each aspect and reviews, as shown in Table 6.

Table 6. The output template

Output Template						
ID	Overall	Actors	Directors	Plot	Music	Sentiment Classifer
1						
2						
3						
4						
5						
6						

The next process of modelling is to assess the correlation of the four aspects (actors, directors, plot, and music) and the overall. To simplify the model, an assumption is made as a premise, that is the relation between aspects score and overall is linear (Wang *et al.*, 2020a). Hence, the relationship between the aspects and the overall can be expressed by the correlation coefficient:

$$\rho(A,B) = \frac{cov(A,B)}{\sigma_A \sigma_B} \tag{1}$$

$$R = \begin{pmatrix} \rho(A,A) & \rho(A,B) \\ \rho(B,A) & \rho(B,B) \end{pmatrix} \tag{2}$$

Where:

A: the movie aspect score matrix

B: the overall movie score matrix

R: the correlation coefficient

3.4.4.2 SVM

Based on the conclusion of Section 2.2.3, SVM has the characteristics of high precision and stability (Moraes *et al.*, 2013). Meanwhile, due to the massive datasets will be used in the project, SVM is selected to construct the model in the project. Compared with the study of Bronchal (2016) and , the data preprocessing techniques used in the SVM model are the same as VADER, therefore, the processed data which have been done in Sections 3.4.1 and 3.4.2, and 3.4.3 can be used in SVM directly. The whole process of constructing a classification model of machine learning as shown in Figure 39.

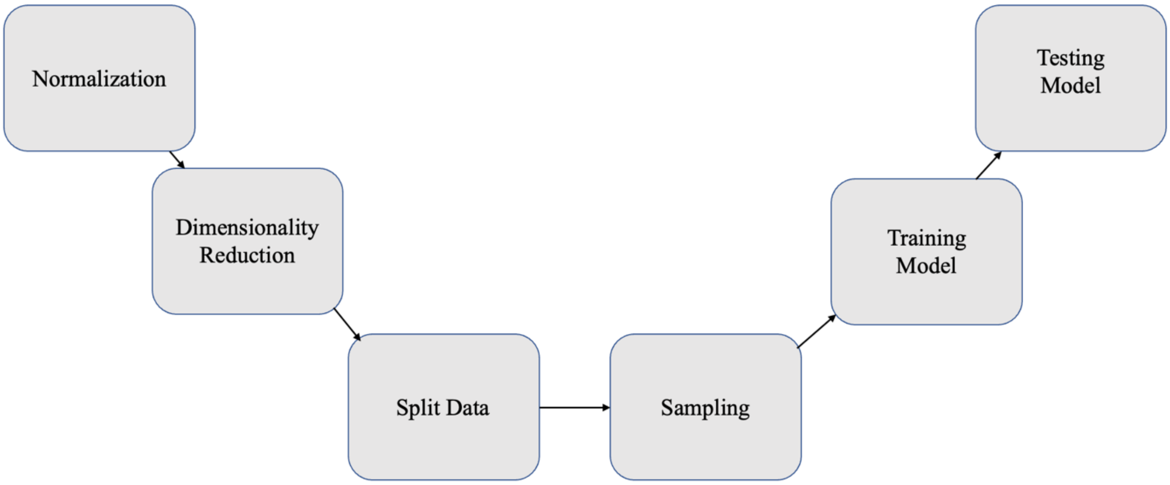


Figure 39. The general process for constructing a model of machine learning (Theodoridis, 2015)

The first and second step is to normalize the dataset and reduce its dimensionality, which aim to reduce the computational load of the model and improve the computational efficiency (Scikit Learn, 2022b). However, almost all features are in string format (movie reviews, titles), hence, normalization and dimensionality reduction which only focus on numerical features will not work for sentiment analysis. The third step is split into training data and testing data, in which the training data is prepared for constructing the model, and testing data is for prediction, which is shown in Figure 40 (Refaeilzadeh *et al.*, 2009; Theodoridis, 2015). In Python, it can be easily done by *model_selection* in *sklearn* library.

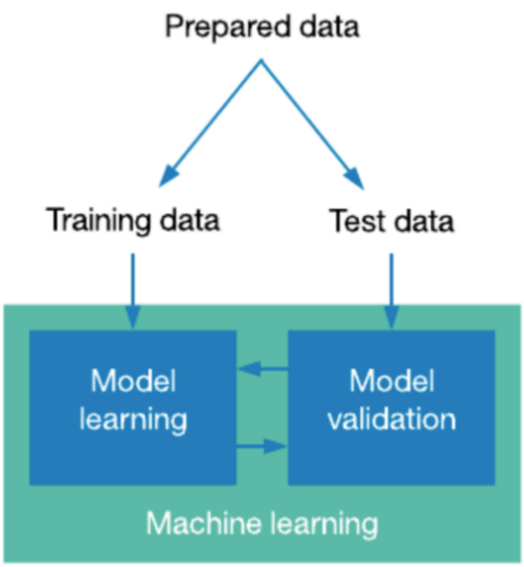


Figure 40. Splitting data in machine learning (Volpi, 2019)

For constructing SVM model, the s*klearn* library should be imported. there are several parameters which can be tuned, as shown in Table 7. To tune the parameters of SVM, *GridSearchCV* will be called which can perform an exhaustive search for input parameters (Scikit Learn, 2022a).

Table 7. The parameters of SVM (Scikit Learn, 2022c)

SVM Parameters			
Optional	Kernel	Gamma	Score
	linear	scale	roc_auc

	poly	auto	fi_macro
	rbf		fi_micro
	signmoid		accuracy
	precomputed		

Based on the best parameters which search by *GridSearchCV*, the result which is shown as classification of the sentiment will be output.

3.5 Result Correctness Verification and Evaluation of Models

This section mainly discusses that how to verify the correctness of the result and evaluate the models. In first objective, the features from the original collected data will be utilized to verify the correctness of VADER and SVM. Besides, the contents of Section 2.2.4 will be used as the benchmark to measure the performance of VADER with comparing with SVM.

3.5.1 Result Correctness Verification

There are two outputs of models (VADER and SVM) should be verified. For VADER, as mentioned in Section 3.3.4, the output of the model is each aspect rating, the overall rating, and the correlation coefficient. The first stage of proof is to verify the correctness of the conclusions drawn by the models for the total score, which will be accomplished by comparing with the ratings of the original collected data. Based on Table 2, if the collected data is balanced, the first consideration is to measure with Accuracy and Error, which are the most intuitive comparison standard for results (Birjali *et al.*, 2021). The formulas of Accuracy and Error are:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Error = 1 - Accuracy \tag{4}$$

Otherwise, if the collected data is imbalanced, F_1-score will be taken to measure the performance of the model, which is calculated from precision and recall and reflects the sensitivity of binary classification (Gao *et al.*, 2015). The formula of F_1-score is:

$$F_1 = \frac{2}{recall^{-1}+precision^{-1}} = 2 * \frac{precision*recall}{precision+recall} = \frac{TP}{TP+0.5*(FN+FP)} \tag{5}$$

Based on the conclusion of Ayata *et al.* (2020), Zhao *et al.* (2017), and He and Zhou (2011) an acceptable model should maintain Accuracy or F_1-score above 60%, which is also the benchmark of this project. Moreover, when the experimental data has a score of more than 70%, it can be concluded that the model is excellent.

3.5.2 Evaluation of Models

Based on the study of Section 2.2.4, the quality of the solution, running time, robustness, flexibility, and simplicity should be considered for evaluating the models (Barr *et al.*, 1995; Chakraborty *et al.*, 2019). In the project, VADER is the core model for solving research questions, and SVM is to assist the evaluation. For each evaluation aspect, the two models will be tested for performance as a cross-comparison variable.

The quality of the solution in the project is measured by the score of Accuracy or F_1-score, which has been discussed in Section 3.3.1. Meanwhile, based on the result of Raj P M and Sai D (2022), it can be conducted that VADER can outperform machine learning models in Accuracy. To compute the running time of the models, the time will be recorded during the whole process. Meanwhile, the robustness of the models refers to stability for different input text data. Hence, according to the principle of machine learning cross-validation, the original data set will be divided into several small subsets, which will be input into the model to obtain the output, and then the robustness of the model will be evaluated by Accuracy and F_1-score (Refaeilzadeh *et al.*, 2009).

3.6 Result Analysis and Reflection

To analyze the experiment outcome, a comparison will be conducted. The result of Wang *et al.* (2020a) has indicates that a successful movie is made by excellent plot, followed by the actors, the director, and music. Nevertheless, in their study, the films were not classified, so different experimental results may emerge in this project. In addition, the project will make some business insights based on the experimental results. For example, how to better improve the quality of movies? What do audiences expect from the movie?

3.7 Limitation

Although the analysis method and evaluation criteria of the whole project have been established in Chapter 3, there may still be some deficiencies. Based on Figure 14, this section chiefly discusses the limitations in each stage of the workflow.

3.8 Model Construction and Result Analysis

There are two sentiment classification models in the project, which is VADER and SVM. According to the conclusion of Birjali *et al.* (2021), the main disadvantage of lexicon-based model is that it mainly cares about a single word and ignores the context, so the VADER model is not acceptable at analyzing extreme expressions such as sarcasm. Whereas in SVM, if the number of features is too large, the model will not perform well (Al Amrani *et al.*, 2018). In the research, the defects of the two models need to be avoided, the accuracy in VADER needs to be improved by the SVM model comparison, and when building the SVM model, it is necessary to simplify the feature selection.

