Åbo Akademi University

BIBEK KADEL

Sentiment Analysis on YouTube Comments: Analysis of prevailing attitude towards Nokia Mobile Phones

Master's Thesis on Information Systems
Supervisor: Professor Jozsef Mezei
Faculty of Social Sciences, Business and Economics, and Law (FSEJ)
Åbo Akademi University (ÅAU), 2023

Subject: Information Systems
Writer: Bibek Kadel
Title: Sentiment Analysis on YouTube Comments: Analysis of prevailing attitude towards Nokia Mobile Phones
Supervisor: Professor Jozsef Mezei
Abstract: The volume of textual data, more specifically, the magnitude of opinionated text on social media, has increased the interest of companies to closely analyze what their customers have to say about them and their products. This thesis explores the possibility of performing aspect-based sentiment analysis with YouTube comments. The comments on Nokia Mobile phones are the subject of the study in this thesis. First, manual labeling was performed to identify the aspect terms and sentiment and then categorize the aspects based on the aspect's functionality on the phone. From the categorization, it was found out that people mainly have shown negative sentiment towards multiple aspects of the phone with maximum negative attitude towards the price of the phone. On the other hand, the only aspect that could gather a positive attitude was the phone's-built quality. The result shows that there are multiple phone aspects that HMD Global can consider for current and future product improvement. Further, this study used the labeled data to perform supervised learning to classify the aspects and the aspect sentiment from the comments. With two features extraction techniques, BoW and TF-IDF, this paper has explored the performance of different machine learning models on YouTube comments. The models show good results for aspect classification; however, the model's performance could be further improved for aspect sentiment classification. Overall, little attention to this area has been discussed because of the complexity, highly unstructured, and noisy nature of text on YouTube. However, despite the challenges, this platform can be valuable in producing insightful analysis, as presented in this thesis.
Keywords: Sentiment Analysis, Aspect-based sentiment analysis (ABSA), YouTube comments, Aspects Terms, Aspect Category, Aspect Sentiment, Nokia Mobile phones, Machine Learning Models
Date: 10th May 2023
Number of Pages: 98

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
BoW	Bog of Words
CRF	Conditional Random Field
CV	Cross-Validation
DT	Decision Tree
FM	Frequency Modulation
FN	False Negative
FP	False Positive
HD	High Definition
HMM	Hidden Markov Model
ID	Identification Document
IDF	Inverse Document Frequency
IOB	Inside Outside Beginning
JST	Joint model of Sentiment and Topic
KNN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LTSM	Long Short-Term Memory
ME	Maximum Entropy
ML	Machine Learning
NB	Naïve Bayes
NBSVM	Naïve Bayes Support Vector Machine
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OS	Operating System
POS	Part of Speech
RF	Random Forrest
RNN	Recurrent Neural Network
RQ	Research Question
Sd	Secure Digital
CNN	Convolution Neural Network

SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency Inverse Document Frequency
TN	True Negative
TP	True Positive
URL	Uniform Resource Locator
USB	Universal Serial Bus

1. INTRODUCTION

The advent of the internet and its rapid growth in recent years has completely changed how people communicate and interact. Furthermore, the rise of social media has accelerated the progress of creating and sharing content through the internet in a much larger and broader way. People can choose from various internet sites and platforms to interact and share their feelings and opinions regarding any topic. Further, interaction has never been so easy with the increase in size and influence of social media platforms such as Facebook, YouTube, Twitter, Instagram, TikTok, and other social networking sites. Millions of people use numerous internet sites and thus produce a magnitude of user-generated opinionated content. This opinionated content now has gathered significant interest from various fields, including academics and research, as this data significance in producing valuable insights is very high.

Initially, obtaining customers' opinions towards specific products or services was limited to surveys and questionaries and required manual review. However, today, the user-generated opinionated content present on the internet has been at the heart of understanding public opinion regarding specific products, services, or a topic. Opinionated content is of great significance, in particular to companies. Companies must examine and understand what customers and consumers think about their products and services in today's global and competitive business environment. A clear understanding of how people perceive their likes and dislikes can significantly impact a company's market presence regarding its growth and survival. Not just limited to companies, emotions, and opinions shared on the internet can be equally significant for individuals in making judgments and decisions. As highlighted by (Schwartz, 1977; Bandura, 1989; Arnold & Silvester, 2005), human behavior is largely influenced by subjective feelings and beliefs such as attitude, emotion, or sentiment. Also, Campbell-Meiklejohn et al. (2010), in their study, highlighted the notion that other people's thinking, opinion, perception, and sentiment significantly influence an individual decision-making process. Therefore, obtaining insights into the emotions and sentiments of others on a specific topic of interest is essential in deriving public opinion toward that particular subject matter.

Social networking sites have been a go-to platform for obtaining such opinionated content as they have a large user base and contain a significant volume of data. These platforms serve as rich data sources composed of different data structures, such as audio, video, images, and text, that form a basis for comprehending users' opinions. One platform that has been at the core of study in understanding opinionated content is YouTube. YouTube is currently one of the most popular social networking platforms worldwide. It allows its users to upload their videos, share the video, and view other video content uploaded by other users.

Further, YouTube allows users to review a video in many ways. They can show their feelings by clicking the 'like' and 'dislike' buttons in the YouTube platform interface. Also, they can share their feelings and opinions towards the video and video content by providing textual feedback called 'comments'. YouTube users can view comments provided by other users and provide their comments or respond to others.

This study focuses explicitly on YouTube in understanding the public opinion of Nokia mobile phones, which is achieved by creating a YouTube comment dataset and performing sentiment analysis on the dataset.

1.1.Background

Sentiment analysis, is used to understand public opinions, sentiments, attitudes, and emotions from textual data (on specific topics). Over the years, sentiment analysis has become one of the most popular fields of study in Natural Language Processing (NLP) because of the ever-increasing amount of online information and its practical application in facilitating decision-making. Sentiment analysis analyses public opinions and sentiments towards products, services, individuals, organizations, issues, events, and topics (Liu, 2012). Sentiment analysis is the subfield of NLP and refers to extracting subjective opinions on a specific subject. The sentiment analysis technique is a technology-based architecture that provides the solution to understanding people's views, reactions, and opinions regarding polarity, e.g., positive, negative, and neutral in textual data (Kumar & Jaiswal, 2019). Knowing users' sentiments regarding a product or service can answer the question of how well the product or the service is doing in the market. For instance, the normal purchasing behavior of an individual when deciding to buy a product is first to gather information about the product and see what others have commented about it. Then, that individual can see other feedback and analyze manually whether the product is good or bad, then make a purchase decision. This is the case on an individual level;

however, when considering this on the organizational level, a company wants to know what customers think about their product. The whole analysis of feedback becomes more complicated as much feedback comes from many customers. Thus, the volume of feedback can be processed with sentiment analysis to harness the critical aspects (Pang & Lee, 2008).

Sentiment analysis on a granular level is categorized into document-level sentiment analysis, sentence-level sentiment analysis, and aspect-level sentiment analysis.

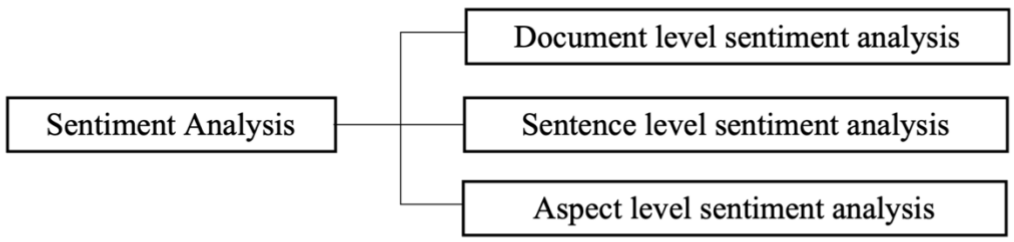

Figure 1: Granularity levels of sentiment analysis

- **Document level:** Document-level sentiment analysis concerns the classification of the entire document as positive, negative, or neutral expressing sentiment. On this level, the understanding is that the analysis assumes that the document articulates opinions or sentiments on one single entity (one product).

  For example, in the following text, *"The new hotel is amazing. It has quite large and open space, the food they serve is great, and service is up to a point",* the sentiment drawn from the entire text is positive as everything being discussed on the text conveys positive opinion towards the hotel.

- **Sentence level:** Sentence level sentiment analysis concerns the task on the sentence level, meaning that the analysis determines whether the sentence expresses a positive, negative, or neutral opinion.

  For example, in the following text, *"The new hotel has a good atmosphere. However, I didn't like their customer service",* the first sentence conveys positive sentiment. In contrast, the second sentence expresses negative sentiment towards the hotel.

- **Aspect level:** The third granularity of sentiment analysis is aspect level sentiment analysis which performs finer-grained sentiment analysis. Aspect-level sentiment analysis is also referred to as feature-level or entity-level sentiment analysis in previous studies (Hu & Liu, 2004; Pang & Lee, 2008; Steinberger et al., 2014). Unlike document and sentence level, aspect-level sentiment analysis goes one step further by examining the opinion itself. The general idea of aspect-level sentiment analysis is that the opinion comprises the target and the sentiment.
  For example, in the following text, *"I like the display of this phone, but I hate those thick bezels."* The user has expressed positive sentiment towards the aspect, 'display', while negative sentiment towards the aspect, '*bezels*'.

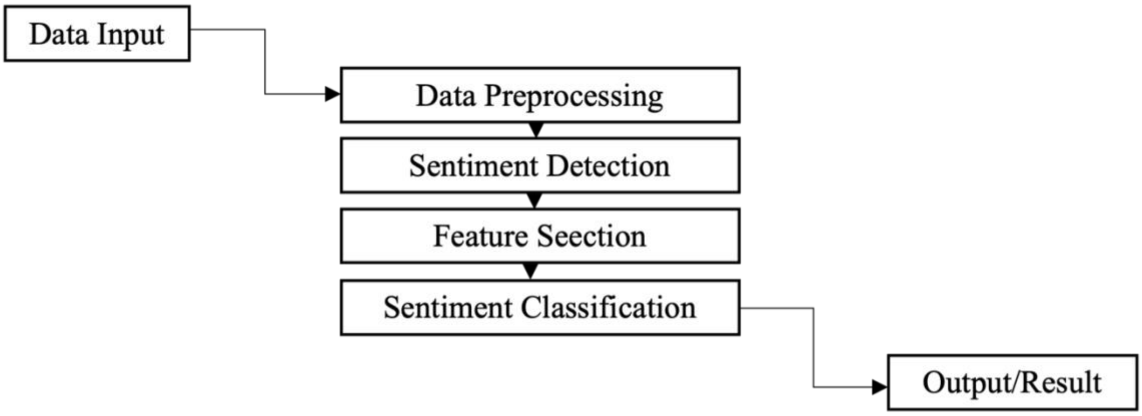

Figure 2: General Workflow of Sentiment Analysis Processing

The idea behind sentiment classification is to label the opinionated text (document, paragraphs, sentences) as positive or negative. Multiple studies in the literature can be found using supervised and unsupervised methods to classify the sentiments. Earlier studies have focused on determining sentiment labels on entire sentences or documents; however, aspect-level sentiment classification is not limited to sentence or document level, but rather it aims to consider aspects in a sentence.

1.2. Purpose and Motivation for the Thesis

Not long ago, companies would spend a significant amount of their resources to gather customer feedback. However, with the growth of the internet and social media platforms, reaching out to customers or getting feedback has become relatively easy. However, though a huge volume of opinionated data is available online, harnessing the essentials from those data is still challenging.

Through sentiment analysis, opinion-rich information can be harnessed from the plethora of data available on the internet. Aspect-based sentiment analysis, a granular level of sentiment analysis especially, can be a valuable tool for companies in understanding users' sentiments towards specific aspects or topics. This approach provides far more specific insights on which companies can build targeted strategies.

YouTube, one of the most popular social media platforms, hosts billions of users and thus contains a tremendous amount of data generated from user interaction. Studies have tried to analyze YouTube data with various tools and techniques, including sentiment analysis. However, aspect-based sentiment analysis, in particular, has received very little attention because of the complexity of the data generated in this platform. Further, the lack of domain-specific labeled data for aspect-based sentiment analysis presents a significant hurdle.

This thesis has developed a manually annotated dataset that takes comments from Nokia Mobile phone videos to address these issues. It has looked into various aspects and their associated sentiments to understand people's current perceptions of the Nokia mobile phone. With a particular reputation for design and dependability, Nokia mobile phones were once the dominant player in the world mobile market share. In 2007, Nokia's smartphone market share was 49.4% (Lee, 2013). However, the company fell into serious turmoil, and Nokia had to offload its mobile production. HMD Global took charge and once again started phone production upholding the heritage of Nokia mobile phones. The idea of understanding public opinion towards the re-emergence of Nokia mobiles and the aspects of today's produced Nokia mobile phones was a strong drive to perform this study. Also, the possibility of contributing to the phone domain labeled data (from the YouTube platform) was a critical component of motivation for this thesis.

1.3.Research Questions and Objectives

The objective of this thesis is to perform aspect-based sentiment analysis with YouTube comments related to Nokia Mobile phones and understand the current perceptions of people towards the phones with regards to their different aspects. The extracted YouTube comments are manually labeled to understand the prevailing discussion on Nokia Mobile phone aspects and the associated sentiments. The labeled dataset is then fed into a machine learning model to understand its quality, as the analysis's scalability depends upon the data's quality. The thesis can contribute to further studies in aspect-based sentiment analysis on YouTube comments related to mobile phones. The labeled dataset can classify aspects and sentiments on new unseen data.

The thesis addresses the study topic by answering the four major research questions. The four research questions established for this thesis are:

RQ 1: What is the present YouTube domain landscape of aspect-based sentiment analysis?

Objective: This research question seeks to understand the overall aspect-based sentiment analysis domain study on YouTube data. Previous studies are reviewed, and a synopsis of the studies carried out so far is presented to answer the question. Further, approaches taken to perform aspect-based sentiment analysis have been looked upon to derive a conclusion on the most used methods in performing aspect-based sentiment analysis with YouTube data.

RQ 2: What aspects have been frequently mentioned in the comments, and what are the sentiments expressed about these aspects?

Objective: The comments from YouTube users are inferred to be feedback given to the product or the company. This question seeks to understand two critical properties of aspect-based sentiment analysis by looking at those comments. Firstly, it aims to find out the most frequent and discussed aspects of Nokia Mobile phones, and secondly, it aims to find out the sentiment or the attitude that YouTube users have shown towards those mentioned aspects. This is achieved by manually annotating comments retrieved from the YouTube platform regarding Nokia Mobile Phones. Through manual annotation, the aspect terms and sentiment towards those aspect terms are extracted.

RQ 3: How does using different feature extraction techniques impact the performance of aspect-based sentiment analysis on YouTube comments?

Objective: This thesis uses two feature extraction techniques: Bag-of-Words (BoW) and Term Frequency and Inverse Document Frequency (TDIF). The performance of these two-feature extraction techniques is analyzed with regard to the YouTube comments dataset used. Furthermore, for each algorithm used, the input features are numerically represented through these two-feature extraction techniques, and the performance of algorithms are evaluated with different performance measures technique.

RQ 4: Which ML models among the chosen ones perform best for this data set?