4. Case Study

4.1 Data Collection

The movie review dataset used in this project was collected by Fragkis (2021). Since the entire dataset can be downloaded directly on Kaggle, which is public dataset collected website and has been introduced in Section 3.4. In accordance with Appendix C, the collection, download and use of secondary data involved in the project no longer requires ethical approval, hence the data collection has been completed. Appendix D declare the dataset which is downloaded from Kaggle and Figure 41 states the head of the dataset.

	id number	Greek title	original title	category	director/creator	movie lenght	movie date	author	review date	review title	review	label	mean of stars	number of reviews	full reviews average stars
0	0	Lekin...	Lekin...	Drama	Gulzar	2 hours 51 minutes	1990	dwnpiyush-336-40603	8-Apr-12	For those who don't mind a slow pacing to the...	"Yaara Sili Sili Virah Ki Raat Ka Jalna"'Lekin...	8	9.16	6	7.5
1	1	Lekin...	Lekin...	Drama	Gulzar	2 hours 51 minutes	1990	cseabhi	10-Jul-21	Lekin - Gulzar's haunting masterpiece	Gulzar is at his best when he is telling such ...	9	9.16	6	7.5
2	2	Lekin...	Lekin...	Drama	Gulzar	2 hours 51 minutes	1990	abbott32	2-Jul-04	Haunting film but would love to have been abl...	I was completely mesmerized by Lekin and espec...	9	9.16	6	7.5
3	3	Lekin...	Lekin...	Drama	Gulzar	2 hours 51 minutes	1990	monoglot	29-Nov-05	An intriguing story well told.	Greatly enjoyed the development of the story I...	9	9.16	6	7.5
4	4	Lekin...	Lekin...	Drama	Gulzar	2 hours 51 minutes	1990	Kammu	27-Nov-99	It's a classic	The lines of time are very blurry. Past, prese...	10	9.16	6	7.5

Figure 41. The head of the collected dataset

The 320,000 pieces of data are composed of different categories of movie reviews, each with its own number and creation time. Each piece of data has 16 features, in which the 'Category' in the dataset allows researchers to efferently identify action movie reviews, and the 'Director' provide the information and convenience when split the whole reviews into four aspects. The 'Label' marks the author's rating for the corresponding movie, and 'URL' enables the viewer to go back to the IMDB website to find relevant and reliable information. 'Mean of stars' can be explained as the average ratings for a same movie which collected in the dataset, 'number of reviews' refer to the number of reviews of one movie collected in the dataset, and 'full reviews average stars' state the rating of the movie on the websites. Since the feature requirement of the project for the dataset is to include at least movie reviews, the names of all movies, ratings, and movie types, the dataset meets the expectations. However, the specific sentiment (positive or negative) is not given. To construct the sentiment classifier, the data in the 'Label' are simply divided into two sections, that is 1-5 is rated as 'Negative', and 6-9 is defined as 'Positive'.

4.2 Data Processing

Cleaning raw data, extracting, selection, and translating significant features into vectorized data (structured data) that machines can understand are the basic objectives of data processing. In this section, 4.3.1 mainly describes the process of data cleaning, tokenization, stemming, lemmatization, and removals of punctuation and stop words, which are all addressed by *NLTK*.

4.2.1 Data Preprocessing

The objective of data preprocessing is to construct an efficient model which can answer the research question. Hence, the first step to deal with the dataset is data cleaning, in which all action movie reviews are extracted, other gerne movies and useless data features are removed. Since reviews and corresponding ratings will be utilized in the feature extraction, feature selection and model construction, hence the rest features including 'Greek title', 'author', 'URL', 'id number', 'movie date', 'movie length', 'review date', and 'review title' are deleted, which is shown in Figure 42. As mentioned above, feature 'director' provides the information and basis for splitting unstructured data into director aspect, and rest features are about ratings of movies.

```
#Delete useless coumlns
#1. Greek title
#2. author
#3. url
#4. id_number (use index)
#5. movie date
#6. movie length
#7. review date
#8. review title

#remain the original data
df = data.copy()
df.drop(columns = ['Greek title', 'author', 'url','id number','movie date', 'movie lenght',
                   'review date', 'review title'], inplace = True)
```

```
df.head()
```

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars
0	Lekin...	Drama	Gulzar	"Yaara Sili Sili Virah Ki Raat Ka Jalna"'Lekin...	8	9.16	6	7.5
1	Lekin...	Drama	Gulzar	Gulzar is at his best when he is telling such ...	9	9.16	6	7.5
2	Lekin...	Drama	Gulzar	I was completely mesmerized by Lekin and espec...	9	9.16	6	7.5
3	Lekin...	Drama	Gulzar	Greatly enjoyed the development of the story I...	9	9.16	6	7.5
4	Lekin...	Drama	Gulzar	The lines of time are very blurry. Past, prese...	10	9.16	6	7.5

Figure 42. The dataset after removing useless features

The second step is to extract all the action movie reviews from the whole dataset. To realize the purpose, the feature 'category' is regarded as filter, and the result is expounded in Figure 43. In total, there are about 70,000 action movie reviews in the dataset, and authors have mixed reviews for the same movie. According to the data collection requirements in Section 3.4, the first 1000 pieces of data are extracted, which can ensure the efficiency of data analysis and model construction.

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	8	8.0	1	5.8
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	1	1.0	1	4.3
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	1	5.83	6	6.6
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	6	5.83	6	6.6
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	6	5.83	6	6.6
...	...	...	...	...	...	...	...	...
70248	Mario Bros.	Action	Shigeru Miyamoto	After Mario's debut as "Jumpman" in the origin...	7	6.88	8	7.5
70249	Mario Bros.	Action	Shigeru Miyamoto	This game must be very confusing to the younge...	7	6.88	8	7.5
70250	Mario Bros.	Action	Shigeru Miyamoto	A classic Nintendo game for old-school video g...	8	6.88	8	7.5
70251	Mario Bros.	Action	Shigeru Miyamoto	This is another honorable mention in my favori...	9	6.88	8	7.5
70252	Mario Bros.	Action	Shigeru Miyamoto	In your face Sega this was before Sonic & Tail...	10	6.88	8	7.5

70253 rows × 8 columns

Figure 43. Reviews of action movies in the dataset

Besides, texts relating to four factors (plot, director, actors, and music) are filtered out to prevent the meaning of the original data expression from being modified throughout the data processing and leading to imprecise analytical results. A filter should be created which can figure out most related four aspects elements. Table 8 is provided by Wang *et al.* (2020a), in which movie aspects related words provide information about text related to the four aspects of the search, while specific terms indicate text that may be missed. Necessarily, the director's name of each movie is supplied in the dataset, allowing for a more accurate classification of the aspect of director. Nonetheless, other specific terms including, the names of actors, as well as some nicknames and abbreviations of names, become the limitations of this project, which will be mentioned later.

Table 8. A dictionary for creating the filter of text extraction (Wang *et al.*, 2020a)

Aspects	Movie aspects related words	Specific terms
Actors	Actors, actress, acting, cast, casting, role, performance	Actor names
Directors	Director, directed, filmed, edition	Director name
Plot	Plot, story, character, romance, revenge, ending, script	Script Writer name
Music	Lyrics, songs, ending song, background music, sound	Singer name

In Python, all text is firstly converted to lowercase, and the program is set to index the text data by keywords according to each aspect, and once the keywords are found, the paragraph is stored, and the grouped results are shown in Figure 44.

	original title	category	director/creator	review	plot	actor	director	music	label	mean of stars	number of reviews	full reviews average stars	lower_review
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	the plot centers on a triad gangster ripping ...				8	8.0	1	5.8	you know what i mean. hong kong action comedie...
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	the main character, played by frank stallone,...	the main characters are an annoying couple of...	it was directed by porn director john t. it w...		1	1.0	1	4.3	lethal games is an absolutely awful attempt at...
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	but the story is stupid, inconsistent,contrive...	the acting is hammy ,full of overacting.joyce...			1	5.83	6	6.6	the film is about royal hong kong police fight...
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	sammo hung also has a role, but unfortunately...	the acting was quite good, especially those of...			6	5.83	6	6.6	this is an intense action movie from hong kong...
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...					6	5.83	6	6.6	one of my favorite fight scene is when joyce k...

Figure 44. The dataset after filtering texts of four aspects

There are some blanks of aspect of each review, hence a counter is used to compute the number of texts in each aspect, which is shown in Table 9. The text has the least information about the director, and the audience talks the most about the plot.

Table 9. Count for texts of each aspect

Count for each aspect				
total	Plot	Director	Actor	Music
1000	905	504	611	771

The next step is to process the text content, and the feature review as an example to show in this section. For a more convenient display, the researcher only compiles and operates the feature 'review' to complete data processing, and the text content of the four aspects can be programmed by the same code.

Figure 45 demonstrates the process of remove punctuation, where the NLP library is not involved, instead all punctuation marks can be effortlessly recognized in using the *string* library.

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars	re_punc
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	8	8.0	1	5.8	You know what I mean Hong Kong action comedies...
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	1	1.0	1	4.3	LETHAL GAMES is an absolutely awful attempt at...
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	1	5.83	6	6.6	the film is about royal hong kong police fight...
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	6	5.83	6	6.6	This is an intense action movie from Hong Kong...
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	6	5.83	6	6.6	One of my favorite fight scene is when Joyce k...

Figure 45. Remove punctuation in the reviews

Secondly, all letters in the text should be converted to lowercase, which has already been implemented in extracting the four related aspects. Then, tokenization is the third step in text processing, which significantly reduces the storage space of data and improves the efficiency of sentiment analysis by splitting a sentence into words. After tokenization, each review stores in an array, and each word in one review is an element of the array. The result of lowering the letters and tokenization is expounded in Figure 46.

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars	re_punc	lower	tokenization
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	8	8.0	1	5.8	You know what I mean Hong Kong action comedies...	you know what i mean hong kong action comedies...	[you, know, what, i, mean, hong, kong, action,...
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	1	1.0	1	4.3	LETHAL GAMES is an absolutely awful attempt at...	lethal games is an absolutely awful attempt at...	[lethal, games, is, an, absolutely, awful, att...
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	1	5.83	6	6.6	the film is about royal hong kong police fight...	the film is about royal hong kong police fight...	[the, film, is, about, royal, hong, kong, poli...
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	6	5.83	6	6.6	This is an intense action movie from Hong Kong...	this is an intense action movie from hong kong...	[this, is, an, intense, action, movie, from, h...
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	6	5.83	6	6.6	One of my favorite fight scene is when Joyce k...	one of my favorite fight scene is when joyce k...	[one, of, my, favorite, fight, scene, is, when...

Figure 46. Lowering letters and Tokenization of text reviews

As mentioned in Section 3.5.1, the stop words also have no meaning in the text, which should be removed as well. In this stage, the *NLTK* library is used, which stores a dictionary for stop words such as personal and demonstrative pronouns. (Avinash & Sivasankar, 2019). Figure 47 elaborates the process of removing stop words.