Objective: Similar to the third research question, this question aims to see the performance of different machine learning models for the labeled YouTube comments dataset in performing aspect-based sentiment analysis. The performance of each machine learning model is evaluated through different performance measurement techniques.

1.4.Thesis Structure

This thesis document comprises six chapters, each exploring and detailing the necessary steps to achieve the objectives set in section 1.3 of this chapter (chapter 1).

Chapter 1- *Introduction*- This chapter introduces the topic of the study, provides background material, establishes the research problem, and explains the research objectives and the motivation for this study.

Chapter 2- *Related Work*- This chapter discusses the state-of-art practices and application of aspect-based sentiment analysis. The primary focus of this chapter is the approaches undertaken in previous studies regarding aspect-based sentiment analysis, particularly with the use of YouTube data in this domain. Furthermore, the chapter discusses the algorithms used in this thesis.

Chapter 3- *Methodology*- This chapter proposes the underlying architecture that has been followed to achieve the thesis objectives. The critical aspects of this chapter are the approaches taken from data collection, aspect terms, and aspect sentiment extraction to data preprocessing and data uses with different algorithms.

Chapter 4- *Results*- This chapter explains the experiments performed with the algorithms and the results obtained with different settings.

Chapter 5- *Discussion*- This section of the thesis explains the experiments performed, analyzes the results, and discusses the results obtained. This chapter describes people's attitudes towards the aspects of Nokia mobile phones and provides a comprehensive performance synopsis from various machine learning algorithms.

Chapter 6- *Conclusion*- This chapter summarizes this study's findings and contribution to the study domain. Also, along with the study's limitations, this section provides a clear explanation of the need for further research with YouTube data for aspect-based sentiment analysis.

Chapter 7- *Limitations of the study*- This chapter discusses some of the limitations of the thesis and outlines the possibility of improvement for future studies.

Chapter 8- *Considerations for Future Work*- This chapter provides suggestions and discussions on the considerations and possibility of future work with aspect-based sentiment analysis with YouTube comments.

2. RELATED WORKS

Though document and sentence-level sentiment analysis have tremendous application use, they are insufficient in explaining the necessary details for application as they do not consider sentiment targets or assign sentiment to the targets (Liu, 2012). If we consider document-level sentiment analysis, a positive document does not necessarily mean that all opinions towards that document's aspects are positive. At aspect level sentiment analysis, the focus is on the opinion itself rather than the construct of the document. Finding out the polarity of the opinions is just not enough, and thus requiring identification of the opinion target is crucial (Steinberger et al., 2014).

The core idea of aspect-based sentiment analysis is to use extracted aspects to characterize the product and the strength of sentiment towards that aspect. This chapter reviews previous studies discussing aspect extraction and suitable methods of sentiment classification to understand the aspect extraction method used in this study. Aspects extraction is a crucial step in aspect-based sentiment analysis as the understanding of public attitude is largely affected by the quality of aspects extracted.

2.1.Aspect-Based Sentiment Analysis

In recent years, aspect-based sentiment analysis has gathered tremendous attention as a large volume of opinionated data from various resources available can be accessed today (Liu, 2012). For example, opinionated data can now be obtained from blogs, e-commerce sites, social media websites, news portals, and many other sources.

Performing aspect-based sentiment analysis requires comprehending and realizing three sub-tasks: aspect term extraction, aspect term categorization, and aspect term sentiment classification. Figure 1 shows the three sub-tasks of aspect-based sentiment analysis.

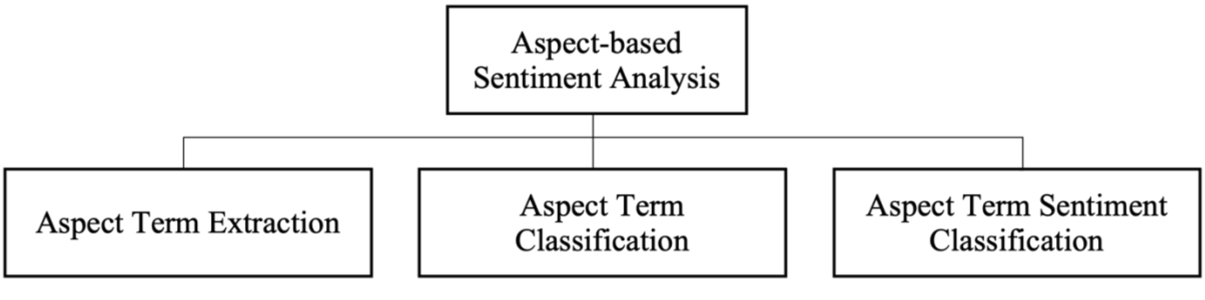

Figure 3: Sub-tasks of aspect-based sentiment analysis

Aspect Term Extraction: Aspect term extraction aims to identify the aspects of a given sentence and is the most critical step in performing aspect-based sentiment analysis (Da'u & Salim, 2019). Türkmen et al. (2016) also stated that aspect extraction is a cornerstone for the powerful development of sentiment analysis systems. Thus, the aspect extraction process needs to be carried out successfully.

Aspect Term Classification: This is the second sub-tasks of aspect-based sentiment analysis, and it deals with clustering synonymous aspect terms into categories. The categories thus formed represent a single aspect, which we refer to as an aspect category (Mukherjee & Liu, 2012).

Aspect Term Sentiment Classification: This is the last task of aspect-based sentiment analysis after aspect term extraction and aspect term classification. Aspect term sentiment classification refers to the identification of the sentiment associated with the aspect terms identified in the first step (aspect-term extraction) of aspect-based sentiment analysis.

A general approach to the three sub-tasks of aspect-based sentiment analysis can be understood from the following two review sentences.

Review sentence 1 (R.S 1): *"The phone's **design** is **great,** but I **hate** the **battery capacity**."*

Review sentence 2 (R.S 2): *"Those **bezels** are **no** for me. Also, I am **not convinced** by the **battery life.**"*

No.of aspects terms in review sentences	Aspect Terms	Aspect Category	Aspect Term Sentiment
1- (for R.S 1)	Design	Design	Positive
2-(for R.S 1)	Battery Capacity	Battery	Negative
1-(for R.S 2)	Bezels	Design	Negative
2- (for R.S 2)	Battery Life	Battery	Negative

Table 1: Aspect terms, Aspect category, and Aspect sentiment classification

The table above explains a general approach to the three sub-tasks of aspect-based sentiment analysis. From the two review sentences, first, the aspect terms were identified *(two aspect terms in each review sentence)*. Then, categorization of those aspect terms was performed based on the synonymous meaning or related meaning of the aspect terms *(design and bezels categorized as "Design" and battery capacity and battery life categorized as "Battery")*. Lastly, the sentiment towards each aspect term was noted based on the opinionated words in the sentence. For instance, for the aspect-term 'battery life', the sentiment shown is negative because of the presence of opinionated word(s) 'not convinced' in the sentence.

2.2.Approaches to Aspect-based Sentiment Analysis

Previous studies on aspect-based sentiment analysis have identified four distinct approaches to address aspect-based sentiment analysis problems: *frequency-based, syntax-based, supervised learning, and unsupervised learning.*

2.2.1. Frequency-Based

The study by Hu & Liu (2004) is considered one of the first studies exploring this aspect-term extraction using a frequency-based approach. The study focused on mining product reviews and provided a comprehensive distinction between review sentences' explicit and implicit aspects. In a given review sentence, aspects can be explicitly mentioned or articulated via other expressions (implicitly). For example, in the sentence *"The design is beautiful, but I hate the battery life of this phone"*, 'design' and 'battery life' as aspects have been explicitly mentioned. However, in the sentence, "*The phone is very expensive*", there aren't any explicitly mentioned aspects. Instead, the aspect 'price' has been expressed with adjectives. Here, the aspect 'price' is implicitly articulated. Hence, aspects in a sentence can also be expressed indirectly through implicit aspects (Chen & Chen, 2016). They only used explicit aspects for the analysis. The research so far in this field has mainly focused on extracting explicit aspects as a standard dataset for testing and evaluating implicit extraction algorithms is not available (Tubishat et al., 2018).

Using the frequency-based approach, Hu & Liu (2004) defined two subsets in their study: product feature set and sentiment set. The product feature set contained the adjectives, and the sentiment set had the sentiments. Further, the product feature set was expanded by adding frequent nouns or noun phrases or the one closest to the opinion word on the review. For the sentiment, they started with 30 positive and negative adjectives as seeds, and later the set was expanded using Wordnet. The Wordnet added the seed words' synonyms to the set. Lastly, the sentiment towards the product aspect/feature identified in the product review sentence was assigned by the dominant sentiment value of the opinion words.

2.2.2. Syntax-Based

Qiu et al. (2011) proposed a syntax-based approach called double propagation. Similar to the studies by Hu & Liu, this study as well used two subsets: opinion target set (aspects) and opinion words (sentiment). Further, these subsets were expanded using the bootstrapping strategy with rules based on dependency relation between words. For instance, in a sentence, if a noun or a noun phrase has a dependency relationship with the opinionated word, the noun or the noun phrase is regarded as the target (aspect). The bootstrapping strategy would stop when no more opinion word or target was identified. This syntax-based approach is greatly dependent on syntax parsing. It thus carries criticism of not performing well when the input data is non-standard, i.e., when the data has spelling or grammar errors.

2.2.3. Supervised learning

Identifying aspects and sentiments can be stated as a sequential labeling problem. Many supervised learning approaches have used sequential labeling methods like Hidden Markov Model and Conditional Random Fields to address the opinion target extraction. The Hidden Markov Model (HMM) represents the probability distributions over a sequence of observations (Rabiner, 1989) and has been applied in POS-tagging and named entity recognition (NER) problems (Liu, 2015). Jin et al. (2009) proposed lexicalized HMM approach to extract aspects and the sentiments words. They used linguistic features and integrated that with the surrounding contextual cues of words for learning. Two tag sets: {Word, POS (Word)} were defined to tag each sentence representing the patterns between the aspect terms and the sentiment or the opinion words. In the tag, the POS (Word) represented the POS of Word. Thus, the idea was to find the appropriate sequence of tags that would maximize the conditional probability. The reported drawback of HMM is that it only works appropriately for sentences with linear sequence structure. The model does not work adequately for review sentences having aspects 2-3 words before or after the associated opinion words. Hence, to address the limitation of HMM, an undirected sequence model was needed.

Conditional Random Fields (CRF) is an undirected sequence model which, unlike HMM, can be arbitrarily structured. In CRF, tokens in a sentence are labeled using the IOB scheme ('B' indicates the entity's beginning, 'O' indicates that the token is outside of the entity, and 'I' indicates inside of the entity. In the CRF model, generally, the features used are the words themselves, POS tags, word stylistics, word dependency relation in the sentence, etc. As the model largely depends on these hand-designed features, it requires a lot of features generating efforts. One of the widely spoken limitations of the CRF model is that it cannot adequately capture long-range dependencies (occurrence of multiple words in between the aspect and the opinion words). A study by Qiu et al. (2011) reported that aspects words and opinion word pairs in a review sentence have long-range dependencies. Li et al. (2010) proposed a structure-aware CRF method that would overcome the long-range dependency issue of the CRM method. They integrated two variations of CRF methods to extract the aspects and the sentiment by providing aspects lists as input. The aware CRF method addresses the problem of long-range dependency by considering conjunctions and deep syntactic aspect dependencies. The learning in the classical CRF method was based on word sequence; however, in the proposed model, the learning was based on linguistic aspect structure.

More recently, the limitations of CRF methods have been addressed by deep neural networks, which have shown considerable improvement in aspect extraction techniques (Poria et al., 2016, Wang et al., 2017). The performance of supervised learning is high; however, the drawback associated with this approach is that it requires annotated data for training the model.

Existing supervised learning approaches like Logistic Regressions, Support Vector Machines, Naïve Bayes, and Neural Networks can be used to predict the sentiment of unlabeled documents. For supervised learning, a labeled dataset is required where aspects and sentiment polarity towards that aspects are highlighted. Once the data is fed into the model, the above algorithms learn the pattern from the training data and then generalize the learned pattern in the new unseen test data.

For aspect-level sentiment classification, similar to sentiment classification problems, neural networks have proven to be very powerful. For example, two single LTSMs were used by Tang et al. (2019), where the model combined left and right contexts to detect the aspects and classify the sentiments associated with the aspects. Another major work in aspect-based sentiment analysis can be said to be of Xue & Li (2018). They were among the first to employ a CNN-based model for aspect-based sentiment analysis. Similarly, a novel approach that could model specific segments of aspect-level sentiment classification in a reinforcement learning environment was proposed by Wang et al. (2019).

2.2.4. Unsupervised learning

The need to label a large amount of data for training purposes and domain restriction problems has seen the emergence of arrays of unsupervised learning approaches based on topic modeling techniques like Latent Dirichlet Allocation (LDA) and its variations for aspect extraction and categorization. LDA so far has found its way into multiple studies (Brody & Elhadad, 2010; Lu et al., 2011). An extended model of LDA called the Joint model of Sentiment and Topic (JST) was proposed by Lin et al. (2012) to detect the topic and the sentiment in the review sentence.

Unsupervised lexicon-based approaches avoid some of the concerns of supervised learning. In the lexicon-based approach, the polarity of the sentiment-carrying word is determined by looking at the sentiment lexicon. The sentiment lexicon contains a list of words associated with sentiment orientation. A sentiment lexicon can be manually created by annotating a large text corpus or using existing resources like SentiWordNet. SentiWordNet has been widely used for aspect-based sentiment analysis; it has high coverage of English terms and sentiment information (Esuli and Sebastiani, 2006).

Once the lexicon is created, this can be used to determine the sentiment of the aspect in the review text by determining the sentiment scores of the associated words. The critical factor for consideration is the determination of the scores. This can be performed in multiple ways. One such method is to sum up the sentiment scores for all the words associated with the aspect and then average to obtain the sentiment score for that aspect. Similarly, a weighted counting approach can also be used. In this approach, words more strongly associated with aspects are assigned higher weights. The lexicon-based method is quite popular as it does not require labeled training data. However, this approach also has limitations, like not being able to handle sarcastic texts, idioms, and other figurative language.

Though the application of aspect-based sentiment analysis is broad and significant, there are still some concerns about performing the analysis as some issues still need to be explained and sorted out. In most studies regarding aspect-based sentiment analysis, it is assumed that pre-specified aspects are derived from the keywords (Wang et al., 2011; Li et al., 2015). The study by Ding et al. (2008) used a lexicon-based approach for aspect-based sentiment analysis, assuming the aspects are known beforehand. Mate (2015) suggested a framework for ranking aspects; however, the study predefined the aspects before the classification. Liu (2012) highlighted that aspect-based sentiment analysis's accuracy is low as the technique still hasn't found its way to dealing with complex sentences.

2.3. Sentiment Analysis and YouTube

With the unprecedented growth of social media, the research community working on sentiment analysis showed interest in more complex data like Twitter tweets (Mishra & Singh 2018), Facebook status and microblog comments (Alfaro et al. 2016), reviews, and YouTube comments. Though there have been proportionally more studies with data from other platforms, YouTube has seen very little exposure in this field. This can be attributed to the type of data generated on YouTube. The comments are noisy and utterly complex. The comments provided by users can be long and unstructured and have multiple aspects or topics in them (Mai & Le, 2020). Below are some studies that have tried to study data from YouTube.

Siersdorfer et al. (2010) performed sentiment analysis by training a classification model to predict the comment ratings. This study studied the connection between the comment ratings and the sentiment terms, and an in-depth analysis was carried out. They used a linear Support Vector Machine and thesaurus to obtain the degree of polarity of each word in the comments. Similarly, to understand the commentator's behavior and attitudes, Schultes et al. (2013) performed classification to categorize YouTube comments into ten different categories. Further, Filippova and Hall (2011) conducted a similar study to construct a text-based classifier to categorize YouTube videos based on the comments.

All these studies provided a significant understanding of the importance of analyzing YouTube comments; however, they did not focus on analyzing YouTube comments' topics. Much of the answer for the neglect of this study was attributed to YouTube comments being full of spam and unrelated to the topics. Uryupina et al. (2014) created a corpus called SenTube, where comments from YouTube were annotated for comment type classification and sentiment detection. Severyn et al. (2014) then used SenTube in their study to remove spam and irrelevant information from the comments.

Poche et al. (2017) studied the behavior of commentators on coding tutorial videos. Two broad categories: Content Concerns and Miscellaneous, were created using SVM and Naïve Bayes model from the collected comments. This study has used some aspects of this study as two classes on top of the three sentiments established in this paper. The details of those classes will be discussed in a later section. Another study on YouTube data was carried out by Madden et al. (2014). They provided a scheme of classifications for YouTube comments in their study. Their study reported that commenting habits of users differ between groups. Their study suggested that some users comment for the promotion, some comment to provide information, and others comment just for pleasure. Their work contributed to providing ten categories and 58 subcategories in classification comments; however, this study's criticism is that it merely provides the schema for the classification.

Savigny and Purwarianti (2017) used YouTube comments to classify emotion in Indonesian. They collected 8,115 comments from 10 different YouTube videos and manually labeled the data with six types of emotions: happy, sad, surprised, disgusted, fearful, and angry. After preprocessing the data, they used four-word embedding techniques (average word vector, average word vector with TDIF, paragraph vector, and CNN). Support Vector Machine (SVM) and unigram with TDIF were used as baseline methods to evaluate the performance of the four-word embedding techniques. The result showed better performance with SVM, with 76.2%.

Similarly, Trinto & Ali (2018) used YouTube comments to perform sentiment analysis on the Bangla language. They manually labeled the collected data into three sets: with three sentiment classes (positive, negative, and neutral), with five sentiment classes (strongly positive, positive, neutral, negative, and strongly negative), and with emotion class (anger, joy, sadness, fear, none). They deployed two deep learning models, LSTM and CNN, to perform sentiment analysis. The set with three sentiment classes achieved the highest accuracy of 65.97%.

Perikos & Hatzilygeroudis (2017) proposed ensemble classifiers on aspect-based sentiment analysis to analyze the aspects from the comments. First, they collected hotel review data from 417 users and manually labeled it. Then, LDA was used to perform topic modeling on the comments. Further, the data was preprocessed, and POS tagging was performed before representing the text as bag of words (BOW). Finally, the Stanford parser was used to determine the dependency of the words in the text. They formulated an ensemble classifier for machine learning using three machine learning algorithms (NB, ME, SVM). The highest performance was obtained with SVM, with both POS and word dependencies utilized. Also, the result indicates that the ensemble classifier outperformed other classifiers by 5.8%.

Marrese-Taylor et al. (2017) used the attention RNN model on YouTube comments to extract the aspects and classify the sentiments. The study discussed a method to mine fine-grained opinions from the closed caption of YouTube videos. The model was trained on YouTube captions and their associated sentiment labels. The study evaluates the model performance with several other benchmark datasets. The result showed the proposed model outperformed the state-of-art method for sentiment analysis.

In the study by Tanesab et al. (2017), they used Support Vector Machine to perform lexicon-based sentiment analysis on YouTube comments with 1000 comments in the Indonesian language to categorize the comments as positive and neutral comments. The performance evaluated with the confusion matrix produced an accuracy of 84%. In the study by Muhammad et al. (2019), a combined method where the machine learning models Naïve Bayes (NB) and Support Vector Machine (SVM) were combined to detect the sentiment in YouTube comments was proposed. The combined model was called NBSVM, and it produced an accuracy of 87%.

In most previous work, the attitude associated with the sentence, or the aspects has been limited to three class systems: positive, negative, and neutral. This study has used two more attitudes on top of these three to depict commentators’ sentiments accurately. The two attitudes the class added are imperative and interrogative. These two classes have been introduced and explained in several studies before but haven’t been studied at the aspect level. Khoo et al. (2006) performed experiments on 14 different sentence classes with multiple models for different sentence classes. Imperative class was discussed in this study.

Similarly, referencing their study, Pokharel & Bhatta (2021) used imperative and interrogative sentence classes to classify YouTube comments into different classes. However, in this thesis, the imperative and interrogative as two classes have been used to define the commentator's attitude towards aspects mentioned along with the positive, negative, and neutral. For instance, in this study, attitude towards an aspect in a comment can be either positive, negative, neutral, imperative, or interrogative. The details of imperative and interrogative attitudes have been explained in the next chapter of this thesis.

2.4.Key factors affecting the choice of Mobile Phones

As this thesis concerns the study of predicting mobile phone aspects and the sentiment towards them, it's important to define the key factors that affect the choice of mobile phone purchase.

Studies performing aspect-based sentiment analysis for mobile phone reviews are relatively low compared to other research exploring the domains like restaurant reviews. However, past studies on restaurant management have uncovered a range of crucial factors such as price, food, variety, reputation, promotion, location, and information sources that drive customer choice of restaurant (Pedraja & Yagüe (2001); Chiciudean et al. (2019); Cullen (2005); Harrington et al. (2011).

Identifying key factors is critical to understanding if the factors discussed in the reviews are the determinant factors. For example, the study by Twenefour (2017) reported that the popularity of the phone, quality of the phone, battery life, affordability, and the presence of more features in the phone are influential factors that people consider before making a purchase decision. Also, the study by Karjaluoto et al. (2005) reported that though the buying choice is very subjective to the buyer, there are some general factors such as price, the brand, interface, and different properties of the phone that influence people buying behavior.

Similarly, Trivedi & Raval (2016) studied the factors influencing student choice in purchasing mobile phones. They reported that design, the latest technology, operating system versions, applications, and hardware features significantly affect their purchase choice. The study by Mack & Sharples (2009) reported that aesthetics, cost, and feature are the factors that have significant implications for the purchase decision of mobile phones. Also, Sata (2013) studied six factors (price, social influence, durability, product feature, brand, and after-sales service) identified from the literature to determine the most influential factors and found the price to be the most significant factor in influencing people to make buying decisions. The product features followed the price and then the durability of the phone. Further, several studies confirmed through their studies that price is one of the most influencing factors to affects the purchase decision of mobile phones (En & Balakrishnan (2022), Rakib et al. (2022), Trivedi & Raval (2016)).

2.5.Aspect Categorization (Labels)

The SemEval dataset is the most referenced dataset for aspect-based sentiment analysis, which contains reviews from e-commerce websites (Li et al., 2019; Chen & Qian, 2020). For instance, data from different domains have been stored in this SemEval dataset.

- SemEval 2014 task 4, ASBA 14 dataset (Pontiki et al., 2014) comprises the data from restaurant and laptop reviews from the e-commerce websites.
- SemEval 2015 task 12, ASBA 15 dataset (Pontiki et al., 2015) extended the work of ASBA 14 from SemEval 2014 by describing the aspect category as entity type combined with the attribute type (Kim et al., 2021).
- SemEval 2016 task 5, ASBA 16 dataset (Pontiki et al., 2016) worked on ASBA 15 from SemEval 2015 by adding new Domains like Hotels, Museums, Electronics, Telecom, and other languages (Kim et al., 2021).

The labels for Mobile Phones were introduced by SemEval 2016 task 5, ASBA 16 dataset. The labels were adopted from the laptop domains, and 17 labels for mobile phones were introduced. The identified 17 labels are explained in the figure below. The mobile phone domain explained in the dataset supported Chinese and Dutch languages.

PHONE, DISPLAY (=screen, touchscreen, including **GRAPHICS**), **BATTERY, CPU** (=processor), **MEMORY** (everything relating to RAM and working memory), **HARD DISC** (everything related to storage and storage capacities), **POWER_SUPPLY, KEYBOARD** (*volume adjustment keys, start button, camera button, virtual keyboard, numpad etc.)*, **MULTIMEDIA_DEVICES, PORTS** (*USB*), **HARDWARE, OS, SOFTWARE, WARRANTY, SHIPPING, SUPPORT, COMPANY.**

Figure 4: Entity and Attributes pair identified in SemEval 2016 Task 5, ASBA 16 by (Pontiki et al., 2016)

Referencing the defined labels from the ASBA 16 dataset, multiple studies have developed the aspect category (labels) for their studies. For instance, in the study by Phan et al. (2021), five annotators were used to label the data based on the defined rules and guidelines for the annotation process. The annotation guideline provided guidance on labeling the mobile phone reviews' aspects and sentiments. Ten aspects and three sentiments were defined for the study. The ten aspects were: "screen", "camera", "features", "battery", "performance", "storage", "design", "price", "general", "service," and "accessories". The three sentiment labels were positive, negative, and neutral.

Similarly, Kim et al. (2021) adopted the same aspect category labels as identified in the study by Phan et al. (2021) in their research to perform span detection in aspect-based sentiment analysis. The labeling was performed by identifying the words or phrases that indicated the opinion of the users regarding the aspect category. So, when a word or a phrase was identified in a review sentence, the aspect category and the sentiment were labeled ASPECT (category)#SENTIMENT.

Further, in the study by Singh & Mishra (2016), they manually identified eight aspects from the mobile review dataset. The eight aspects discussed in the study were "cost", "size", "battery", "camera", "operating system", "processor", "storage", and "screen". They analyzed the reviews from three different phones.

Similarly, Yiran & Srivastava (2019) performed LDA topic modeling on 400,000 Amazon unlocked phone reviews to cluster the topic words in the review sentence and to perform sentiment analysis. Through the LDA topic modeling approach, they were able to find three aspect categories in the study: "screen", "camera", and "battery".

This thesis identifies 14 aspects (category) from the aspect terms extracted during the manual annotation. The 14 aspects were determined by categorizing the aspect terms present in the comments. The presence of these aspects suggests that the aspects and the associated sentiment towards those aspects are significantly relevant to understanding as these aspects have been studied and identified as the key determinant factors in making the purchase decision.

Studies	**Aspect Category**
(Pontiki et al., 2016) (SemEval 2016)	Phone, Display, Battery, CPU, Memory, Hard Disc, Power Supply, Keyboard, Multimedia devices, Ports, Hardware, Operating System, Software, Warranty, Shipping, Support, Company
Phan et al. (2021)	Screen, Camera, Features, Battery, Performance, Storage, Design, Price, General, Service, Accessories.
Kim et al. (2021)	*adopted Phan Et al. (2021) aspect categories* (Screen, Camera, Features, Battery, Performance, Storage, Design, Price, General, Service, Accessories)
Singh & Mishra (2016)	Cost, Size, Battery, Camera, Operating System, Processor, Storage, And Screen.
Yiran & Srivastava (2019)	Screen, Camera, Battery

Table 2: Summary of previous studies defining the aspect categories.

The sentiments towards the key factors or aspects are the key to understanding people's attitudes towards the product. This enables companies to make necessary changes to address the negatives or some expectations shown towards the product. The details on the 14 defined aspect categories from the comments from Nokia Mobile Phone YouTube videos are provided in section 3.3.1.

3. METHODOLOGY

This study uses aspect-based sentiment analysis to understand the public perception of Nokia mobile phones. The research seeks to identify the aspects people mostly talk about and their attitudes toward them. On top of the three most widely used sentiments, positive, negative, and neutral, this study encompasses two classes (imperative and interrogative) to label the data correctly. These two classes were included so that a clearer picture of commentator attitudes could be captured. The fact that commentators not only show a positive or negative attitude towards something but also have many questions, queries, requests, and interest in the topic motivated me to include these two classes in this study. The study presents an overall comparison between the aspect category with the sentiments that people have shown towards those aspects. The diagram for the proposed methodological framework (overview) can be found in the figure below.

The first step in this study is to collect the data. YouTube comments were extracted to perform the analysis. The data were preprocessed before the labeling was done. Labeled data was needed as the study aimed at performing supervised machine learning. Hence, aspect terms, aspect sentiment/class, and aspect category are manually labeled for each comment. The annotation guideline was prepared for the data labeling process. Feature extraction was performed on the cleaned and labeled dataset before running classification models to evaluate the model's performance. Several evaluation metrics like accuracy, precision, F1-score, and Recall have been used to evaluate the classification model in this study.

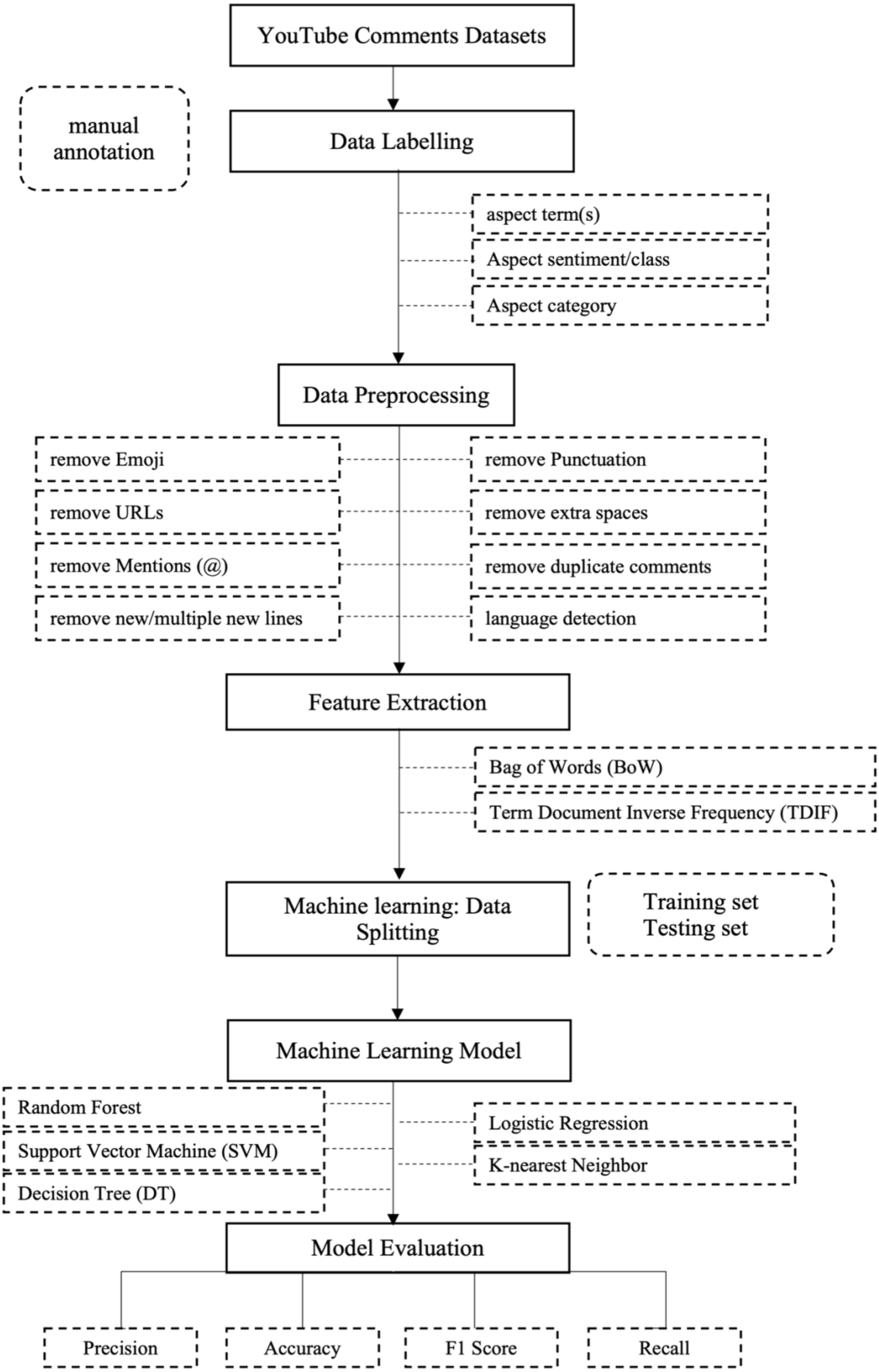

Figure 5: Proposed methodological framework of this thesis

3.1.Data Source

YouTube is one of the most popular social media platforms, where millions of users create and share their content daily. As reported by Kemp (2023), YouTube has over 2.5 billion users, and a single user spends around 23.1 hours on average (on Android phones) using the YouTube app in a month. Because of the volume of users and the interaction thus produced, YouTube has become a rich data source for user-generated content, which can be used to understand public opinion regarding the content discussed on the platform. YouTube's data (user feedback) can be recorded via three different means.