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars	re_punc	lower	tokenization	stop_words
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	8	8.0	1	5.8	You know what I mean Hong Kong action comedies...	you know what i mean hong kong action comedies...	[you, know, what, i, mean, hong, kong, action,...	[know, mean, hong, kong, action, comedies, ten...
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	1	1.0	1	4.3	LETHAL GAMES is an absolutely awful attempt at...	lethal games is an absolutely awful attempt at...	[lethal, games, is, an, absolutely, awful, att...	[lethal, games, absolutely, awful, attempt, ac...
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	1	5.83	6	6.6	the film is about royal hong kong police fight...	the film is about royal hong kong police fight...	[the, film, is, about, royal, hong, kong, poli...	[film, royal, hong, kong, police, fighting, vi...
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	6	5.83	6	6.6	This is an intense action movie from Hong Kong...	this is an intense action movie from hong kong...	[this, is, an, intense, action, movie, from, h...	[intense, action, movie, hong, kong, featuring...
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	6	5.83	6	6.6	One of my favorite fight scene is when Joyce k...	one of my favorite fight scene is when joyce k...	[one, of, my, favorite, fight, scene, is, when...	[one, favorite, fight, scene, joyce, kicks, ag...

Figure 47. Removing stop words

The last step is lemmatization, which is implemented by the *NLTK* library to find the root of the word and restore it to the most concise form, this can greatly reduce the running time of the program and speed up the construction of the model.

	original title	category	director/creator	review	label	mean of stars	number of reviews	full reviews average stars	re_punc	lower	tokenization	stop_words	lemmatization
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	8	8.0	1	5.8	You know what I mean Hong Kong action comedies...	you know what i mean hong kong action comedies...	[you, know, what, i, mean, hong, kong, action,...	[know, mean, hong, kong, action, comedies, ten...	[know, mean, hong, kong, action, comedy, tend,...
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	1	1.0	1	4.3	LETHAL GAMES is an absolutely awful attempt at...	lethal games is an absolutely awful attempt at...	[lethal, games, is, an, absolutely, awful, att...	[lethal, games, absolutely, awful, attempt, ac...	[lethal, game, absolutely, awful, attempt, act...
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	1	5.83	6	6.6	the film is about royal hong kong police fight...	the film is about royal hong kong police fight...	[the, film, is, about, royal, hong, kong, poli...	[film, royal, hong, kong, police, fighting, vi...	[film, royal, hong, kong, police, fighting, vi...
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	6	5.83	6	6.6	This is an intense action movie from Hong Kong...	this is an intense action movie from hong kong...	[this, is, an, intense, action, movie, from, h...	[intense, action, movie, hong, kong, featuring...	[intense, action, movie, hong, kong, featuring...
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	6	5.83	6	6.6	One of my favorite fight scene is when Joyce k...	one of my favorite fight scene is when joyce k...	[one, of, my, favorite, fight, scene, is, when...	[one, favorite, fight, scene, joyce, kicks, ag...	[one, favorite, fight, scene, joyce, kick, agn...

Figure 48. Lemmatization

The final step, shown in Figure 48, completes the text preprocessing. Essentially, in the text processing stage, unnecessary features have been eliminated, the sentences in the text have been divided into four aspects based on the keywords, and all of the texts have been processed into root form, which is in the word-based units.

4.2.2 Feature Extraction and Selection

Acquired from Sections 3.5.2 and 3.5.3, the feature extraction and selection should be construction when completing data preprocessing. The BoW model is the fundamental model for the feature extraction in the project, which will be implemented by the function *CountVectorizer* in the *sklearn*, which also integrates the process of in text processing including tokenization, stop word removal and lowering the letters (Scikit Learn, 2022e). Figure 49 expounds the process and result of constructing *CountVectorizer*. The most important features in each review are extracted, meanwhile, the reviews are converted into a vector matrix, which can be recognized by the machine and used in the sentiment analysis model.

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

CountVec = CountVectorizer(# to use bigrams ngram_range=(2,2)
                           stop_words='english')
    #transform
Count_data = CountVec.fit_transform([sample['review'][0]])

#create dataframe
cv_dataframe=pd.DataFrame(Count_data.toarray(),columns=CountVec.get_feature_names())
print(cv_dataframe)
```

```
   able  accident  action  angel  auntie  better  bloody  boss  cash  catch  \
0     1         1       2      1       1       1       1     2     1      1

   ...  story  support  tend  triad  veer  way  wildly  yin  young  yukari
0  ...      1        1     1      1     1    1       1    1      1       1
```

Figure 49. The process of constructing the BoW model

Since a vector matrix is generated for each review, the program runs slower at this stage. Furthermore, the BoW model can autonomously create dictionaries as the filter in each review, and all aspects of text have been extracted in text processing, so all required features have been stored. The researchers did not further use feature selection to process the data again, which may cause the data to become obscured and the analysis results to be inaccurate.

4.3 Model Construction

Evolved from Section 3.5.4, the project focus on two models: the support vector machine (SVM) and VADER, which will be constructed in this section. As a two-category machine learning model, SVM does not meet the multivariate scoring mechanism required in the project, instead, it can only be used as an auxiliary model to roughly explore the relationship between the four aspects and the success of the movie. The latter, VADER, is indispensable analytical tool in the project, which answer research questions bluntly. Furthermore, SVM can be used as a benchmark whose runtime, correctness evaluation, and efficiency will be compared with VADER, supporting the conclusions drawn by the project and enhancing the improvement of the model construction.

4.3.1 SVM

4.4.1.1 Dataset Splitting

Since SVM belongs to supervised learning of the machine learning, the model needs to find the best parameters through training, and then put the trained model structure into the actual prediction (Wang & Zhao, 2020). Therefore, the process of constructing SVM starts from dataset splitting.

```
train, test = train_test_split(svm_df, test_size=0.2, random_state=1)
X_train = train['review'].values
X_test = test['review'].values
y_train = train['label']
y_test = test['label']
```

Figure 50. Dataset splitting

As shown in Figure 50, the training set and test set are split in a ratio of 8:2, which means 800 reviews will be put into the training process, and *random_state* ensures the consistency of each data split. Moreover, to make training on the data more efficient, the *StratifiedKFold* function is called.

```
kfolds = StratifiedKFold(n_splits=5,
                         shuffle=True,
                         random_state=1)
```

Figure 51. The setting of the *StratifiedKFold*

There are three parameters in this function, including *n_splits*, *shuffle*, and *random_stateint*. *n_splits* splits the training set into n folds, *shuffle* determines the disorder or order of the data, while *random_stateint* control the randomness of each fold (Scikit Learn, 2022a). The setting of those three parameters is shown in Figure 51.

4.3.1.2 Model Construction for SVM

Combined with the output of data preprocessing, feature extraction, feature selection and dataset splitting, the text has been transformed into structured data, which can be applied to the SVM model. Figure 52 clarifies the process of model construction, where the *make_pipeline* function is utilized to concatenate text processing, data vectorization, dataset segmentation and the SVM model.

```
train, test = train_test_split(svm_df, test_size=0.2, random_state=1)
X_train = train['review'].values
X_test = test['review'].values
y_train = train['label']
y_test = test['label']
def tokenize(text):
    tknzr = TweetTokenizer()
    return tknzr.tokenize(text)
def stem(doc):
    return (stemmer.stem(w) for w in analyzer(doc))
en_stopwords = set(stopwords.words("english"))
vectorizer = CountVectorizer(
    analyzer = 'word',
    tokenizer = tokenize,
    lowercase = True,
    #ngram_range=(1, 1),
    stop_words = en_stopwords)
kfolds = StratifiedKFold(n_splits=5,
                         shuffle=True,
                         random_state=1)
np.random.seed(1)
pipeline_svm = make_pipeline(vectorizer,
                            SVC(probability=True, kernel="linear", class_weight="balanced"))
grid_svm = GridSearchCV(pipeline_svm,
                    param_grid = {'svc__C': [0.01, 0.1, 1]},
                    cv = kfolds,
                    scoring="recall",
                    verbose=1,
                    n_jobs=-1)
grid_svm.fit(X_train, y_train)
time_end=time.time()
print(grid_svm.score(X_test, y_test))
print('totally cost time(s)',time_end-time_start)
grid_svm.best_params_
```

```
Fitting 5 folds for each of 3 candidates, totalling 15 fits
0.9375
totally cost time(s) 2.5704538822174072
{'svc__C': 0.1}
```

Figure 52. Model construction for SVM

In this model, only one parameter *svc_C* is changed, that is, through the training of the model, the machine needs to find the most suitable *svc_C* for the dataset, which is 0.1. Combined with Table 2, recall is selected to measure the performance when constructing the model, and the testing score is 0.9375, which means the predicting ability of the model is powerful and it can be applied to further prediction directly. On the other hand, the running time of the SVM model has also been computed, which is about 2.57s. To further evaluate the performance of the model, Figure 53 displays F1-score, accuracy, precision, and recall.

```
def report_results(model, X, y):
    pred_proba = model.predict_proba(X)[:, 1]
    pred = model.predict(X)

    #auc = roc_auc_score(y, pred_proba)
    acc = accuracy_score(y, pred)
    f1 = f1_score(y, pred)
    prec = precision_score(y, pred)
    rec = recall_score(y, pred)
    result = { 'f1': f1, 'acc': acc, 'precision': prec, 'recall': rec}
    return result
```

```
report_results(grid_svm.best_estimator_, X_test, y_test)
```

```
{'f1': 0.9146341463414636,
 'acc': 0.86,
 'precision': 0.8928571428571429,
 'recall': 0.9375}
```

Figure 53. The scores of the SVM model for F1-score, accuracy, precision, and recall

Therefore, due to the excellent performance of the SVM model, it is applied to predict sentiment polarity in four aspects. The researcher filters out the data with text in four aspects and predicts the positive and negative sentiments in turn. The results are shown in Table 10.

Table 10. Predicting result of the SVM model

	plot	actor	director	music	label
0	1	1	1	1	1
1	1	0	1	1	1
2	0	0	1	0	0
3	1	1	1	1	0
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	0	1	1	1	1
9	1	1	1	1	1
10	0	1	1	1	1
11	0	1	0	1	1
12	0	1	1	1	0
13	1	1	1	1	0

It should be emphasized that SVM is a binary classification model, so in the project 0 is used to replace negativity and 1 refers to positive sentiment in the analysis results.

4.3.2 VADER

The second model built in the project is VADER, which is altogether encapsulated in Python. Subject to Section 3.5.4, the preliminary work that needs to build the VADER model has been completed in Section 4.3, hence the package *NLTK* can be straight to utilize to construct VADER. Figure 54 describe the primarily output of VADER, which includes four elements: negative, neutral, positive, and compound. In these four elements, the compound represents the composite score of the sentence, therefore, it is chosen to clarify the polarity and final score of each text data (Hutto & Gilbert, 2014).