1. If the user likes the content uploaded to the platform, the user gives a thump up, showcasing their positive sentiment towards the content.
2. If the user does not like the content shared on the platform, the user gives a thump-down button indicating a negative opinion.
3. If the users feel the urge to communicate with the creator or other users on the platform, they write their feedback in a textual form. This is referred to as a comment. Comments can be either a parent comment or a child comment. Parent comments are the main comments users give in response to the content of the video. Child comments are the reply made to the parent comments. The comments as well can be liked and disliked by users.

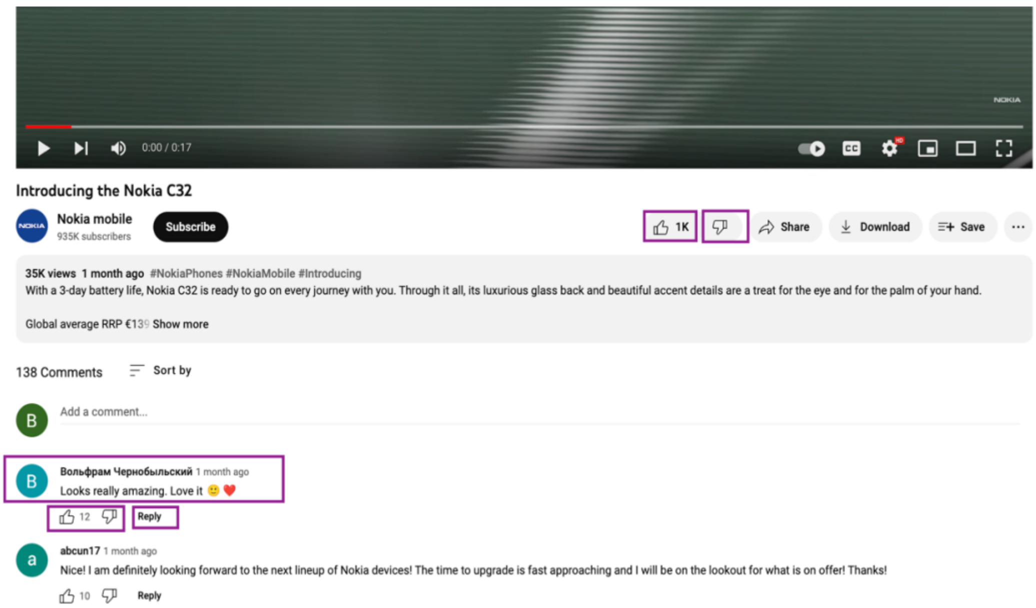

Figure 6: Thumps up, thumps down button and parent-child comment feature in YouTube

Data gathered from these feedback-capturing features of the platform can be used to perform sentiment analysis. Especially the comments, the textual data can be used to understand users' attitudes toward any topic discussed in the video, especially the comments.

3.2.Data Collection

For data collection, this study has used YouTube API to crawl the data from several YouTube videos. The selection of YouTube videos was carried out in two layers. In the first layer, YouTube videos were selected from Nokia Mobile's official YouTube Channel. Videos uploaded in the channel from January 2022 to December 2022 were selected, and videos related to mobile phones were put into the selection pool. A total of 68 videos were listed in the pool in the beginning, and after filtering the videos related to mobile phones, 46 videos were finally selected. In the 46 videos selected, 17 different Nokia mobile phones were discussed. The rest 22 videos were opted out because the content of the video was related to other products like tablets. Using YouTube API, a total of 6,800 comments were extracted. The second data collection layer was performed based on the 16 phone models.

No.	Phone Model
1	Nokia C12
2	Nokia C32
3	Nokia G4005G
4	Nokia C22
5	Nokia 5710
6	Nokia C100
7	Nokia C200
8	Nokia C21
9	Nokia C21 Plus
10	Nokia G100
11	Nokia G11 Plus
12	Nokia G22
13	Nokia G60 5G
14	Nokia X30 5G
15	Nokia G11

No.	Phone Model
16	Nokia G50

Table 3: List of 16 different phone models

For each phone model, a search query on YouTube was made, and the first video on the search was selected for each phone model searched. While searching, out of 16 different phone models, the search result showed three videos from the Nokia Mobile channel itself; hence, those three videos were rejected in this data collection layer as they would have duplication issues. The remaining 13 videos were selected from the pool, and their comments were extracted. A total of 2,937 comments were extracted from this layer. In total, from both layers, a total of 9,737 comments were extracted. The final dataset has two columns, “Comments” and “Phone Models”. The interest of the study is only in the comment column; however, the phone models discussed in the videos are also listed in the dataset for further study if needed. A detailed illustration of the data extraction approach and the selection of comments is provided in the figure below.

The following sections illustrate the methodological approach used in this study to label the data (finding aspect terms, aspect category, and aspect sentiment/class) and perform sentiment analysis classification.

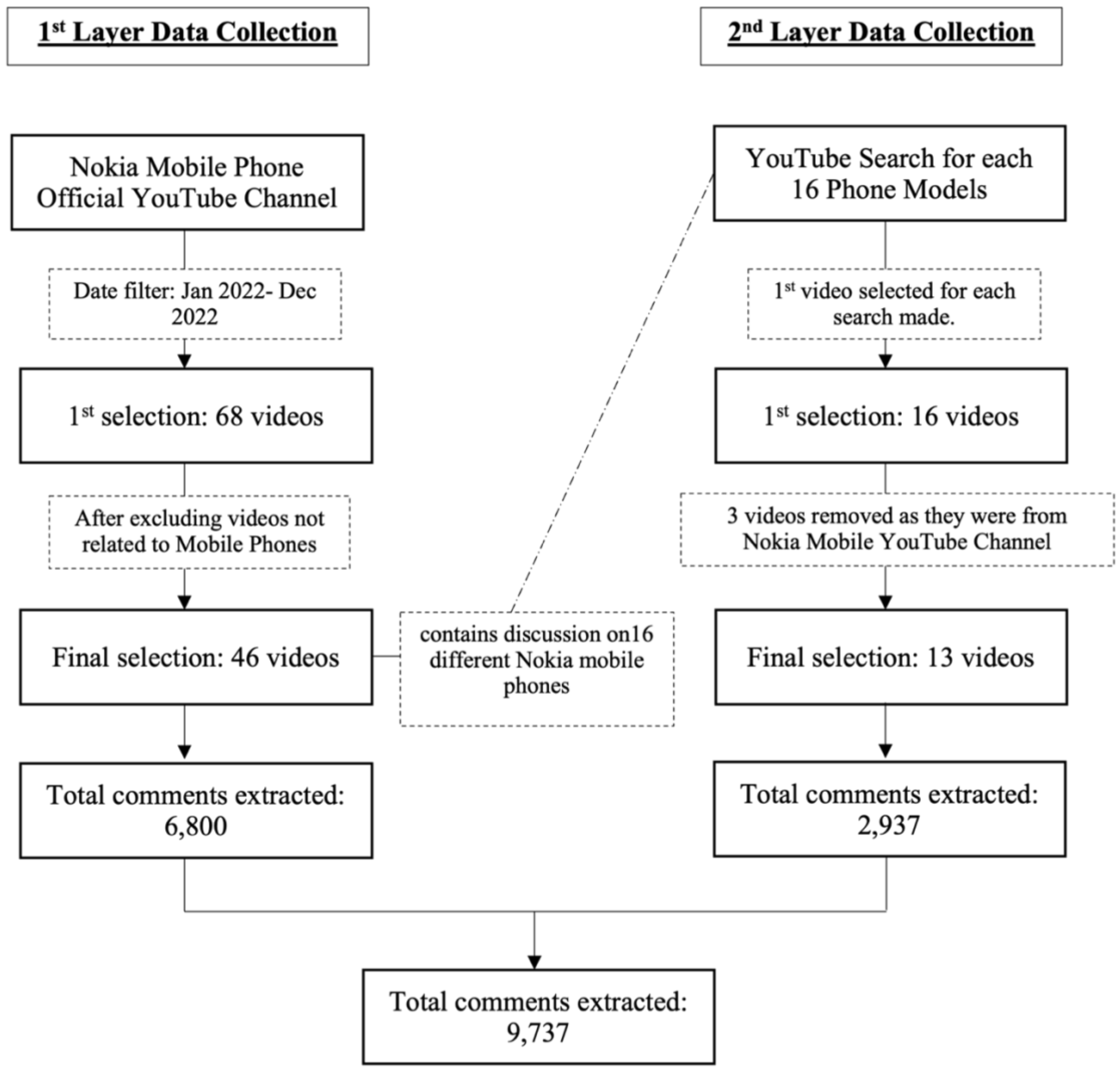

Figure 7: Data Collection Framework of the Thesis

3.3.Data labeling

Data labeling in aspect-based sentiment analysis is a critical step in the whole analysis journey, as the machine learning model's performance heavily depends on the quality of the labeled data. In this study, data labeling was carried out manually and thus was a laborious process to annotate the aspects terms and the aspect sentiment each comment carried. Before labeling the comments, from a quality standpoint, an annotation guideline was prepared for annotating aspect terms and aspect sentiment, which formed the basis of data annotation.

3.3.1. Annotation Guidelines

These annotation guidelines have been prepared to detect aspects, aspect categories, and aspect sentiment. The task of the annotator here is to identify the following types of information from the comments which have been preprocessed.

- Aspect terms: aspects terms in the comments can be single or multi-word terms that name a particular aspect of the entity.
  Example 1: in the sentence, “The *silicone tip* buds are horrible for me. They fall out too easy.”, there is only one but a multi-word aspect: *silicone tip*.
  Example 2: in the sentence, “I love that integration of the *headphones*, so smart.”, the aspect term is *headphones*.

- Aspect term polarity: After identifying aspect terms, each aspect term must be assigned with aspect polarity. The following polarities should be assigned based on the sentiment of the aspects.
  - Positive
  - Negative
  - Neutral

  For instance, in the above examples, in the first comment, the polarity associated with the aspect “silicone tip” is *negative* as the “horrible” word mentioned in the comment denotes the negative opinion of the commenter towards that entity (aspect). Similarly, in the second comment, the aspect term identified is “headphones”, and it has *positive* polarity as the commenter has used the word “love” about the integration of the headphones.

- Aspect class: In this study, along with the three polarities, after seeing the data, the need to include more classes was essential. Thus, from the literature guide, this study has used two more classes, “imperative” and “interrogative”, besides three polarities to annotate aspect terms. The general idea behind including these two classes is that positive, negative, and neutral sentiments are insufficient to explain people’s sentiments when the space for expressing oneself is wide open. Hence, the aspect in the comment in this study can be either positive, negative, neutral, imperative, or interrogative. For an aspect to be imperative, the study assumes the commenter is requesting, making a command, or has an expectation about the aspect.

Example 1: in the sentence, "I wish it would be expanded to *1tb sd card*", the aspect term is "*1tb sd card*". Since this comment reflects more of an expectation of the commentator, the aspect "1tb sd card" has *imperative* class.

Example 2: in the sentence, "Can it play *flac format audio files*?", the aspect term is *"flac format audio files"*. Here, the commentator seeks information, queries, and asks questions. Hence, the aspect term 'flac format audio files' has an *interrogative* class.

- Aspect category: This step is performed only after identifying aspect terms, as this step is classifying the identified aspect terms in the comments. The aspect category in this study has not been predefined. Thus, a grouping of aspect terms based on the closeness of functionality is performed initially, and then a category is formed. A comment sentence can be classified into one or more aspect categories.

  For example, in the sentence, *"The price is great if we are being honest. ram is a bit low, but if you do not use a ton of apps, then you should be fine, which is my case.",* the aspect terms, aspect sentiment/class and aspect category are.

Sentence ID	Sentence	Aspect Term(s)	Aspect Sentiment/Class	Aspect Category
101	the price is great, if we're being honest. ram is a bit on the low side, though.	price	positive	Price
101	the price is great, if we're being honest. ram is a bit on the low side though.	ram	negative	Storage

Table 4: Annotating aspect terms, aspect sentiment/class, and aspect category

As seen in the table above, the aspect terms, aspect category, and aspect term sentiment have been labelled from the dataset. A total of 14 aspect categories were formed by grouping the aspect terms identified. Below is the list of 14 different aspect categories defined for this study.

No	Aspect Category	Sample Aspect Terms
1	design	size, screen, body, bezel, casing, color, design, notch
2	software	android, updates, software, own software, bootloaders, kai os, symbian
3	processor	CPU, processor, chipset, configuration, snapdragon, unisoc
4	camera	front camera, rear camera, megapixel, sensor, camera bump, zeiss lens
5	price	price, affordability, budget, cost, money
6	display	display, refresh rate, amoled, oled, pureview
7	availability	available, release, launch
8	accessories	Headphones, earbud, Bluetooth, Wi-Fi, USB, network
9	media	audio, video, FM, media
10	battery	battery, charging, mah, battery capacity, charging port
11	built quality	Durability, built quality, repairability
12	storage	sd card, storage, memory card, ram
13	sustainability	eco-friendly, environment, recycled plastic, booklets, sustainability
14	service and support	customer care, customer service, customer support, maintenance

Table 5: List of Aspect Category defined in this study

Even though the aspect categories were not predefined, the aspect categories formed in this study are similar to those found in the previous literature. Though some aspects might not seem to be studied in the past, they were covered under other aspects as studies have used different approaches in defining what each aspect category means. Also, some categories in previous studies have terms other than the ones mentioned in the table above, such as “General”. In previous studies, the category “General” is for the general comments made towards the phone. However, this thesis does not define the general category, as all aspect terms are categorized under 14 categories.

3.4. Data Preprocessing

Once the data was extracted and labeled, it needed some preprocessing before moving into the next step. The data extracted from YouTube now is in an unstructured format, which is very difficult to work with. For example, the comments extracted might contain emojis, URLs, mentions, abbreviations, empty spaces, special characters, and duplicates. Therefore, the data needs to be preprocessed and cleaned to further progress with sentiment analysis. Various pre-processing techniques were applied to clean the data and extract appropriate information. Data preprocessing helps achieve higher performance of machine learning algorithms (Rustam et al., 2021).

3.4.1. General Preprocessing techniques

Here below are the steps that were taken during this phase of data pre-processing.

- Removing emojis: delete all the emojis in the comments using regular expressions.
- Removing URLs: delete all the URLs (uniform resource locator) in the dataset.
- Removing mentions: the @ sign and username were deleted from the dataset.
- Removing new lines: the comments containing new lines and multiple new lines were trimmed by removing those new lines.
- Removing non-English comments: comments written in languages other than English were removed.
- Removing punctuation: punctuation in the comments unnecessarily complicates the process; hence, the punctuation was removed from the dataset.
- Removing duplicates: the dataset might contain spam comments; hence, those comments must be removed. Thus, duplicate comments found were removed from the dataset.
- Lowercasing the comments: the words in the comments must be in lowercase as the machine would otherwise read 'camera' and 'Camera' as different words affecting the model's training for the classification. Hence, all the comments were lowercase.

The tables below showcase different preprocessing techniques for the sample comments from the dataset.

Sentence ID	Input Data	After emoji removal
101	🥰😍🤩 stunning design and colors!	stunning design and colors!
102	Thank you @Nokia ♡ Built quality and it's durability is amazing.	Thank you @Nokia Built quality and it's durability is amazing.
103	Your built quality and design is awesome 👌	Your built quality and design is awesome
104	https://youtu.be/Jfqf9s7TRa0	https://youtu.be/Jfqf9s7TRa0
105	بطلوسرفالكم التعانة تنتغلواعلى لهوتف الرائدة و الخرافية	بطلوسرفالكم التعانة تنتغلواعلى لهوتف الرائدة االخرافية
103	Your built quality and design is awesome	Your built quality and design is awesome

Table 6: Sample comments before and after emojis removal

Sentence ID	Input Data	After URL removal
101	stunning design and colors!	stunning design and colors!
102	Thank you @Nokia Built quality and it's durability is amazing.	Thank you @Nokia Built quality and it's durability is amazing.
103	Your built quality and design is awesome	Your built quality and design is awesome
104	https://youtu.be/Jfqf9s7TRa0	
105	بطلوسرفالكم التعانة تنتغلواعلى لهوتف الرائدة و الخرافية	بطلوسرفالكم التعانة تنتغلواعلى لهوتف الرائدة االخرافية
103	Your built quality and design is awesome	Your built quality and design is awesome

Table 7: Sample comments before and after URL removal

Sentence ID	Input Data	After mentions removal
101	stunning design and colors!	stunning design and colors!
102	Thank you @Nokia Built quality and it’s durability is amazing.	Thank you Nokia Built quality and it’s durability is amazing.
103	Your built quality and design is awesome	Your built quality and design is awesome
105	بطلوسرفهلالكم التعلة نتغلواعلى لهوتف الرئدة و الخرافية	بطلوسرفهلالكم التعلة نتغلواعلى لهوتف الرئدة االخرافية
103	Your built quality and design is awesome	Your built quality and design is awesome

Table 8: Sample comments before and after mentions removal

Sentence ID	Input Data	After the removal of non-english comments
101	stunning design and colors!	stunning design and colors!
102	Thank you Nokia Built quality and it’s durability is amazing.	Thank you Nokia Built quality and it’s durability is amazing.
103	Your built quality and design is awesome	Your built quality and design is awesome
105	بطلوسرفهلالكم التعلة نتغلواعلى لهوتف الرئدة و الخرافية	
103	Your built quality and design is awesome	Your built quality and design is awesome

Table 9: Sample comments before and after the removal of non-English text

Sentence ID	Input Data	After removing duplicates
101	stunning design and colors!	stunning design and colors!
102	Thank you Nokia Built quality and it’s durability is amazing.	Thank you Nokia Built quality and it’s durability is amazing.
103	Your built quality and design is awesome	Your built quality and design is awesome
103	Your built quality and design is awesome	

Table 10: Sample comments before and after removing duplicates

Sentence ID	Input Data	After punctuation removal
101	stunning design and colors!	stunning design and colors
102	Thank you Nokia Built quality and it’s durability is amazing.	Thank you Nokia Built quality and its durability is amazing
103	Your built quality and design is awesome	Your built quality and design is awesome

Table 11: Sample comments are removing the punctuations

Sentence ID	Input Data	After lowercasign
101	stunning design and colors	stunning design and colors
102	Thank you Nokia Built quality and its durability is amazing	thank you nokia built quality and its durability is amazing
103	Your built quality and design is awesome	your built quality and design is awesome

Table 12: Sample comments before and after lowercase

3.4.2. Stop Word Removal

Removing stop words from the data also falls under the data preprocessing process. Stop word removal helps enhance the algorithm's learning performance during the data training. Articles such as 'a', 'an', and 'the' and helping words such as 'are', 'am', and 'is' carry no useful information and thus are removed in this process. Similar to these words, other words do not necessarily contain useful information. Thus, in this process, those words are removed from the sentences. Having these stop words in the dataset would mean needing more space and increased processing time for the model. Stop words can be removed using the NLTK in Python, which lists stop words for multiple languages. Further, the list can be modified and customized per specific tasks. For example, in this study, the stop words list from NLTK has been customized to address the need to include certain words as they contribute to the context of the sentence. For instance, stop words such as 'why', 'what', 'when', 'not', and 'don't' have been excluded from the stop word lists as they give meaning to the interrogative and negative attitude in the sentence.

Sentence ID	Input Data	After removing stopwords
101	stunning design and colors	stunning design color
102	thank you nokia built quality and its durability is amazing	thank nokia built quality durability amazing
103	your built quality and design is awesome	built quality design awesome

Table 13: Comment samples before and after data preprocessing

3.4.3. Lemmatization and Stemming

Lemmatization and Stemming are part of data preprocessing techniques where derivates of words are converted to their base words. Performing lemmatization and steaming helps increase text analysis accuracy without creating any learning complexity for the machine learning algorithms. Lemmatization involves reducing a word to its base form. Stemming, on the other hand, involves reducing a word to its base form by removing the suffixes or the prefixes. For example:

Durability > Lemmatization > durable

Durability > Stemming > durabl

In this study, the lemmatization technique has been used for its better performance than stemming. This is because the base form of the word produced from the lemmatization technique is more valid in the language. Stemming might produce truncated words that might or might not be valid in the language. The incorrect word form produced from stemming can affect the accuracy of the model. Furthermore, lemmatization considers the context of the word in the sentence while producing the base form. Hence, for these reasons, lemmatization has been adopted in the study.

3.5. Feature Extraction

Text in raw form cannot be fed into the machine learning models. Computers cannot comprehend textual information. Feature extraction techniques in natural language processing are used to represent words in vectors during text analysis. Hence, with feature extraction techniques, words and documents are represented in numeric vector forms to make them machine-readable. It does this by tokenizing each word in the text as a sequence and converting them into vectors. In this approach, words with similar meanings are given similar vector representations, and the approach aims to preserve the semantic and syntactic meaning of words in the text.

Several feature extraction techniques are available; however, this study uses two feature extraction techniques: BoW and TF-IDF.

The two feature extraction techniques (BoW and TF-IDF) are chosen to compare the performance of the simplest and most effective methods in text classification. BoW approach is the simplest technique to vectorize the data (Xu et al., 2013). It is used in cases where the requirement of the classification task is to identify the presence or the absence of the keywords. On the other hand, TF-IDF is regarded as one of the most effective vectorization techniques for text classification (Salton & Buckley, 1988) and is generally used to show the relative importance of the word in the text document (Abubakar & Umar, 2022).

The idea for selecting these two techniques is to compare the performance of each technique, i.e., comparing the capability of each of the two-feature extraction techniques in understanding whether the context of the words in YouTube comments affects the performance of the machine learning models. Also, the notion of the TF-IDF feature extraction technique performing better than other approaches is tested in this thesis for the YouTube comments dataset.

3.5.1. Bag-of-Words

Bag-of-Words is one the most common feature extraction techniques used in natural language processing to extract features from a given raw text (Rustam et al., 2021). In the BoW model, the sentences or the document are represented as the bag of its words. In this model, only word duplicates are considered, and grammar or the order of the words are ignored (Qader et al., 2019). The model describes the occurrence of the words in the text, and it concerns two key steps, (i) vocabulary of known words and (ii) count of the known words present. The working of this model can be understood through the result representation of two of the sample comments after preprocessing.

Comment (C 1): *thank nokia built quality, durability amazing*

Comment (C 2): *built quality design awesome*

Step 1: Determining the vocabulary

(thank, nokia, built, quality, durability, amazing, design, awesome)

Step 2: Counting the words present in the text

C	thank	nokia	built	quality	durability	amazing	design	awesome
C 1	1	1	1	1	1	1	0	0
C 2	0	0	1	1	0	0	1	1

Table 14: BoW result on the two sample comments.

3.5.2. Term Frequency-Inverse Document Frequency (TF-IDF)

TDIF is a machine learning model used in text analysis by giving scores to the word in the text. It’s a statistical measure of finding the relevance of the words. The model statistically finds word relevance by looking at the occurrence of the words in multiple text documents. The word's relevance is evaluated by multiplying the Term Frequency (TF) metrics with Inverse Document Frequency (IDF) metrics.

The Term Frequency model is used to measure the occurrence of words in a document (Hakim et al., 2014). For instance, let's assume that we have a document 'D1' containing 1000 words, and the word 'camera' is present 15 times. The length of documents may vary from small to large; thus, a word in a large document might appear multiple times compared to documents with a small length. So, to determine the Term Frequency, the occurrence of a word in a document is divided by the total number of words present in the document. Therefore, in this example, the term frequency of the word 'camera' in the document 'D1' is

$$TF = 15/1000 = 0.015$$

The whole idea with Inverse Document Frequency is that some words in the document might occur more frequently and might contain no or significantly less information compared to some words that might appear less in the documents. It can be observed that when performing Term Frequency, the model treats all words equally, even the stop words if present. For instance, if a stop word 'not' is present in the document 1500 times, the Term Frequency model determines the Term frequency for the word 'not', which is not informative to the task. This is where the notion of IDF comes in. IDF model assigns low weightage to those frequent words and high weightage to the rare words. For example, we have 15 documents, and the word 'design' appears in 5 of those documents. Therefore, the inverse document frequency of the word 'design' is

$$IDF = log\_e\ (15/5) = 0.477$$

To determine the TF-IDF in this case, we have to multiply the TF and IDF, i.e., 0.015 * 0.477, which is 0.0072. The TD-IDF score for each word in the preprocessed comments are:

D	**thank**	**nokia**	**built**	**quality**	**durability**	**amazing**	**design**	**awesome**
C 1	0.176	0.176	0.117	0.117	0.176	0.176	0.000	0.000
C 2	0.000	0.000	0.176	0.176	0.000	0.000	0.176	0.176

Table 15: TF-IDF results on the two sample comments.

3.6. Machine Learning Algorithms

Several machine learning algorithms can be applied when working with feature-level sentiment analysis. The learning algorithms can be classified into three sub-categories: supervised learning algorithms, unsupervised learning algorithms, and semi-supervised or hybrid learning algorithms. This section will only discuss four different types of supervised learning algorithms, as only a supervised learning approach has been adopted in this study. The machine learning models used for the classification in this study are Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor.

Optimizing the hyperparameters can enhance the results of the machine-learning models (Pokharel & Bhatta, 2021). Hence, to understand how different parameters significantly affect the performance of the machine learning models, this study has experimented with different subsets of those parameters for each model using a grid search approach.

3.6.1. Logistic Regression

Logistic regression is one of the most common machine learning algorithms used for regression and classification. Logistic regression is based on linear regression; however, the difference is that linear regression is used for solving regression problems while logistic regression is used for classification problems. The model typically uses the Sigmoid function to map out the output, which must be either categorical or discrete. The output value produced is between 0 and 1 and represents the probability of the output class. The mathematical function that represents logistic regression is (Aslam et al., 2022):

$$P = \frac{1}{1 + e^{-}\,(a + bX)}$$

where,

'P': represents the output which is the probability between 0 and 1

'e': is the mathematical constant that represents Euler's number (base of natural logarithms)

'a' and 'b': are the parameters

'X': is independent variable

3.6.2 Support Vector Machine (SVM)

A Support Vector Machine is a supervised machine learning algorithm that functions the task of regression, non-linear classification, and outlier detection (Bennett & Campbell, 2000). Initially, SVM was proposed to perform binary classification by Cortes & Vapnik (1995); however, the model has now been expanded to multi-class classification. This classifier model works by distinguishing the best possible boundary called hyperplane between data points of different classes. Therefore, the model aims to find the best hyperplane for separating the classes with maximum margin (Neethu & Rajasree, 2013). The margin is the determined by computing the distance between the hyperplane and the closet data points of each class. So, when the margin is low, meaning when the data points are closer to the hyperplane, they are more likely to be misclassified. Hence, the model works on finding the maximum margin.

3.6.3. Decision Tree

A decision Tree is also one of the most commonly used and powerful supervised machine learning algorithms used for prediction and classification tasks (Charbuty & Abdulazeez, 2021). The decision tree model is constructed by repeatedly dividing the data according to the split criteria. The decision tree model consists of nodes and branches, where nodes represent the test on a specific feature while the branches represent the possible outcome of the test. There are different types of nodes in a decision tree (Tan et al., 2016). The node at the top of the decision tree is called the root node, and it represents the entire dataset, while the node at the bottom of the tree is called the leaf and represents the prediction.

The most informative features must be selected to construct a decision tree. For this, entropy (uncertainty in a dataset) and information gain (a measure of entropy) with respect to the target variable should be calculated for each feature (Aslam et al., 2022). Entropy and information gain are used to determine the best feature, which is then used to split the data at each node.

3.6.4. Random Forest

Random Forest is a supervised learning algorithm used for classification and regression (Zahoor et al., 2020). While dealing with classification problems, the model handles the categorical data; for regression problems, the model handles continuous data. This model combines multiple decision trees, each giving the class prediction (Naeem et al., 2022). The class with the maximum prediction is the model's prediction for the classification.

The model uses a method called bagging, where each decision tree is allowed to select the sample dataset randomly. The aim is to have different types of trees to increase the model performance by increasing the accuracy and minimizing the variance. Further, in the ensemble method, feature randomness is used to increase diversification and decrease correlation with trees (Aslam et al., 2022).