```
tem = sid.polarity_scores(sample.review[0])
print (tem)
{'neg': 0.093, 'neu': 0.779, 'pos': 0.128, 'compound': 0.9429}
```

Figure 54. The output of VADER model

Figure 55 declares the specific score evaluation for each text through the VADER model. Since *compound* is a float which is between -1 and 1, the original labels of the data need to be transformed to compare with the results of the VADER model. Therefore, the classification criteria in 4.2 are refined, in this section, a label which is greater than 5 is set as positive, equal to 5 as neutral, otherwise negative. For the evaluation result, *compound* which is greater than 0 is positive, equal to 0 as neutral, and less than 0 is negative.

	original title	category	director/creator	review	plot	actor	director	music	label	mean of stars	number of reviews	full reviews average stars	compound
0	Yue gui xing dong	Action	Benny Wong	You know what I mean. Hong Kong action comedie...	the plot centers on a triad gangster ripping ...				8	8.0	1	5.8	0.9496
1	Lethal Games	Action	John T. Bone	LETHAL GAMES is an absolutely awful attempt at...	even scriptwriter daniel benton used a pseudo...	the main character, played by frank stallone,...	it was directed by porn director john t. it w...		1	1.0	1	4.3	-0.9769
2	Huang jia nu jiang	Action	Corey Yuen	the film is about royal hong kong police fight...	but the story is stupid, inconsistent,contrive...	the acting is hammy ,full of overacting. prob...			1	5.83	6	6.6	0.9528
3	Huang jia nu jiang	Action	Corey Yuen	This is an intense action movie from Hong Kong...	sammo hung also has a role, but unfortunately...	the acting was quite good, especially those of...			6	5.83	6	6.6	-0.9565
4	Huang jia nu jiang	Action	Corey Yuen	One of my favorite fight scene is when Joyce k...	after she recovers she tries to kick joyce bu...				6	5.83	6	6.6	-0.7391

Figure 55. The evaluation of VADER for each review

In this case, the VADER, which has three classes, is a multi-classification model in the project, the accuracy, precision, and recall cannot be computed at this time. However, since the values of labels and *compound* have been normalized so that it can roughly estimate the accuracy of the model. The formula of accuracy is

$$Accuracy = \frac{Successfully\ evaluated\ movie\ reviews}{Total} \tag{6}$$

Where

Total is the number of the sample

The result of accuracy is 0.762, which means the model has reached the eligibility criteria in the evaluation, and it can be applied for finding the relationship between the texts of four aspects and each review. First of all, the data whose polarity is wrongly judged is deleted, which would seriously affect the assessment of correlations. Then, following the same steps as evaluating film reviews, all texts in the four areas are evaluated, and *compound* is selected as the final score, and the basic data frame is shown in Figure 56.

```
        plot   actor  director   music  compound
0     0.0000     NaN       NaN     NaN    0.9496
1     0.4927 -0.4215    0.0000     NaN   -0.9769
2    -0.9452 -0.4595       NaN     NaN   -0.9565
3    -0.5267     NaN       NaN     NaN   -0.7391
4        NaN  0.6124    0.8313     NaN    0.8268
..       ...     ...       ...     ...       ...
757   0.7517 -0.9783   -0.8750  0.2975   -0.9161
758  -0.3286 -0.1695    0.6690 -0.1695    0.9641
759      NaN  0.9831       NaN     NaN    0.9887
760   0.3182  0.3182       NaN     NaN   -0.7215
761   0.3536  0.6059    0.0516  0.0516    0.9429

[762 rows x 5 columns]
```

Figure 56. The VADER output

After removing all false results, a total of 762 data remained, which matched the accuracy. However, due to the lack of text data in four aspects, the final evaluation results may be affected. Based on the Formulas (1) and (2), the final relationship is shown in Figure 57.

```
relation_df.corr(method = 'spearman')
```

	plot	actor	director	music	total
plot	1.000000	0.310922	0.309358	0.342707	0.533740
actor	0.310922	1.000000	0.409754	0.554324	0.513330
director	0.309358	0.409754	1.000000	0.333707	0.409272
music	0.342707	0.554324	0.333707	1.000000	0.436860
total	0.533740	0.513330	0.409272	0.436860	1.000000

Figure 57. The relationship between reviews and four aspects

The direct relationship of each variable is stated in this two-dimensional matrix, where 'total' represents each complete film review. In VADER's evaluation results, the plot has the greatest impact on the success of action films, followed by the actors, the music, and finally the director.

5. Model Evaluation & Result Analysis

5.1 Model Evaluation

The evaluation for the SVM and the VADER is constructed in this section. Formulated on Section 3.6, basically there are two parts should be computed for model evaluation: result correctness verification and performance measurement. In result correctness verification, due to the different types of the two models, the project utilizes distinctive method to verify the correctness. Since the SVM model is a binary classification supervised machine learning method, a 2x2 confusion matrix is built, and accuracy, recall, F1-score, and precision are computed to demonstrate the correctness of the SVM model. For the VADER model, Section 4.4.2 has stated the formula of verification. However, whether the dataset is balanced needs to be counted before using accuracy. All above will be done in Section 5.2.1, and when result correctness verification completed, the two models are allowed to predict the contents of four aspects. In Section 5.2.2, the performance of the two models is evaluated, which follows the benchmarks in Section 3.6.2. In the project, the quality of the solution, running time, robustness, flexibility, and simplicity should be considered for measuring the performance of the two models.

5.1.1 Result Correctness Verification

5.1.1.1 SVM

Since the SVM model is binary classification supervised model, the sentiment classifier of the SVM only includes positive and negative. Therefore, a 2x2 confusion matrix, which can be computed for binary classification model, can be constructed for displaying the final output of the SVM model, as shown in Figure 58.

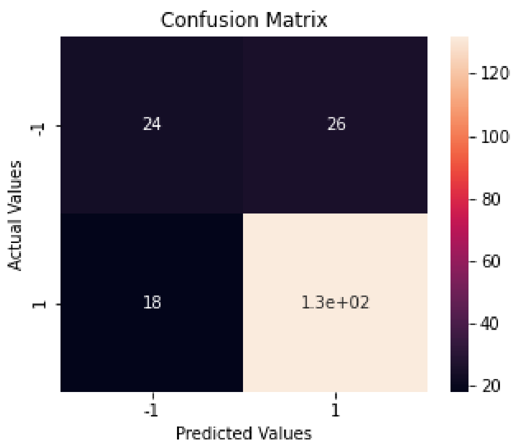


Figure 58. The confusion matrix for testing set of the SVM model

The confusion matrix consists of a rectangular coordinate system, where the abscissa represents the predicted label, and the ordinate refers to the true label. Generally, Figure 58 can be divided into 4 parts, and according to Section 3.6.1, from left to right, the four parts respond to *TN*, *FP*, *FN*, and *TP* respectively. Meanwhile, the shades of color represent the number of datasets in each part, where the lighter the color, the greater the amount of data predicted. Therefore, it can be found that the testing dataset is considered an imbalanced dataset, where the true positive labels are much more than true negative labels so derived from Section 3.6.1, the SVM model cannot be evaluated by accuracy. Meanwhile, the model can predict positive reviews sufficiently, since a large number of true positive labels are successfully predicted. Nevertheless, for the ability to predict the negative reviews, other evaluating standards such as recall and F1-score should be computed. Figure 53 has already states the scores of the SVM model for F1-score, accuracy, precision, and recall, and based on Section 3.6.1, for measuring the imbalanced dataset result, F1-score, which is calculated from precision and recall and reflects the sensitivity of binary classification, is the best standard. As shown in Figure 53, the F1-score of the SVM model is about 0.86, which proves that the model built in the project far exceeds the qualified standard and has a strong ability to judge text sentiment.

5.1.1.2 VADER

Different from the SVM model, the VADER is not a supervised machine learning model for binary classification, so the traditional 2x2 confusion matrix cannot be visualized to represent its classification results. On the other hand, for measuring the performance of VADER, a method for constructing accuracy has been proposed in Section 4.4.2, and the score of the accuracy is 0.762. Before using this evaluation criterion, the number of labels of different types in the dataset needs to be counted, and the accuracy cannot be accepted if the dataset is imbalanced. Hence, *Counter* from *Collections* library is called for counting the dataset, and the result is shown in Figure 59.