3.6.5. K-Nearest Neighbor

K-Nearest Neighbor is a supervised machine-learning model that predicts the class by calculating the distance between the training data points and the test data (Harrison, 2019). The model is used for regression and classification (Sarker, 2021). The distance between the training data points and the test data is computed using Euclidean distance (Aslam et al., 2022), the most commonly used distance metric in computing the similarity between two instances (Nguyen et al., 2016). Other measures like Hamming and Manhattan distance can also be used to compute the distance. Once the Euclidean distance is computed, the best K data points are selected, and the model performs the prediction using the most repeated classes in K data points.

The working mechanism of this model is that it figures out the nearest neighbor of the new data. For instance, if the $k = 3$, the model checks the three closest neighbors in the training set and assigns the new data the class of the majority of the k nearest neighbor.

3.7.Evaluation Parameters

Different evaluation metrics could be used to measure the performance of machine learning algorithms. The evaluation metrics help analyze and compare the performance of the machine learning models. They further assist in decision-making. In this study, the performance evaluation of the classifiers used is evaluated with accuracy, precision, recall, and F1 score evaluation metrics.

When discussing these evaluation metrics, the confusion matrix must first be understood. A confusion matrix is an error matrix that indicates four quantities: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Table 12 shows the confusion matrix where each row represents actual labels, and each column represents predicted labels (Landy & Szalay, 1993).

		Predicted Labels	
		Negative	**Positive**
Actual Labels	**Negative**	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 8: The confusion matrix

Reading the Confusion Matrix

To apprehend the measure of the evaluation matrices, the four elements of the confusion matrix must be understood (Rokach & Maimon, 2006).

- True Negative (TN): when a review initially labeled as negative is predicted negative by the classifier too.
- True Positive (TP): when a review initially labeled as positive is predicted positive by the classifier too.
- False Positive (FP): when a review initially labeled as negative is indicated positive by the classifier.
- False Negative (FN): when a review initially labeled as positive is labeled negative by the classifier.

3.7.1. **Accuracy**

Accuracy is one of the most widely used performance measure matrices for classification algorithms in machine learning models that predicts the ratio of true (correct) predictions to the total predictions. The accuracy of a classifier is determined as follows:

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN}$$

3.7.2. **Precision**

Precision is another evaluation measure used to measure classification algorithms' performance. It focuses on measuring the accuracy of the classifier in predicting the positive class. This evaluation measure shows the ratio of the predicted positive class over the initially positive class.

$$Precision\ (P) = \frac{TP}{TP + FP}$$

3.7.3. **Recall**

Another evaluation metric that measures the performance of classification algorithms is recall. It is the ratio of the true positive class to the total positive class. Recall can be determined by the mathematical formula given below.

$$Recall\ (R) = \frac{TP}{TP + FN}$$

3.7.4. **F1-score**

F1-score also measures the performance of classification algorithms. Recall as an evaluation metric uses both precision and recall and thus is regarded as more significant than precision and recall alone (Bruce et al., 2002). F1- score is the harmonic mean of the precision and the recall (Aslam et al., 2022).

$$F1 - score(F) = \frac{Precision(P) * Recall\ (R)}{Precision(P) + Recall\ (R)}$$

4. RESULTS

This section of the thesis presents an overview of the aspects and the sentiment towards those aspects that YouTube users have mentioned in the comment section of the YouTube platform. It also explains the results of different machine learning models for classification tasks.

The thesis aims to analyze the YouTube comments; therefore, we must train our classifiers with labeled data, i.e., comments with aspect category and sentiment labels. The dataset used to analyze the comments is from different YouTube videos (comments extracted in two layers: details in Chapter 3). A total of 9,737 comments were extracted from various YouTube videos. The comments were annotated with aspect terms, aspect categories, and aspect sentiment with the help of annotation guidelines. After labeling the data, the dataset underwent data preprocessing techniques. The final dataset contained 3,877 comments. However, as the thesis aims at identifying sentiment at the aspect level, each comment might include multiple aspects. Out of 3,877 labeled comments, around 38% had more than one aspect discussed. Therefore, the number of unique comments in the dataset was 2396. The number of unique aspect terms in the labeled dataset was 1,599. The table below details the annotation process.

Description	**Details**
Number of labeled comments	3877
Comments with multiple aspects	38 %
Number of unique comments in the dataset	2396
Number of unique aspect terms	1599
Number of aspect category class	14
Number of aspect sentiment class	5

Table 16: Labelled Data description

All experiments run for this thesis were performed using Python programming language and executed using Jupyter Notebooks. The comments were extracted from the YouTube platform using YouTube API. The “pandas” library was used to import the extracted data. The preprocessing of the data (emojis removal, punctuation removal, mentions removal, new lines removal) was performed using regular expressions. Stop words were removed using the ‘nltk’ library: however, the default stop words were customized to fit the thesis scope. Similarly, only English comments were used in this study, and the filtering was performed using the "langdetect" library in Python. The comments were transformed into their base form using the lemmatization method, and "WordNetLemmatizer" was used to achieve the base form of each word in the comments.

The results of research questions 2, 3, and 4 are presented in this section.

4.1. Identifying aspects and associated sentiment to explain the public perception of different Nokia Mobile phone aspects.

The second research question of this thesis aims to understand the aspects and the associated sentiment that people mentioned in comments for Nokia Mobile phone YouTube videos.

- Aspect Terms and Aspect Category

As this thesis performed the manual annotation, the mobile phone aspects from the comments were manually identified. Initially, aspect terms from the comments were identified. After the identification of aspect terms, they were grouped into multiple categories. Finally, the categorization was performed based on the feature that aspect terms explained. For instance, the size of the phone, screen, color, notch, etc., were grouped into the category "design". The details on the categorization of aspects are explained in Chapter 3.

Altogether, 14 aspects were created for all the aspect terms labeled in the dataset. The figure below illustrates the number of aspect terms in each category. As can be seen, 570 aspect terms were categorized under the "design" aspect category. Similarly, 535 aspect terms in the "software" aspect category, 446 aspect terms in the "software" aspect category, 446 aspect terms in the "processor" aspect category, 317 aspect terms in the "camera" aspect category, 305 aspect terms in the "price" aspect category, 298 aspect terms in the "display" aspect category, 279 aspect terms in the "availability" aspect category, 270 aspect terms in the "accessories" aspect category, 242 aspect terms in the "media" aspect category, 236 aspect terms in the "battery" aspect category, 196 aspect terms in the "built quality" aspect category, 103 aspect terms in the "storage" aspect category, 41 aspect terms in the "sustainability" aspect category and 39 aspect terms in the "service and support" aspect category were categorized.

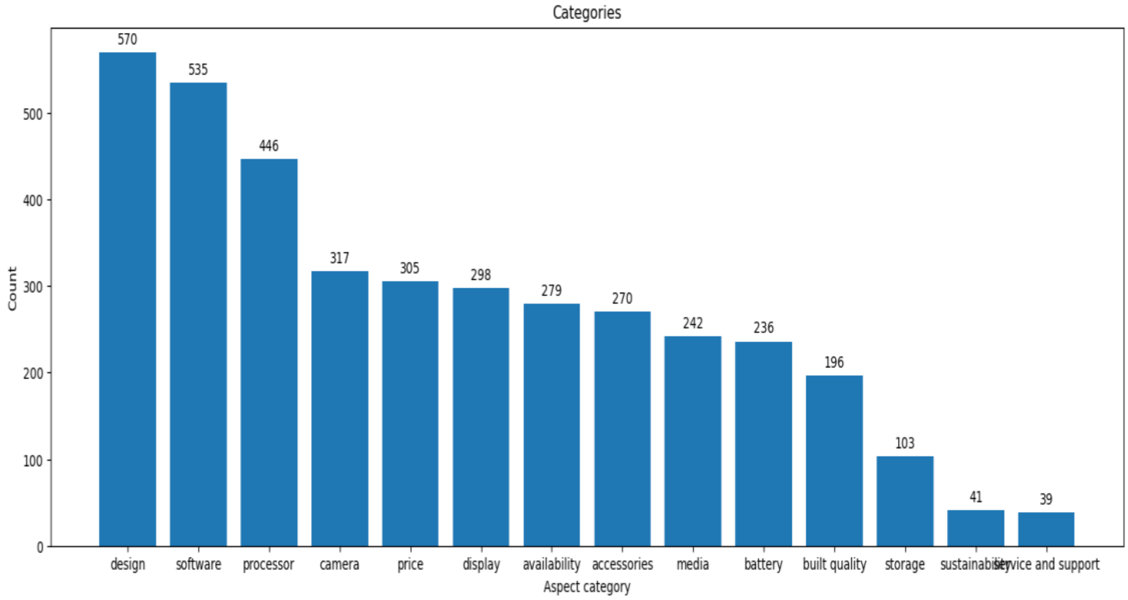

Figure 9: 14 aspect categories constructed from the labeled aspect terms

- Aspect Sentiment

Similar to aspect terms, aspect sentiments were manually labeled for the YouTube comments dataset. An associated sentiment was labeled for each aspect term identified in the comment. A total of five aspect sentiment classes (positive, negative, neutral, imperative, and interrogative) were created to explain people's attitudes toward the mentioned aspect in the comment. A detailed explanation of how each sentiment class was identified has been provided in Chapter 3.

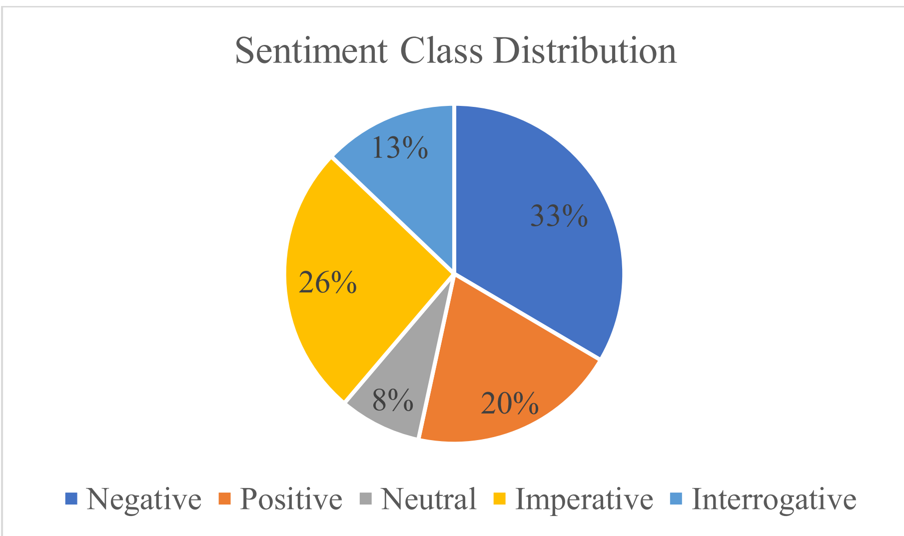

Figure 10: Sentiment Class Distribution

A total of 3,877 comments were labeled; hence, the total comments with sentiment labels are 3,877 as well. As can be inferred from the chart above, the YouTube comments dataset had a maximum number of "negative" sentiments or attitudes (33%) from people towards the mobile phone aspect they were referring to. The "imperative" sentiment class was the second highest, with 26% of overall sentiment labeled. The aspect terms with positive and interrogative sentiment labels in the dataset were 20% and 13%, respectively. The comments with aspect terms labeled with neutral sentiment were only about 8% of the dataset.

The absolute number for each sentiment class label is given in the chart below:

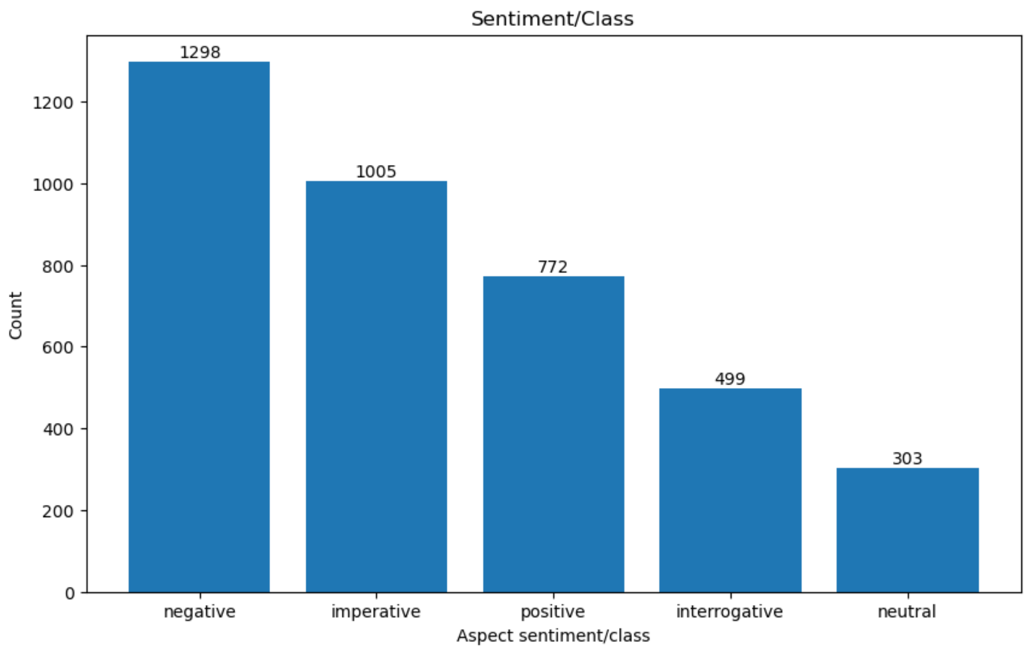

Figure 11: Aspect Sentiment Class Distribution

- Aspect and the associated sentiment

As stated earlier, this thesis identified 14 aspect category classes and five sentiment class labels. One of the objectives of this study was to understand what people are discussing about Nokia Mobile phones and their attitude towards that discussed aspect of the phone.

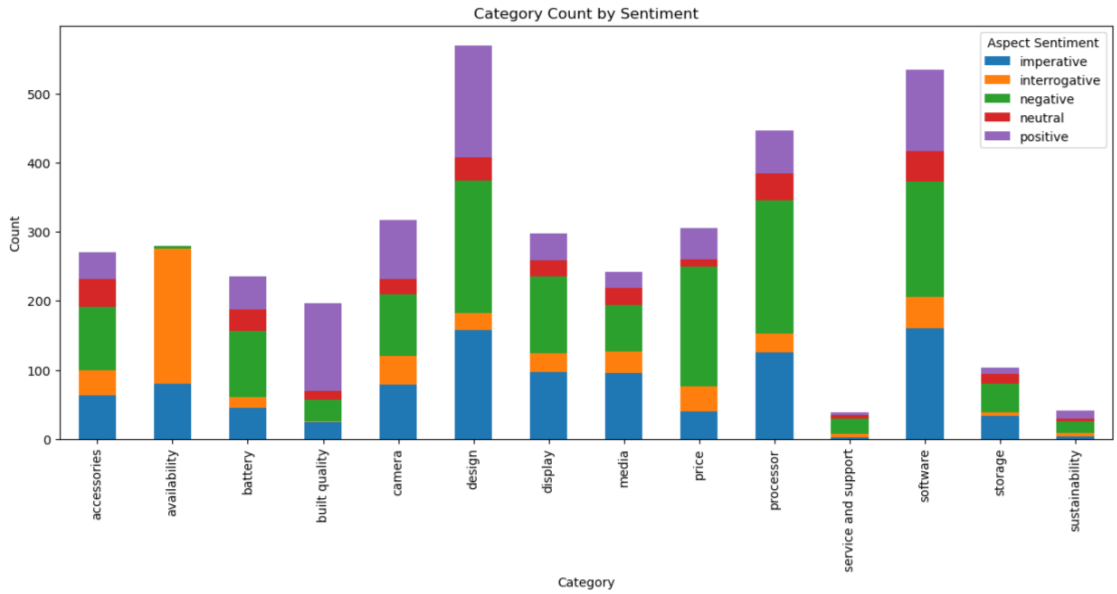

Figure 12: Aspect Sentiment distribution per aspect category class

The sentiment class for each aspect category needed to be determined to understand this. Hence, a table is presented with each sentiment class for each aspect category. For example, in the table, we can see for the "accessories" aspect of the phone, the maximum number of comments had negative sentiments, i.e., 34% of comments where aspects related to "accessories" were mentioned had negative sentiments. Similarly, for the "availability" aspect, the maximum number of comments had interrogative sentiment or attitude class (70%), meaning that people mostly asked questions concerning phone availability in the comments. Further, for the aspect "built quality", the maximum number of comments had positive sentiment (64%).

The aspect category "design" is the most discussed aspect of Nokia Mobile phones. People's sentiment or attitude towards this aspect shows that most people are happy with the design aspect of the phone. However, it can also be argued that people are also showing an imperative attitude towards this aspect: meaning they are requesting a change in the design aspect of the phone.

The aspect with the highest imperative sentiment shown is towards "media". 39% of comments where media aspects were discussed show imperative sentiment. Similarly, the aspect with the highest interrogative sentiment is "availability". 70% of comments with availability aspect show that people are asking questions about the availability of the phone. The “price” aspect received maximum negative sentiment from people, while the “built quality” aspect was discussed mainly positively.

The sentiment towards each identified aspect category can be inferred from the table. A detailed discussion on each aspect category and the sentiment is provided in Chapter 5.

The table below shows the absolute number of sentiments for each aspect category.

Aspect Category	imperative	interrogative	negative	neutral	positive
accessories	63	37	92	40	38
availability	80	196	3	0	0
battery	45	16	96	31	48
built quality	24	2	30	14	126
camera	79	41	90	22	85
design	158	24	192	34	162
display	97	27	112	23	39
media	95	31	68	24	24
price	40	36	174	10	45
processor	125	27	194	39	61
service and support	2	5	22	5	5
software	33	6	41	14	9
sustainability	4	5	17	3	12

Table 17: Number of sentiments labeled under each aspect category

The table below shows the percentage distribution of five sentiments for each aspect category.

Aspect Category	imperative	interrogative	negative	neutral	positive
accessories	23 %	14 %	34 %	15 %	14 %
availability	29 %	70 %	1 %	0 %	0 %
battery	19 %	7 %	41 %	13 %	20 %
built quality	12 %	1 %	15 %	7 %	64 %
camera	25 %	13 %	28 %	7 %	27 %
design	28 %	4 %	34 %	6 %	28 %

Aspect Category	imperative	interrogative	negative	neutral	positive
display	33 %	9 %	38 %	8 %	13 %
media	39 %	13 %	28 %	10 %	10 %
price	13 %	12 %	57 %	3 %	15 %
processor	28 %	6 %	43 %	9 %	14 %
service and support	5 %	13 %	56 %	13 %	13 %
software	30 %	9 %	31 %	8 %	22 %
storage	32 %	6 %	40 %	14 %	9 %
sustainability	10 %	12 %	41 %	7 %	29 %

Table 18: Percentage distribution of sentiment for each aspect category

4.2. Results of Machine Learning Models

The section of the study now presents the results of the machine learning models. The machine learning models are run with the best hyperparameters, as shown in the table above. The models are evaluated with the four-evaluation metrics: accuracy, precision, recall, and F1-score.

4.2.1. Classification Tasks

After the data preprocessing step, the comments are clean and ready for use in machine learning models. This study has used five machine learning models (logistic regressions, support vector machine, random forest, decision tree, and k- nearest neighbor) for the classification tasks. In addition, two feature extraction techniques have been used to extract the features from the dataset. The input and the output variables for this classification task are:

Input Variables: “cleaned comments” and “aspect terms”

Output Variables: “aspect category” and “aspect sentiment”

The approach used in classification tasks with the machine learning model in this thesis is given below.

1. Defining the input and output variables. The input variables, aspect terms, and aspect category are vectorized using BoW and TF-IDF feature extraction techniques.
2. The input variables, "*aspect terms*" and "*aspect category*" features, are combined.
3. Splitting the dataset into training (80%) and test set (20%).
4. Fitting the aspect category into classification models (LR, SVM, RF, DT, KNN).
5. Predicting the aspect category on the test dataset using the same classification model.
6. Fitting the aspect sentiment into classification models (LR, SVM, RF, DT, KNN).
7. Predicting the aspect sentiment on the test dataset using the same classification model.
8. Evaluating the aspect category classification model using evaluation techniques such as accuracy, precision, recall, and F1-score.
9. Evaluating the aspect sentiment classification model using the same evaluation techniques.

(Performing grid search for the output variable: "aspect category")

1. Creating a machine learning model (LR, SVM, RF, DT, KNN).
2. Defining the hyperparameters for the search.
3. Creating grid search object.
4. Fitting the grid search object to the data (for *aspect_category*).
5. Printing the best hyperparameters for aspect category.
6. Predicting the aspect category on the test data using the best hyperparameters.
7. Evaluating the performance of the model with the best hyperparameters.
8. Printing the evaluating metrics (accuracy, precision, recall, and F1-score) of the model with the best parameters.

(Performing grid search for the output variable: "aspect sentiment")

1. Creating a machine learning model (LR, SVM, RF, DT, KNN).
2. Defining the hyperparameters for the search.

3. Creating grid search object.
4. Fitting the grid search object to the data (for *aspect_sentiment*).
5. Printing the best hyperparameters for aspect sentiment.
6. Predicting the aspect sentiment on the test data using the given best hyperparameters.
7. Evaluating the performance of the model with the best parameters.
8. Printing the evaluating metrics.

The above explained steps are performed for each of the five machine-learning models used in this thesis. Furthermore, the above steps are performed twice for each model, i.e., one for BoW and the next for TF-IDF feature extraction techniques.

- Hyperparameters

The performance of the machine learning models can be enhanced by optimizing the hyperparameters. With the help of grid search, this thesis has experimented all the models with the minimum subsets of hyperparameters to know how different parameters affect the performance of the selected models. The parameters were tuned first, and the best parameter was used for the model in the classification tasks.

Model: Feature Extraction	**Hyper-Parameters Tunning**	**Best hyperparameters for Aspect Category**	**Best hyperparameters for Aspect Sentiment**
LR: BoW	C: [1.0,2.0,3.0,4.0,5.0]; max_iter: [50,100,1000]	{‘C’: 5.0, ‘max_iter’: 100}	{‘C’: 4.0, ‘max_iter’: 1000}
LR: TF-IDF	C: [1.0,2.0,3.0,4.0,5.0]; max_iter: [50,100,1000]	{‘C’: 5.0, ‘max_iter’: 50}	{‘C’: 5.0, ‘max_iter’: 1000}
SVM: BoW	C: [1.0,2.0,3.0,4.0,5.0]; kerne: [poly, linear, sigmoid]	{‘C’: 3.0, ‘kernel’: ‘linear’}	{‘C’: 1.0, ‘kernel’: ‘linear’}
SVM: TF-IDF	C: [1.0,2.0,3.0,4.0,5.0]; kerne: [poly, linear, sigmoid]	{‘C’: 2.0, ‘kernel’: ‘sigmoid’}	{‘C’: 1.0, ‘kernel’: ‘linear’}

Model: Feature Extraction	Hyper-Parameters Tunning	Best hyperparameters for Aspect Category	Best hyperparameters for Aspect Sentiment
RF: BoW	n_estimators: [10,100,1000]; max_depth: [50,100, 200, 300, 400, 500]	{‘max_depth’: 500, ‘n_estimators’: 1000}	{max_depth’: 200, ‘n_estimators’: 1000}
RF: TF-IDF	n_estimators: [10,100,1000]; max_depth: [50,100, 200, 300, 400, 500]	{max_depth’: 100, ‘n_estimators’: 1000}	{‘max_depth’: 100, ‘n_estimators’: 100}
DT: BoW	max_depth: [50,100, 200, 300, 400, 500, 1000]	{‘max_depth’: 300}	{‘max_depth’: 400}
DT: TF-IDF	max_depth: [50,100, 200, 300, 400, 500, 1000]	{‘max_depth’: 500}	{‘max_depth’: 300}
KNN: BoW	n_neighbors: [2,3,4,5,6,7,8,9,10]; weights: [‘uniform’, ‘distance’]	{‘n_neighbors’: 3, ‘weights’: ‘uniform’}	{‘n_neighbors’: 3, ‘weights’: ‘distance’}
KNN: TF-IDF	n_neighbors: [2,3,4,5,6,7,8,9,10]; weights: [‘uniform’, ‘distance’]	{‘n_neighbors’: 5, ‘weights’: ‘distance’}	{‘n_neighbors’: 8, ‘weights’: ‘distance’}

Table 19: Hyperparameters used in the Machine learning models

4.2.2. Results from BoW feature extraction technique for Aspect Category classification

The table below contains the results of the machine learning models using BoW features for aspect category classification.

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	80.15 %	80.53 %	80.15 %	80.17 %
Support Vector Machine	79.89 %	80.44 %	79.89 %	79.83 %
Random Forest	69.32 %	69.68 %	69.32 %	69.12 %

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree	74.22 %	75.16 %	74.22 %	74.36 %
K-Nearest Neighbor	50.77 %	54.04 %	50.77 %	49.90 %

Table 20: Result from ML models with BoW feature extraction method for Aspect Category classification

The model's performance can be considered good as the Logistic Regression (LR) achieved an accuracy of 80.15% and F1 score of 80.17%. Logistic Regression (LR) is followed by Support Vector Machine (SVM), which received an accuracy of 79.89% and a 79.83% F1 score. The good performance of Logistic Regression and Support Vector Machine can be attributed to the large feature set generated by Bag of Words (BoW) feature extraction techniques. On the other hand, the lowest performance score among the five models was from K-Nearest Neighbor, which showed 50.77% accuracy and 49.90% F1 score.

The classification report for the Logistic Regression model is given below.

Aspect Category Classification Report:

	precision	recall	f1-score	support
accessories	0.75	0.75	0.75	56
availability	0.88	0.98	0.93	59
battery	0.87	0.87	0.87	38
built quality	0.63	0.73	0.67	44
camera	0.80	0.73	0.76	59
design	0.78	0.84	0.81	117
display	0.79	0.83	0.81	53
media	0.76	0.69	0.72	49
price	1.00	0.85	0.92	66
processor	0.77	0.81	0.79	79
service and support	0.60	0.67	0.63	9
software	0.88	0.81	0.84	112
storage	0.73	0.67	0.70	24
sustainability	0.56	0.45	0.50	11
accuracy			0.80	776
macro avg	0.77	0.76	0.76	776
weighted avg	0.81	0.80	0.80	776

Figure 13: Aspect Category Classification Report of Logistic Regression (BoW)

From the above classification report, it can be seen that the overall accuracy of predicting aspect category is 80%. The model shows good performance for subclasses like availability, price, battery, software, design, and display, with F1 scores of 93%, 92%, 87%, 84%, and 81% for both design and display, respectively. The low performance is for the sub-class sustainability, which has an F1 score of 50%.

4.2.3. Results from BoW feature extraction techniques for Aspect Sentiment

The table below reports the performance of machine learning models using BoW features for aspect sentiment classification.

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	68.04 %	69.51 %	68.04 %	68.58 %
Support Vector Machine	65.59 %	66.11 %	65.59 %	65.74 %
Random Forest	70.48 %	70.46 %	70.48 %	70.16 %
Decision Tree	59.27 %	61.15 %	59.27 %	59.87 %
K-Nearest Neighbor	63.53 %	64.06 %	63.53 %	63.46 %

Table 21: Result from ML models with BoW feature extraction method for aspect sentiment classification

Random Forest classifier better predicted the aspect sentiment in this dataset. The model achieved 70.48% accuracy and a 70.16% F1 score. Random Forest is followed by Logistic Regression and Support Vector Machine with accuracy of 68.04% and 63.59%, respectively. Finally, the lowest-performing model is the Decision Tree, with an accuracy of 59.27% and F1 score of 59.87%.

The good performance of the Random Forrest classifier can be attributed to its capability to handle complex data with noise and non-linear relationships. Hence, it can be stated that in this classification task of predicting aspect sentiment, Random Forrest performed better by capturing more complex relationships between the variables for the training data set.

```
Aspect Sentiment Classification Report:
               precision    recall  f1-score   support

   imperative       0.78      0.78      0.78       199
interrogative       0.83      0.70      0.76       106
     negative       0.64      0.80      0.71       254
      neutral       0.52      0.41      0.46        59
     positive       0.73      0.56      0.64       158

     accuracy                           0.70       776
    macro avg       0.70      0.65      0.67       776
 weighted avg       0.71      0.70      0.70       776
```

Figure 14: Aspect Sentiment Classification Report of Random Forrest (BoW)

The overall accuracy of aspect sentiment classification is 70%. The model performance for imperative, interrogative, and negative could be considered relatively good compared to the other two classes. The neutral class has the lowest performance, with a 46% F1 score. The low performance on the neutral class could be because of the low support value or neutral class data being low on the dataset.

4.2.4. Results from TF-IDF feature extraction techniques for Aspect Category

The table below describes the performance of five machine learning models using the TF-IDF feature extraction technique for aspect category classification.

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	89.69 %	90.16 %	89.69 %	89.79 %
Support Vector Machine	86.85 %	88.05 %	86.85 %	87.09 %
Random Forest	73.32 %	73.95 %	73.32 %	73.41 %
Decision Tree	76.41 %	76.99 %	76.42 %	76.56 %
K-Nearest Neighbor	84.66 %	85.17 %	84.66 %	84.69 %

Table 22: Result from ML models with TF-IDF feature extraction method for aspect category classification

The table shows that the performance from the TF-IDF feature extraction technique yielded better performance from all five models compared to the BoW feature extraction technique. The better performance of TF-IDF can be attributed to its capability to capture the importance of words in the document. Also, this feature extraction technique outperforms BoW in taking account of the context of the words in the document.