```
from collections import Counter
Counter(y_test)
```

```
Counter({1: 500, -1: 269, 0: 231})
```

Figure 59. The number of each label in the dataset

Since VADER is not a supervised machine learning model, the dataset is not split into training set and testing set, therefore, the total number of tested data is 1000. Figure 59 depicts that the data with different labels is basically uniform, so that the accuracy can be accepted to be used to measure model performance. Combined to the benchmark which is proposed in Section 3.6.1, the final score of the VADER model is higher than 60%, the further work, including appraising the four aspect classifiers, can be computed and the result correctness of the VADER model is verified.

5.1.2 Performance Measurement

Since the SVM model cannot answer the research question directly, in the project it is regarded a cross-comparison variable for the VADER model. Subject to Section 3.6.2, the quality of the solution, running time, robustness, flexibility, and simplicity should be considered for measuring and comparing the performance of the two models. Table 11 expound the main features and result of the two models.

Table 11. Performance measurement of the two models

Performance Measurement					
	Quality of solution	Running time (s)	Robustness	Flexibility	Simplicity
SVM	F1-score	10.127	Strong	Strong	Medium
	0.857				
VADER	Accuracy	5.2	Strong	Weak	Strong
	0.762				

Note that both models were run on the same computer, which means that all hardware effects have been excluded during the evaluation. In these five criteria, the strengths of each model have been marked in red, while the disadvantages are marked in white. The quality of solution has been evaluated in Section 5.2.1, the conclusion is that since the SVM and the VADER do not belong to the same type of model and the outputs of the two models are different, the strengths of each model can be reflected. The SVM is more limited and is only suitable for binary classification models (positive and negative), so it cannot meet most objectives in sentiment analysis. The VADER evaluates the sentiment of a movie review by specific numerical values, and the program can convert the numerical values into any corresponding output, such as sentiment labels and classifier. For the running time, the SVM took nearly twice as long as the VADER model, which can be conducted that the SVM model is more efficient. The possible reason is that in Python, the VADER is completely encapsulated in the library, and a simple statement can call the entire VADER model, while the SVM model requires programmers to build the process from data preprocessing. Moreover, when constructing a model, SVM needs to debug the most appropriate parameters and hyperparameters through the training stage, but the dictionary-based VADER model automatically completes sentiment analysis. Robustness refers to the ability of the model to perform consistently when involving different input datasets. For the SVM, the project has verified the robustness of the model at training time through *tratifiedKFold*, and the result depicts the F1-score of the SVM model deals with the steady state no matter how the split between training and test sets changes. Meanwhile, there is no need to split dataset for the VADER model, further confirmation is required for the robustness verification of the VADER model. Therefore, the collected dataset is used again. The researchers randomly select 5,000 reviews from 70,000 action movie reviews, divide them into five groups, put each group into the VADER model to predict the sentiment classifiers, and calculate the accuracy of each model according to formula (6), and the result is shown in Table 12, where the maximum and minimum values of accuracy have been marked in the table.

Table 12. The 5 experiments for testing robustness for the VADER model

Accuracy for the VADER model						
	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Accuracy	0.792	0.804	0.693	0.755	0.768	0.762

From the table, it can be conducted that the average accuracy of the five experiments is consistent with that reported by the researchers. Although the results of the model fluctuate slightly, it is still in a stable state, which proves that the VADER model also has strong robustness. Flexibility is assessed by two aspects: (1) the ability to combine with other functions, and (2) the ability to change and modification for different output requirements. In the SVM model *make_pipeline* is used, which allows data preprocessing, dataset segmentation and model construction to be integrated. Moreover, the machine learning model can tune the parameters through *GridSearchCV*, so the two aspects are satisfied by the SVM model, which means the model has strengths on the perspective of flexibility, but the simplicity is a bit complex.

5.2 Result Analysis

In Section 5.1, the result correctness verification has demonstrated the superiority and reliability of the model constructed in the project, therefor, in this section can conduct analysis and reflection based on the output of the two models. Figure 60 depicts the result of the SVM model, which depicts the relationship between four aspects and the label. First of all, rows 3 and 11 are special since when all four are positive, the label is rated as negative, which means that there are other factors that influence the sentiment polarity of movie reviews, and another reason is that these two data is wrong predicted by the SVM model. Therefore, the two data can be removed from the analysis. Meanwhile, row 12 shows the importance of the label, and row 11 states the irrelevance of the label compared with the row 12 and no obvious conclusions can be drawn from the remaining data. In general, the researchers found from the SVM model that in addition to these four aspects, there are still other factors that affect the polarity of labels. Diversely, due to the small sample size and the binary classification of labels, it is difficult to analyzing the results of the SVM model so that the conclusion cannot be conducted based on this model which is related to the research question.

	plot	actor	director	music	label
0	1	1	1	1	1
1	1	0	1	1	1
2	0	0	1	0	0
3	1	1	1	1	0
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	0	1	1	1	1
9	1	1	1	1	1
10	0	1	1	1	1
11	0	1	0	1	1
12	0	1	1	1	0
13	1	1	1	1	0

Figure 60. Predicting result of between the label and four aspects in the SVM model

However, the result of the VADER model provides a strong indication of the research question. Figure 61 is obtained by visualizing Figure 57, showing the relationship between the four aspects and the success of the action movies in the form of a heatmap. Compared with the output of the SVM model, the precise numerical value intuitively reflects the connection between the variables. From a subjective point of view, the experimental results depict that the audience for the success of an action film is primarily concerned with the development of the plot, followed by the actors and music, and finally the director, which is also consistent with the conclusions observed in the SVM model.

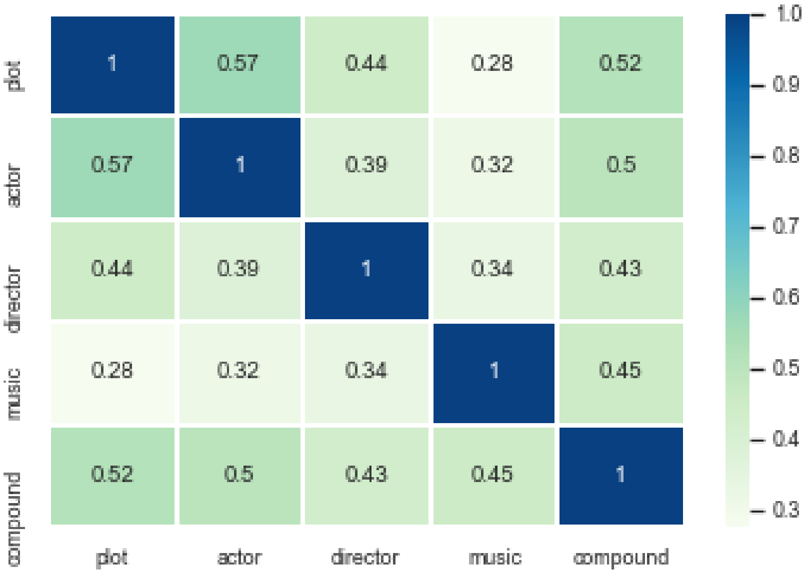


Figure 61. The relationship between reviews and four aspects in the VADER model

Nevertheless, compared with the study of Wang *et al.* (2020a), the conclusions of this project are somewhat conflicting, which may be caused by following reasons: (1) The gernes of movies are not filtered by the study of Wang *et al.* (2020a). Since the variety of movies, the target audiences attracted by different movies are far disparate, and the focus of these audiences also changes accordingly. (2) In this project, to extract more effective sentences about the four aspects, the feature extraction is advanced in the data processing, which has some deviations from the project prototype, hence the input of the model also has been influenced. (3) The conclusion is also influenced by the collected dataset and the analysis method.

Overall, the answer to the research question can be explained as audiences are more likely to be drawn to the storyline in action films. Therefore, in order to improve the quality of an action movie, it is even more necessary for the producer to capture the audience's interest through the plot design. Of course, excellent actors and moving music can attach more highlights to the elaboration and performance of the action movie.

5.3 Limitation

In Section 5.3, researchers have analyzed the results of the project, and it can be conducted that although the models used in the project perform relatively well, there are still some flaws that can be promoted. This section mainly describes the limitations in the process of programming and the stages that can be enhanced, which will be built based on the Section 3.8, explaining the evaluation following the steps of the construction of sentiment analysis.

5.3.1 Data Processing

There are three phrases in the data processing: data preprocessing, feature extraction and feature selection. In data preprocessing, removing punctuations and stop words, computing tokenization, normalization and lemmatization are done. However, the text is divided into word-based units, so the order of the sentences and the meaning of the context are separated, which leads to the construction of the entire project bases on a dictionary-based analysis method, and the analysis results lack the understanding of coherent sentences. In feature extraction and selection, although the data set is suitable for the research objectives of this project, there is abundant incomplete data in the dataset after extracting the texts of the four aspects, which is accomplished by using the dictionary introduced in Table 8. Additionally, the extracted text may duplicate appearing several times, making the efficiency of the data processing stage is reduced. The reasons for the obstruction are: (1) The dictionary used in data extraction does not effectively segment the text in four aspects. For example, for a person's name, the system cannot identify whether it belongs to the director's name or the actor's name; (2) A paragraph contains many aspects of content. Meanwhile, Section 5.3 has proved that there may be other factors in the text that affect the labels, and these factors affect the final analysis results.

5.3.2 Model Construction

The fully encapsulated VADER makes it difficult to implement parameter tuning during model building, so that the modification of the output only depends on the input (dataset). Additionally, based on the study of Wang *et al.* (2020a), it is challenging for VADER to accurately assess complex sentiment like sarcasm and criticism, which could cause inaccuracies in the findings of experiments. For SVM, derived from Section 2.2.3, it is an efficient and stable low-dimensional classification model. However, in order to solve the research problem, the method selected for data processing in the project is more suitable for the VADER model. Therefore, the analysis of coherent sentences by the SVM model is affected. Also, the binary label of the SVM model makes the control experiments of the whole project incomplete, and the evaluation of the results is not perfect.

6. Conclusion

The project started by proposing the research question "What factor determines the success of an action movie?" and Chapter 2 made the literature review of the British film industry, including the development of the film industry in the United Kingdom, the evaluation of the British film industry, and the impact of social media and film reviews on the film industry, which conducted that social media has become a communication platform between film companies and audiences, and film reviews played the role as communication medium through direct feedback. Therefore, how to evaluate the reviews on the social medias became the curial problem, and the sentiment analysis was found, which is one the most advanced technologies in the Language Natural Processing (NLP) and is the best suitable for addressing the research question. For introducing and understanding the sentiment analysis, the project began with its development, functionality, and extends to the commercial application of sentiment analysis. To implement sentiment analysis, the project also expounded its establishment process in detail, including data collection, data preprocessing, feature capture, feature selection, model establishment and performance evaluation.

Based on the study of Chapter 2, Chapter 3 described methodology where specified the research question as "Among the four factors of director, plot, actors and music, which factor determines the success of an action movie". To implement the project efficiently, there are two method of data collection are decided to construct alternatively, which are collecting data from public dataset Kaggle or through the web scraping. Furthermore, two models were utilized to construct the VADER model and the SVM model, where the former answered the research question essentially, and the latter would be the standard against which the main model of the project is compared. To enable the model to analyze the data effectively, data preprocessing including text preprocessing, feature extraction, and feature selection will be carried out before modeling. The output of the VADER model includes the sentiment classification of each review as well as the correlation between the four factors (plot, music, director, and actors) and the overall rating of each film. The SVM model, meanwhile, serves two purposes. One is to assess the effectiveness of the two models by comparing them to the VADER, and the other is to confirm the reliability of the VADER study results. When the answer to the research question was constructed, the defined methods of correctness verification and performance measurement were constructed, in which the credibility of the model and the availability of the results were stated.

Chapter 4 took the designed methodology into the case study, where introduced how to realize the project objectives. In general, a high-quality dataset that matched the project specifications was firstly obtained from Kaggle and used for the project. To assure the accuracy and authenticity of the data, some extraneous elements in the dataset were deleted during the implementation phase. The feature extraction process, which turns the unstructured input into vectors and incorporates the BoW model, kept the most crucial features after feature selection. All of the aforementioned actions served as basic setups for model development, in which Python creates the SVM model and the VADER model. The SVM is a standard binary classification model, hence it does not directly address the research question. Instead, a sentiment classifier (positive or negative) was developed derived from the ratings of the dataset to mark each review, and the dataset was divided into training set and testing set for model creation. The researchers first investigated the potential link between four factors and movie success using the SVM model. Moreover, the output of the VADER model provided a direct response to the research topic.

Chapter 5 made use of the conclusions of Chapter 4 to verify the correctness of the two models, and visualize the model output for analysis. The contents of the analysis include: (1) Answering the research question: In action films, the plot is the aspect that the audience pays the most attention to, but it is the actors and music, and finally the director. (2) Compared with the experimental results of Wang *et al.* (2020a), the conclusions obtained by the project are different, which may be caused by the fact that the gerne of movies are limited in the project, and the method of data processing has changed. (3) Because the two models of the project performed well in the evaluation, they can be used successfully to assist the production companies to improve the quality of their action films. (4) Since the analysis of film reviews can be compared to reviews in other industries, the models in the project can be flexibly improved and applied to other fields. However, there is also limitation proposed, which provide ideas for the future direction of the project.

References

Abbasi, A., Chen, H. & Salem, A., (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM transactions on information systems (TOIS), 26*(3), 1-34.

Abd El-Jawad, M. H., Hodhod, R. & Omar, Y. M., (2018). Sentiment analysis of social media networks using machine learning.

Aggarwal, C. C., (2015). Mining text data.

Agrawal, R., (2021). Must Known Techniques for text preprocessing in NLP. Available from: https://www.analyticsvidhya.com/blog/2021/06/must-known-techniques-for-text-preprocessing-in-nlp/ (Accessed 18 April 2022).

Ahmad, I. S., Bakar, A. A. & Yaakub, M. R., (2020). Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews. *Information Processing & Management, 57*(5), 102278.

Ahmad, S. R., Bakar, A. A. & Yaakub, M. R., (2019). A review of feature selection techniques in sentiment analysis. *Intelligent data analysis, 23*(1), 159-189.

Airani, R. & Karande, K., (2022) How social media effects shape sentiments along the twitter journey? A Bayesian network approach. *Journal of Business Research*, 142, 988-997.

Al Amrani, Y., Lazaar, M. & El Kadiri, K. E., (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127, 511-520.

Arnheim, R., (1957). *Film as Art: 50th anniversary printing*. Univ of California Press.

Ashari, A., Paryudi, I. & Tjoa, A. M., (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications (IJACSA), 4*(11).

Avinash, M. & Sivasankar, E., (2019). A Study of Feature Extraction Techniques for Sentiment Analysis.

Ayata, D., Yaslan, Y. & Kamasak, M. E., (2020). Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems. *Journal of Medical and Biological Engineering, 40*(2), 149-157.

Bae, G. & Kim, H.-J., (2019). The impact of movie titles on box office success. *Journal of Business Research*, 103, 100-109.

Barr, R. S., Golden, B. L., Kelly, J. P., Resende, M. G. & Stewart, W. R., (1995). Designing and reporting on computational experiments with heuristic methods. *Journal of heuristics, 1*(1), 9-32.

Bastı, E., Kuzey, C. & Delen, D., (2015). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems*, 73, 15-27.

Basuroy, S., Chatterjee, S. & Ravid, S. A., (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of marketing, 67*(4), 103-117.

Batrinca, B. & Treleaven, P. C., (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY, 30*(1), 89-116.

Berelson, B., (1952). Content analysis in communication research.

Bhowmik, N. R., Arifuzzaman, M. & Mondal, M. R. H., (2022). Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13, 100123.

Bilal, M., Israr, H., Shahid, M. & Khan, A., (2016). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences, 28*(3), 330-344.

Bird, S., Klein, E. & Loper, E., (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Birjali, M., Beni-Hssane, A. & Erritali, M., (2016). A method proposed for estimating depressed feeling tendencies of social media users utilizing their data.

Birjali, M., Beni-Hssane, A. & Erritali, M., (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 113, 65-72.

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

Boiy, E. & Moens, M.-F., (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval, 12*(5), 526-558.

Bräysy, O. & Gendreau, M., (2005). Vehicle Routing Problem with Time Windows, Part I: Route Construction and Local Search Algorithms. *Transportation Science, 39*(1), 104-118.

British Film Institute, (2016). UK film – the vital statistics. Available from: https://web.archive.org/web/20111211150021/http://www.ukfilmcouncil.org.uk/vitalstats (Accessed 9 February 2022).

British Film Institute, (2020). Film distributors ranked by market share in the United Kingdom (UK) and Republic of Ireland in 2019, based on box office gross. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/238258/market-share-of-film-distributors-in-the-uk-and-ireland/ (Accessed 3 March 2022).

British Film Institute, (2021a). Film industry revenue in the United Kingdom (UK) from 1998 to 2019 (in million GBP). Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/297182/film-industry-revenues-in-the-uk/ (Accessed 09 March 2022).

British Film Institute, (2021b). Film industry revenue in the United Kingdom (UK) from 2009 to 2019, by platform (in million GBP). Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/238890/film-revenues-in-the-uk-by-platform/ (Accessed 10 March 2022).

British Film Institute, (2021c). Number of film and video production companies in the United Kingdom (UK) from 1996 to 2019. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/239180/number-of-film-and-video-production-companies-in-the-uk/ (Accessed 10 March 2022).

Casaló, L. V., Flavián, C., Guinalíu, M. & Ekinci, Y., (2015). Do online hotel rating schemes influence booking behaviors? *International Journal of Hospitality Management*, 49, 28-36.

Ceron, A., Curini, L. & Iacus, S. M., (2015). Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. *Social Science Computer Review, 33*(1), 3-20.

Chakraborty, K., Bhattacharyya, S., Bag, R. & Hassanien, A. A., (2019). 7 - Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Techniques. In: Dey, N., Borah, S., Babo, R. & Ashour, A. S., eds. *Social Network Analytics*. Academic Press: 127-147.

Chakravarthy, S., (2020). Tokenization for Natural Language Processing. Available from:

https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4 (Accessed 18 April 2022).

Chen, K. & Yin, J., (2017). Information competition in product launch: Evidence from the movie industry. *Electronic Commerce Research and Applications*, 26, 81-88.

CineVue, (2020). The True Cost of Movie-making. Available from: https://cine-vue.com/2020/11/the-true-cost-of-movie-making.html (Accessed 15 March 2022).

Cordeau, J.-F., Gendreau, M., Laporte, G., Potvin, J.-Y. & Semet, F., (2002). A guide to vehicle routing heuristics. *Journal of the Operational Research society, 53*(5), 512-522.

Dang, N. C., Moreno-García, M. N. & De la Prieta, F., (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics, 9*(3), 483.

Deepanshi, (2021). Text Preprocessing in NLP with Python codes. Available from: https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/ (Accessed 18 April 2022).

Demchenko, A., (2020). ParseHub Review. Available from: https://data-ox.com/parsehub-review/ (Accessed 19 May 2022).

Dhar, V. & Chang, E. A., (2009). Does Chatter Matter? The Impact of User-Generated Content on Music Sales. *Journal of Interactive Marketing, 23*(4), 300-307.

Duan, Y., Liu, T. & Mao, Z., (2022). How online reviews and coupons affect sales and pricing: An empirical study based on e-commerce platform. *Journal of Retailing and Consumer Services*, 65, 102846.

Edington, T., (2022). Downing Street parties: What COVID rules were broken? Available from: https://www.bbc.co.uk/news/uk-politics-59577129 (Accessed 10 March 2022).

Eliashberg, J., Elberse, A. & Leenders, M. A., (2006), The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing science, 25*(6), 638-661.

Eliashberg, J. & Shugan, S. M., (1997). Film critics: Influencers or predictors? *Journal of marketing, 61*(2), 68-78.

Erdmann, A., Arilla, R. & Ponzoa, J. M., (2022). Search engine optimization: The long-term strategy of keyword choice. *Journal of Business Research*, 144, 650-662.

Esteban-Bravo, M. & Vidal-Sanz, J. M., (2021). *Marketing Research Methods: Quantitative and Qualitative Approaches*. Cambridge: Cambridge University Press.

European Audiovisual Observatory, (2019). Consumer revenues for on-demand audiovisual services in the United Kingdom (UK) from 2011 to 2018 (in million euros) [online] Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/438718/online-vod-film-consumer-spending-uk/ (Accessed 14 March 2022).

Ezaldeen, H., Misra, R., Bisoy, S. K., Alatrash, R. & Priyadarshini, R., (2022). A hybrid E-learning recommendation integrating adaptive profiling and sentiment analysis. *Journal of Web Semantics*, 72, 100700.

Fan, Z.-P., Che, Y.-J. & Chen, Z.-Y., (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research*, 74, 90-100.

Film Blerg, (2019). The Increasing Importance of Film And Television Show Reviews. Available from: http://www.filmblerg.com/2019/08/15/the-increasing-importance-of-film-and-television-show-reviews/ (Accessed 16 March 2022).

Gao, K., Xu, H. & Wang, J., (2015). A rule-based approach to emotion cause detection for Chinese micro-blogs. *Expert Systems with Applications, 42*(9), 4517-4528.

Gascón, J., Bernal, P., Román, E., González, M., Giménez, G., Aragón, Ó., Roé, L., López, E., Rodríguez, J. & Morales, C., (2016). Sentiment analysis as a qualitative methodology to analyze social media: study case of tourism. *CIAIQ2016*, 5.

Gavilan, D., Avello, M. & Martinez-Navarro, G., (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, 66, 53-61.

GeeksforGeeks, (2019). Bag of words (BoW) model in NLP. Available from: https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/ (Accessed 22 May 2022).

GeeksforGeeks, (2020). Word Embeddings in NLP. Available from: https://www.geeksforgeeks.org/word-embeddings-in-nlp/ (Accessed 20 April 2022).

GeeksforGeeks, (2022). NLP – Expand contractions in Text Processing. Available from: https://www.geeksforgeeks.org/nlp-expand-contractions-in-text-processing/ (Accessed 20 May 2022).

Gomery, D., (2015). Film and Video Industry. In: Wright, J. D., ed. International Encyclopedia of the Social & Behavioral Sciences (Second Edition). *Oxford: Elsevier*, 156-159.

Goode, W. J. & Hatt, P. K., (1952). Methods in social research.

GOV.UK., (2021). Have you consumed digital films illegally in the past three months. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/291300/illegal-digital-film-consumption-in-the-united-kingdom/ (Accessed 15 March 2022).

Govindan, V. & Balakrishnan, V., (2022). A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection. *Journal of King Saud University - Computer and Information Sciences*.

Gupta, I. & Joshi, N., (2020). Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. *Journal of intelligent systems, 29*(1), 1611-1625.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H., (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

He, Y. & Zhou, D., (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management, 47*(4), 606-616.

Hearst, M. A., (1992). Direction-based text interpretation as an information access refinement. *Text-based intelligent systems: current research and practice in information extraction and retrieval*, 257-274.

Hoque, N., Bhattacharyya, D. K. & Kalita, J. K., (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications, 41*(14), 6371-6385.

Hu, N., Bose, I., Koh, N. S. & Liu, L., (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674-684.

Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V. & Niemann, H., (2000). Recognition of emotion in a realistic dialogue scenario.

Hutto, C. & Gilbert, E., (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Jacobs, J. P., Shahian, D. M., Prager, R. L., Edwards, F. H., McDonald, D., Han, J. M., D'Agostino, R. S., Jacobs, M. L., Kozower, B. D. & Badhwar, V., (2015). Introduction to the STS National Database Series: outcomes analysis, quality improvement, and patient safety. *The Annals of thoracic surgery, 100*(6), 1992-2000.

Jain, P. K., Pamula, R. & Srivastava, G., (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, 41, 100413.

Jindal, K. & Aron, R., (2021). A systematic study of sentiment analysis for social media data. *Materials Today: Proceedings*.

Joachims, T., (1998). Text categorization with support vector machines: Learning with many relevant features.

Johnson, D., (2022). NLTK Tokenize: Words and Sentences Tokenizer with Example. Available from: https://www.guru99.com/tokenize-words-sentences-nltk.html (Accessed 19 May 2022).

Joseph, J., Vineetha, S. & Sobhana, N. V., (2022). A survey on deep learning based sentiment analysis. *Materials Today: Proceedings*.

JustWatch, (2021). Number of titles available on Netflix, Amazon Prime Video, NowTV, and Disney+ in the United Kingdom (UK) as of February 23, 2021, by type. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/1107383/netflix-amazon-prime-now-tv-disney-number-of-titles-uk/ (Accessed 10 March 2022).

Kaggle, (2022). Kaggle Community Guidelines. Available from: https://www.kaggle.com/community-guidelines (Accessed 28 April 2022).

Karthik, R. V., Ganapathy, S. & Kannan, A., (2018). A Recommendation System for Online Purchase Using Feature and Product Ranking.

Kasri, M., Birjali, M. & Beni-Hssane, A., (2019). A comparison of features extraction methods for Arabic sentiment analysis.

Kessler, B., Nunberg, G. & Schütze, H., (1997). Automatic detection of text genre. *arXiv preprint* cmp-lg/9707002.

Khan, Z., Khan, S. M., Dey, K. & Chowdhury, M., (2019). Development and evaluation of recurrent neural network-based models for hourly traffic volume and annual average daily traffic prediction. *Transportation Research Record, 2673*(7), 489-503.

Kharde, V. & Sonawane, P., (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint* arXiv:1601.06971.

Kim, D., Park, K. & Bang, Y., (2022). The effect of TV drama piracy: An analysis of digital piracy users, internet buzz, and TV drama viewership. *Information & Management, 59*(2), 103599.

Kim, E. & Kim, S., (2017). Online movie success in sequential markets: Determinants of video-on-demand film success in Korea. *Telematics and Informatics, 34*(7), 987-995.

Kohli, A. K. & Jaworski, B. J., (1990). Market orientation: the construct, research propositions, and managerial implications. *Journal of marketing, 54*(2), 1-18.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y. & Alsaadi, F. E., (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86, 105836.

Kumar, R. & Kaur, J., (2020). Random forest-based sarcastic tweet classification using multiple feature collection. In: *Multimedia Big Data Computing for IoT Applications*. Springer: 131-160.

Kwartler, T., (2017). The OpenNLP project. *Text Mining in Practice with R*, 237-269.

Leavy, P., (2014). *The Oxford handbook of qualitative research*. Oxford University Press, USA.

Liang, H., Sun, X., Sun, Y. & Gao, Y., (2017). Text feature extraction based on deep learning: a review. *URASIP journal on wireless communications and networking, 2017*(1), 1-12.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D., (2014). The Stanford CoreNLP natural language processing toolkit.

Medhat, W., Hassan, A. & Korashy, H., (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Mei, J. & Frank, R., (2015). Sentiment crawling: Extremist content collection through a sentiment analysis guided web-crawler.

Middleton, M., (2020). How film producers take advantage of IMDb? Available from: https://thecustomercollective.com/how-film-producers-take-advantage-of-imdb/ (Accessed 18 May 2022).

Mikolov, T., Chen, K., Corrado, G. & Dean, J., (2013a). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781,

Mikolov, T., Le, Q. V. & Sutskever, I., (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint* arXiv:1309.4168,

Moraes, R., Valiati, J. F. & Gavião Neto, W. P., (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*(2), 621-633.

Naresh, E., Vijaya Kumar, B., Niranjanamurthy, M. & Nigam, B., (2019). Challenges and issues in test process management. *Journal of Computational and Theoretical Nanoscience, 16*(9), 3744-3747.

Nasreen Taj, M. B. & Girisha, G. S., (2021). Insights of strength and weakness of evolving methodologies of sentiment analysis. *Global Transitions Proceedings, 2*(2), 157-162.

Ngoc, P. V., Ngoc, C. V. T., Ngoc, T. V. T. & Duy, D. N., (2019). A C4. 5 algorithm for English emotional classification. *Evolving Systems, 10*(3), 425-451.

NLTK, (2022). Natural Language Toolkit. Available from: https://www.nltk.org/ (Accessed 12 June 2022).

Pang, B. & Lee, L., (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval, 2*(1–2), 1-135.

Park, D.-H. & Kim, S., (2008). The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications, 7*(4), 399-410.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. & Roth, R., (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic.

Pearce, J., (2022). PIRACY SURGES DURING CORONAVIRUS LOCKDOWN. Available from: https://www.ibc.org/piracy-surges-during-coronavirus-lockdown/5820.article (Accessed 15 March 2022).

Pentheny, J. R., (2015). The Influence of Movie Reviews on Consumers Available from: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1267&context=honors (Accessed 16 March 2022).

Pereira, R. & Tam, C., (2021). Impact of enjoyment on the usage continuance intention of video-on-demand services. *Information & Management, 58*(7), 103501.

Perez, M., (2019). Web Scraper Tutorial: How to Easily Scrape any Website for Free. Available from: (Accessed 17 April 2022).

Pinto, A., Gonçalo Oliveira, H. & Oliveira Alves, A., (2016). Comparing the performance of different NLP toolkits in formal and social media text.

Pipis, G., (2020). How to Run Sentiment Analysis in Python using VADER. Available from: https://python-bloggers.com/2020/10/how-to-run-sentiment-analysis-in-python-using-vader/ (Accessed 24 April 2022).

Polanyi, L. & Zaenen, A., (2006). Contextual valence shifters. In: *Computing attitude and affect in text: Theory and applications*. Springer: 1-10.

Raj P M, K. & Sai D, J., (2022). Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques. *Materials Today: Proceedings*, 51, 576-584.

Rana, S. & Singh, A., (2016). Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques.

Refaeilzadeh, P., Tang, L. & Liu, H., (2009). Cross-validation. *Encyclopedia of database systems*, 5, 532-538.

Rodrigues, R., Camilo-Junior, C. G. & Rosa, T., (2018). A taxonomy for sentiment analysis field. *International Journal of Web Information Systems, 14*(2), 193-211.

Rundell, M., (2005). MacMillan English-Chinese dictionary for advanced learners.

Sadhukhan, S., Banerjee, S., Das, P. & Sangaiah, A. K., (2018). Producing better disaster management plan in post-disaster situation using social media mining. In: *Computational Intelligence for multimedia big data on the cloud with engineering applications*. Elsevier: 171-183.

Salience Search Marketing, (2021). Online product review reading behavior among consumers in the United Kingdom (UK) in 2021. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/statistics/1226424/online-review-reading-behavior-in-the-uk/ (Accessed 16 March 2022).

Sánchez-Marono, N., Alonso-Betanzos, A. & Tombilla-Sanromán, M., (2007). Filter methods for feature selection–a comparative study.

Sauravkaushik, (2016). Introduction to Feature Selection methods with an example (or how to select the right variables?). Available from: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/ (Accessed 20 April 2022).

Scikit Learn, (2022a). sklearn.model_selection. GridSearchCV. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=gridsearchcv#sklearn.model_selection.GridSearchCV (Accessed 26 April 2022).

Scikit Learn, (2022b). sklearn.decomposition.PCA. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (Accessed 26 April 2022).

Siddharth, M., (2021). Feature Extraction and Embeddings in NLP: A Beginners guide to understand Natural Language Processing. Available from: https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/ (Accessed 20 April 2022).

Simonton, D. K., (2011). Film. In: Runco, M. A. & Pritzker, S. R., eds. *Encyclopedia of Creativity (Second Edition)*. San Diego: Academic Press: 509-515.

Statista, (2014). Five most popular film genres in the United Kingdom as of October 2013*. Available from: https://www.statista.com/statistics/286476/most-popular-film-genres-united-kingdom-uk/ (Accessed 14 April 2022).

Statista, (2021). Number of social network users in the United Kingdom from 2017 to 2026 (in millions). Available from: www-statista-com.pugwash.lib.warwick.ac.uk/statistics/553530/predicted-number-of-social-network-users-in-the-united-kingdom-uk/ (Accessed 16 March 2022).

Statista, (2022). Sources of information about products in the UK in 2022. Available from: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/forecasts/997830/sources-of-information-about-products-in-the-uk (Accessed 27 June 2022).

Stegner, B., (2020). IMDb vs. Rotten Tomatoes vs. Metacritic: Which Movie Ratings Site Is Best? Available from: (Accessed 17 April 2022).

Sudhir, P. & Suresh, V. D., (2021). Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings, 2*(2), 205-211.

Swarnkar, N., (2020). VADER Sentiment Analysis: A Complete Guide, Algo Trading and More. Available from: https://blog.quantinsti.com/vader-sentiment/ (Accessed 23 April 2022).

Tabinda Kokab, S., Asghar, S. & Naz, S., (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14, 100157.

TextBlob, (2022). TextBlob: Simplified Text Processing. Available from: https://textblob.readthedocs.io/en/dev/ (Accessed 12 June 2022).

The Numbers, (2022). Movie Production Countries. Available from: https://www.the-numbers.com/movies/production-countries/ (Accessed 9 February 2022).

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S. & Moore, J. H., (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85, 189-203.

Vanaja, S. & Belwal, M., (2018). Aspect-level sentiment analysis on e-commerce data.

Vashishtha, S. & Susan, S., (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138, 112834.

Verma, V., (2020). A comprehensive guide to Feature Selection using Wrapper methods in Python. Available from: https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/ (Accessed 21 April 2022).

Volpi, G. F., (2019). 6 amateur mistakes I've made working with train-test splits. Available from: https://towardsdatascience.com/6-amateur-mistakes-ive-made-working-with-train-test-splits-916fabb421bb (Accessed 25 April 2022).

Wang, G., Sun, J., Ma, J., Xu, K. & Gu, J., (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57, 77-93.

Wang, Y., Shen, G. & Hu, L., (2020a). Importance Evaluation of Movie Aspects: Aspect-Based Sentiment Analysis.

Wang, Y., Wang, M. & Fujita, H., (2020b). Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems*, 190, 105030.

Weng, J., (2019). NLP Text Preprocessing: A Practical Guide and Template. Available from: https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79 (Accessed 18 April 2022).

Xu, Q. A., Chang, V. & Jayne, C., (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, 3, 100073.

Yoo, G. & Nam, J., (2018). A hybrid approach to sentiment analysis enhanced by sentiment lexicons and polarity shifting devices.

Zhao, Z., Rao, G. & Feng, Z., (2017). DFDS: A Domain-Independent Framework for Document-Level Sentiment Analysis Based on RST.

Abbreviation

NLP: Natural Language Processing

SA: Sentiment Analysis

SVM: Support Vector Machine

VADER: Valence Aware Dictionary and Sentiment Reasoner

Appendix

Appendix A: Certificate of Research Integrity


OXFORD UNIVERSITY PRESS

epigeum

Certificate

Number: 1066390291

This is to certify that

Yuan XU

of University of Warwick

Successfully completed the course

Research Integrity: Concise (core course)

as part of the Epigeum Online Course System with a score of 84%.

Dated: 30 April 2022

Copyright Oxford University Press 2022

Appendix B: Certificate of Completing the Ethics Training Courses

Grade item	Calculated weight	Grade	Range	Percentage	Feedback	Contribution to course total
Data Protection (UK GDPR) and Information Security Refresher						
Data Protection (UK GDPR) and Information Security Refresher - Start Here - 80% required to pass	100.00 %	80.00	0–100	80.00 %		80.00 %
Course total Simple weighted mean of grades.	-	**80.00**	**0–100**	**80.00 %**		-

Grade item	Calculated weight	Grade	Range	Percentage	Feedback	Contribution to course total
Information Security Smart						
Information Security Smart	100.00 %	80.00	0–100	80.00 %		80.00 %
Course total Simple weighted mean of grades.	-	**80.00**	**0–100**	**80.00 %**		-

Appendix C: Ethical Approval

2022/6/12 20:15 Email - XU, YUAN (PGT) - Outlook

Ethical Approval Confirmation

wmg-ftmasters@warwick.ac.uk <wmg-ftmasters@warwick.ac.uk>
Thu 2022-05-26 12:05
To: XU, YUAN (PGT) <Yuan.Xu.5@warwick.ac.uk>
Cc: Zhang, Jiayi <J.Zhang.37@warwick.ac.uk>

WARWICK
THE UNIVERSITY OF WARWICK

Ethical Approval Not Required

26 May 2022

Dear Mr Xu,

Warwick ID Number: 2119738

This is to confirm that your Supervisor's Delegated Approval form has been received by the WMG Full-time Master's Office, confirming that your project: Sentiment Analysis on Action Movie Reviews **does NOT require ethical approval.**

You are reminded that you must now adhere to the answers and detail given in the completed WMG SDA ethical approval form (and associated documentation) within your research project. If anything changes in your research such that any of your answers change, then you must contact us to check if you need to reapply for or update your ethical approval before you proceed.

If your data collection strategy changes substantially prior to or during data collection (e.g. you decide to conduct any interviews, surveys, focus groups or anything similar involving human participants or their data), then you MUST stop data collection and reapply for ethical approval before your changes are implemented.

When you submit your project please write N/A against the ethical approval field in the submission pro-forma and include a copy of this email in the appendices of your project.

Best Wishes

Lucy Inman
WMG Full-Time Master's Office
wmg-ftmasters@warwick.ac.uk
go.warwick.ac.uk/wmgftmsc
+44(0)24 7657 4206

https://outlook.office.com/mail/inbox/id/AAQkADAxYTcyYmM1LWEwMjAtNGM3Yy1iZGFmLTllYTFhNGM3ZjMwMQAQAPGaFrOkl%2FpAj%2FzbNlX… 1/1

Appendix D: Data Collection Examples

id number	original title	category	irector/creat	movie lenght	movie date	author	review date	review title	review	label	mean of stars	umber of review	full reviews average stars
0	Lekin...	Drama	Gulzar	ours 51 minu	1990	piyush-336-4	8-Apr-12	n digest a the	t can be hear	8	9.16	6	7.5
1	Lekin...	Drama	Gulzar	ours 51 minu	1990	cseabhi	10-Jul-21	ar's haunting	nd no one els	9	9.16	6	7.5
2	Lekin...	Drama	Gulzar	ours 51 minu	1990	abbott32	2-Jul-04	to have been	d, periodicall	9	9.16	6	7.5
3	Lekin...	Drama	Gulzar	ours 51 minu	1990	monoglot	29-Nov-05	guing story w	dy desert pro	9	9.16	6	7.5
4	Lekin...	Drama	Gulzar	ours 51 minu	1990	Kammu	27-Nov-99	It's a classic	one end and t	10	9.16	6	7.5
5	Lekin...	Drama	Gulzar	ours 51 minu	1990	esskaydas	23-Oct-03	storyline anc	ani culture ar	10	9.16	6	7.5
6	e Lemon Sist	Comedy	Joyce Chopra	our 33 minut	1989	Chaves7777	7-Dec-06	In the name c	s great music	9	9.5	2	5
7	e Lemon Sist	Comedy	Joyce Chopra	our 33 minut	1989	poshjosh	23-Sep-99	d this is my s	ave laughed a	10	9.5	2	5
8	sley: The Que	Drama	chard Michae	our 34 minut	1990	rrobinson-41(	24-Oct-19	vie is not play	this movie ar	2	7.13	15	6.6
9	sley: The Que	Drama	chard Michae	our 34 minut	1990	noonspinner5	16-Oct-09	vorld where t	friendships w	5	7.13	15	6.6
10	sley: The Que	Drama	chard Michae	our 34 minut	1990	ndSAboutMov	29-Jul-20	Goofy	irector Richar	5	7.13	15	6.6
11	sley: The Que	Drama	chard Michae	our 34 minut	1990	mskinne22	4-Mar-05	nance as the	s or still is. si	6	7.13	15	6.6
12	sley: The Que	Drama	chard Michae	our 34 minut	1990	Sylviastel	21-Jan-08	of Mean Reig	f many years	7	7.13	15	6.6
13	sley: The Que	Drama	chard Michae	our 34 minut	1990	Goingbegging	29-Dec-14	eople and Big	a through he	7	7.13	15	6.6
14	sley: The Que	Drama	chard Michae	our 34 minut	1990	orefane-4116	2-Jul-19	e female Don	nances from	7	7.13	15	6.6
15	sley: The Que	Drama	chard Michae	our 34 minut	1990	mark.waltz	26-May-02	a about a ver	h prettier anc	8	7.13	15	6.6
16	sley: The Que	Drama	chard Michae	our 34 minut	1990	MarieGabriell	30-Nov-06	e Pleshette sl	n each one ..	8	7.13	15	6.6
17	sley: The Que	Drama	chard Michae	our 34 minut	1990	classicalsteve	12-Jul-18	onald Trump:	leshette is pe	8	7.13	15	6.6
18	sley: The Que	Drama	chard Michae	our 34 minut	1990	wgranger	2-May-17	ely a guilty pl	explore how	8	7.13	15	6.6
19	sley: The Que	Drama	chard Michae	our 34 minut	1990	rgodfried	15-Jun-12	Good movie	an in a hotel,	8	7.13	15	6.6
20	sley: The Que	Drama	chard Michae	our 34 minut	1990	mellymariefl	21-Jan-06	kinne2 is righ	ould have stil	9	7.13	15	6.6
21	sley: The Que	Drama	chard Michae	our 34 minut	1990	celebrityroyalt	27-Jul-16	rmance by S	work in this	9	7.13	15	6.6
22	sley: The Que	Drama	chard Michae	our 34 minut	1990	Roman11	23-Mar-05	What a show!	were all note	10	7.13	15	6.6
23	e gui xing do	Action	Benny Wong	our 27 minut	1990	charlesws	29-Jul-05	ong Kong acti	ong story sho	8	8	1	5.8
24	r Wicked Wa	Drama	chard Michae	our 35 minut	1991	gadawag	28-Apr-21	Great movie!	cklear! I'm f	7	8.5	2	5.5
25	r Wicked Wa	Drama	chard Michae	our 35 minut	1991	pat-176	7-Jul-02	al life intrigu	ssful. In the p	10	8.5	2	5.5
26	Lethal Games	Action	John T. Bone	our 23 minut	1991	eofwine_dra	9-Aug-16	orst of the w	n the film's p	1	1	1	4.3
27	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	lidiermustntd	15-Aug-10	1 out of 10..	work herself.	1	5.83	6	6.6
28	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	OllieSuave-00	22-Jan-16	se and full of	pearances di	6	5.83	6	6.6
29	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	hite_fang-84	7-Apr-18	t Joyce Gode	rs she tries t	6	5.83	6	6.6
30	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	eofwine_dra	15-Oct-17	lominated st	ots of jealous	7	5.83	6	6.6
31	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	phillip-58	30-Aug-08	e for Joyce G	e film but or	7	5.83	6	6.6
32	ang jia nu jia	Action	Corey Yuen	our 27 minut	1990	sarastro7	18-Dec-06	n doesn't let	were two ac	8	5.83	6	6.6
33	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	LydiaOLydia	11-Oct-09	as Low Budg	efore I go an	3	7	6	6
34	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	sonnyschlaeg	15-Jun-07	ave been mu	ers don't pret	6	7	6	6
35	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	pete36	25-May-19	of a less-kno	lled last U-bo	8	7	6	6
36	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	benkidlingtor	21-Jun-07	while U-boa	U-boat" can't	8	7	6	6
37	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	ArtV∫s	7-Sep-98	way to help	the U-boat re	8	7	6	6
38	s letzte U-Bo	Drama	Frank Beyer	our 39 minut	1993	I.K	14-Aug-99	opeless missi	o is involved	9	7	6	6
39	u jai yim taa	Fantasy	Ngai Choi Lam	our 30 minut	1990	dbborroughs	29-Dec-08	Boring	ed Amy Yip in	2	6.55	11	5.7