Among the five models, Logistic Regression showed the best performance, with an accuracy of 89.69% and F1 score of 89.79%. Logistic Regression is followed by Support Vector Machine and K-Nearest Neighbor with 86.85% and 84.66% accuracy and F1 score of 87.09% and 84.69%, respectively. Though Random Forrest increased its performance compared to the BoW feature, where it produced the best performance, the performance from the model with TF-IDF is the lowest compared to other models.

```
Aspect Category Classification Report:
                     precision    recall  f1-score   support

         accessories       0.74      0.93      0.83        56
        availability       1.00      0.95      0.97        59
             battery       0.95      0.97      0.96        38
       built quality       0.84      0.84      0.84        44
              camera       0.84      0.83      0.84        59
              design       0.83      0.90      0.86       117
             display       0.92      0.87      0.89        53
               media       0.91      0.84      0.87        49
               price       1.00      0.95      0.98        66
           processor       0.92      0.92      0.92        79
 service and support       0.89      0.89      0.89         9
            software       0.97      0.89      0.93       112
             storage       0.91      0.88      0.89        24
      sustainability       0.82      0.82      0.82        11

            accuracy                           0.90       776
           macro avg       0.90      0.89      0.89       776
        weighted avg       0.90      0.90      0.90       776
```

Figure 15: Aspect Category Classification Report of Logistic Regression (TF-IDF)

The aspect category classification from the Logistic Regression model shows an accuracy of 90%. The model's performance has increased for all the sub-classes compared to the performance produced by the BoW approach. The model performed best for price, availability, battery, software, and processor, with F1 scores of 98%, 97%, 96%, 93%, and 92%, respectively. The lowest performance among 14 sub-classes was reported for sustainability with F1 score of 82%.

4.2.5. Results from TF-IDF feature extraction techniques for Aspect Sentiment

The table below shows the result of the models using the TF-IDF feature extraction technique for aspect sentiment prediction.

Models	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	66.23 %	67.43 %	66.23 %	66.72 %
Support Vector Machine	65.85 %	67.44 %	65.85 %	66.51 %
Random Forest	69.20 %	70.14 %	69.20 %	68.89 %
Decision Tree	60.95 %	61.68 %	60.95 %	61.07 %
K-Nearest Neighbor	54.12 %	52.87 %	54.12 %	52.81 %

Table 23: Result from ML models with TF-IDF feature extraction method for aspect sentiment classification

Similar to the performance from the model with the BoW feature, the Random Forrest Model showed the best performance among the five models selected for the aspect sentiment classification. The model yielded an accuracy of 69.20% and F1 score of 68.89%. Logistic Regression follows the performance with 66.23% accuracy and 66.72% F1 score. Finally, k-Nearest Neighbor showed the least performance with an accuracy of 54.12% and F1 score of 52.81%.

```
Aspect Sentiment Classification Report:
               precision    recall  f1-score   support

   imperative       0.78      0.78      0.78       199
interrogative       0.83      0.70      0.76       106
     negative       0.64      0.80      0.71       254
      neutral       0.52      0.41      0.46        59
     positive       0.73      0.56      0.64       158

     accuracy                           0.70       776
    macro avg       0.70      0.65      0.67       776
 weighted avg       0.71      0.70      0.70       776
```

Figure 16: Aspect Sentiment Classification Report of Random Forrest (TF-IDF)

Similar to the BoW classification report, the TF-IDF classification report also showed the accuracy of 70% while predicting aspect sentiment. The interrogative and positive sub-classes have performed relatively better than the other two. The neutral sub-class has the lowest performance, with F1 score of 46%.

4.3.Results from five-fold cross-validation

A cross-validation test was performed on the data to see the machine learning model's ability to predict the result of the new unseen data. Five-fold cross-validation was performed to split the data into five different training and test sets. The models were run with the best hyperparameter identified for that particular model to test the model's ability. The result after the cross-validation test is given in the table below.

- Cross Validation (CV) results for models using BoW technique for aspect category.

Model Name	result before CV (accuracy)	result after CV (accuracy)
LR	80%	78%
SVM	80%	77%
RF	70%	69%

Model Name	result before CV (accuracy)	result after CV (accuracy)
DT	74%	75%
KNN	51%	48%

Table 24: CV result for ML models with BoW method for aspect category classification

The logistic Regressions model performed the best with the cross-validation test as well. The model achieved the accuracy of 78%. The model's overall performance is similar to the result produced with a single train test dataset. The classification report is also similar to the initial model performance report. The high-performing sub-classes are availability, price, battery, camera, and processor, with F1 scores of 92%, 91%, 87%, 84%, and 82%, respectively. The low-performing sub-class is sustainability, with F1 score of 28%.

```
Mean Aspect Category Classification Report:
                     precision    recall  f1-score   support

        accessories       0.61      0.72      0.66       270
       availability       0.89      0.96      0.92       279
            battery       0.90      0.85      0.87       236
      built quality       0.72      0.71      0.71       196
             camera       0.86      0.81      0.84       317
             design       0.77      0.75      0.76       570
            display       0.83      0.73      0.78       298
              media       0.65      0.64      0.65       242
              price       0.95      0.88      0.91       305
          processor       0.81      0.83      0.82       446
service and support       0.48      0.56      0.52        39
           software       0.81      0.79      0.80       535
            storage       0.68      0.77      0.72       103
     sustainability       0.23      0.34      0.28        41

           accuracy                           0.78      3877
          macro avg       0.73      0.74      0.73      3877
       weighted avg       0.79      0.78      0.79      3877
```

Figure 17: Mean Aspect Category Classification Report of Logistic Regression model after Cross Validation (BoW)

- Cross Validation (CV) results for models using the BoW technique for aspect sentiment.

Model Name	result before CV (accuracy)	result after CV (accuracy)
LR	68%	55%
SVM	66%	54%
RF	70%	54%
DT	60%	39%
KNN	64%	40%

Table 25: CV result for ML models with BoW method for aspect sentiment classification

Initially, Random Forest performed better with 70% accuracy; however, after the cross-validation test, the logistic regression model achieved the best result with 55% accuracy. The low-performing model is Decision Tree with 39% accuracy.

The classification report shows the least performance from the model for the neutral sub-class with F1 score of just 12%. The interrogative and negative sub-class seem to perform better than others, but the performance is low for all the sub-classes.

```
Mean Aspect Sentiment Classification Report:
               precision    recall  f1-score   support

   imperative       0.58      0.57      0.57      1005
interrogative       0.61      0.72      0.66       499
     negative       0.62      0.60      0.61      1298
      neutral       0.11      0.14      0.12       303
     positive       0.57      0.50      0.53       772

     accuracy                           0.55      3877
    macro avg       0.50      0.51      0.50      3877
 weighted avg       0.56      0.55      0.55      3877
```

Figure 18: Mean Aspect Sentiment Classification Report of Logistic Regression Model after Cross Validation (BoW)

- Cross Validation (CV) results for models using TF-IDF technique for aspect category.

Model Name	result before CV (accuracy)	result after CV (accuracy)
LR	90%	88%
SVM	87%	86%
RF	73%	71%
DT	76%	73%
KNN	85%	86%

Table 26: CV result for ML models with TF-IDF method for aspect category classification

Similar to the result from the BoW feature extraction technique, the performance from models with the TF-IDF technique produced the best outcome for Logistic regression with accuracy of 88%. The difference in results from the initial single train and cross-validation tests is not significantly different, meaning the training data initially used was a good representative of the dataset.

After cross-validation, the classification report from the Logistic Regression model shows better results from the TF-IDF approach. The best result was for the sub-classes availability, price, battery, processor, and camera, with the F1 score of 98%, 97%, 95%, 93%, and 91%, respectively. Though the model's performance for the sustainability sub-class was increased compared to the result from the BoW approach, the result is the lowest among the 14 sub-classes.

Mean Aspect Category Classification Report:

	precision	recall	f1-score	support
accessories	0.68	0.87	0.77	270
availability	1.00	0.95	0.98	279
battery	0.99	0.92	0.95	236
built quality	0.90	0.86	0.88	196
camera	0.95	0.88	0.91	317
design	0.79	0.87	0.83	570
display	0.90	0.85	0.88	298
media	0.88	0.79	0.83	242
price	1.00	0.95	0.97	305
processor	0.94	0.92	0.93	446
service and support	0.78	0.97	0.86	39
software	0.93	0.87	0.90	535
storage	0.81	0.86	0.84	103
sustainability	0.47	0.59	0.52	41
accuracy			0.88	3877
macro avg	0.86	0.87	0.86	3877
weighted avg	0.89	0.88	0.89	3877

Figure 19: Mean Aspect Category Classification Report of Logistic Regression model after Cross Validation (TF-IDF)

- Cross Validation (CV) results for models using TF-IDF technique for aspect sentiment.

Model Name	result before CV (accuracy)	result after CV (accuracy)
LR	66%	57%
SVM	66%	57%
RF	69%	53%
DT	61%	39%
KNN	54%	48%

Table 27: CV result for ML models with TF-IDF method for aspect sentiment classification

The initial test with a single test and train set showed Random Forrest performed better among the five models with 69% accuracy; however, after a 5-fold cross-validation test, the performance from the model decreased to 53%. The best-performing models after cross-validation tests are the Logistic Regression and Support Vector Machine, with accuracy score of 57%. The Decision Tree produced the lowest performance with 39% accuracy.

From the classification report, it can be observed that the performance for the sub-classes has slightly improved compared to the BoW technique. The model has performed better for interrogative, imperative, and negative sub-classes than the positive and neutral sub-classes. They achieved the F1 score of 67%, 60%, and 63%, respectively. The least performance from the model was for the neutral sub-class, with a 15% F1 score.

```
Aspect Sentiment Classification Metrics:
Accuracy: 0.5721014299966745
Precision: 0.575740093870591
Recall: 0.5721014299966745
F1-score: 0.5709351290703519
Mean Aspect Sentiment Classification Report:
               precision    recall  f1-score   support

   imperative       0.60      0.60      0.60      1005
interrogative       0.67      0.67      0.67       499
     negative       0.62      0.64      0.63      1298
      neutral       0.14      0.15      0.15       303
     positive       0.56      0.52      0.54       772

     accuracy                           0.57      3877
    macro avg       0.52      0.52      0.52      3877
 weighted avg       0.57      0.57      0.57      3877
```

Figure 20: Mean Aspect Sentiment Classification Report of Logistic Regression Model after Cross Validation (TF-IDF)

From the 5-fold cross-validation test, it can be inferred that the result of models for aspect category classification did not differ significantly, highlighting no issue of overfitting the training data for both BoW and TF-IDF feature extraction techniques. However, for aspect sentiment classification, the performance of models shows a significant difference between the initial result of the model with single training and test set and the 5-fold cross-validation test. The significant difference can be explained by the issue of overfitting the training dataset for the initial model run before performing the cross-validation test.

However, upon evaluating the variance of the results from the 5-fold cross-validation, the result showed low variance among different subsets of data. The variance in the evaluation metrics was determined to know how the metrics vary among different datasets. The Logistic Regression model produced the best result for aspect sentiment using BoW; thus, variations in the evaluation metrics were checked for this model. After cross-validation, the accuracy and F1 score for aspect sentiment classification using the

BoW technique were 0.0011 and 0.0012. Also, Logistic Regression produced the best result for aspect sentiment classification using the TF-IDF approach; hence, the variance in the evaluation metrics was computed in the model as well. The variance in accuracy and F1 score across different datasets from the cross-validation test produced the result of 0.0007 and 0.00058. The variance in the evaluation metrics from the two best-performing models is very low. This implies that the model's performance is consistent with the different datasets and evaluations, and the issue of overfitting and underfitting hasn't impacted the data.

However, despite the consistent performance of the model, the performance of the model itself can be considered to be low. There could be multiple reasons for the low performance of the models for aspect sentiment classification. For example, the issue could be due to low data size, the quality of data about the noise, and the complexity of the YouTube comments. The discussion on the possible reasons will be discussed in the next chapter.

5. DISCUSSION

This chapter of the thesis discusses the results produced in this study. The discussion will aim at explaining the answer to the research questions and also the application of this study in the business field.

5.1. RQ 1: Understanding the current study domain of aspect-based sentiment analysis with YouTube comments.

The approaches used for aspect-based sentiment analysis largely focus on using four methods: frequency-based, syntax-based, supervised approach, and unsupervised approach. The frequency-based approach determines the frequency of words associated with the aspects to determine the attitude towards those aspects. The syntax-based approach examines the relationship between words to understand what aspect is discussed and the attitude expressed towards that aspect. The supervised Learning method uses a labeled dataset to train machine learning models to predict the aspect and the sentiment towards that aspect. The unsupervised learning method mainly uses clustering techniques to gather discussed aspects and analyze the sentiment towards those aspects.

YouTube, because of its popularity, has been rigorously studied in the past. For example, some studies have investigated the accuracy of specific important topics (Briones et al., 2012), the communication value of the videos (Lewis et al., 2012), and also the identification of the content of a small set of videos (Desai et al., 2013; Smith et al., 2012). Further, the platform has seen its presence in the study related to content analysis as well. However, multiple challenges have been found in using YouTube and other social media platforms to perform social media analytics.

Though other forms of social media analytics have been performed using YouTube data, from the literature survey, it was identified that aspect-based sentiment analysis on YouTube comments is an untouched area, and the reasons for this is explained by the complexity of the data that is generated in this social media platform to perform aspect-based sentiment analysis. The study by Severyn et al. (2016) highlighted almost no work has underlined the effective opinion mining with YouTube comments. The study also highlights that the closest study that worked with YouTube comments is from Siersdorfer et al. (2010), which proposed different classifications of YouTube comments.

Several issues and challenges to performing aspect-based sentiment analysis on YouTube comments have prevented extensive research in this area. One of the most important tasks in aspect-level sentiment analysis is extracting the aspect terms or keywords. Since there are no or very few domains specific corpora developed for YouTube comments, developing, and designing keyword dictionaries is difficult (Bordoloi & Biswas, 2023). Also, using generalized dictionaries does not give good results as they are developed for specific domain use. Further, the lack of domain-specific dictionaries for YouTube comments challenges the researchers to use generalized sentiment dictionaries as predefined polarity for a word affects the performance of the model as the dictionaries are very domain-specific (Bordoloi & Biswas, 2023).

The noise and unstructured nature of comments (Gilbert & Karahalios, 2009) present further challenges to performing aspect-based sentiment analysis on YouTube comments. This very issue has been realized in this thesis as well. The study extracted 9737 comments, but the unique comments labeled were just 2396, i.e., just about 25% of the data was useable for the analysis. The low useability of the data can further create a big problem of class imbalance in the dataset, where some classes are highly represented while some classes majorly underrepresented. Additionally, the comments are full of emojis, spam, and sarcasm. This presents another hurdle for performing aspect-based sentiment analysis on this platform. Moreover, because of highly unstructured text, the comments can contain multiple simultaneous sentiments (positive and negative) bearing expressions, making it difficult for the machine learning model to learn and predict them (Singh & Tiwari, 2021). Hence, there is a big gap in performing aspect-based sentiment analysis using YouTube data because of all these reasons.

However, it’s essential to utilize the data on social media platforms like YouTube as they hold a large user base and have a magnitude of opinionated content. Text analytical approaches like aspect-based sentiment analysis can be an appropriate approach for companies to dissect the data and gather meaningful information that could assist the company in decision-making.

5.2.RQ 2: Identification of the most discussed aspects and the associated sentiments.

The core idea of this thesis was to understand the prevailing attitude of the public (YouTube users) towards the recently released Nokia mobile phones. The motive was to identify the aspects mentioned in the comments and determine people's sentiments when referring to different phone aspects. This study successfully identified the 14 different aspects category that people mainly discussed. Also, the study identified that most comments show negative attitudes towards different mobile phone aspects. Further, the study found out that a significant number of comments have shown an imperative attitude meaning the expectation is there among the users for the company to provide better products in the future.

The information from this kind of study could be immense for companies (HMD Global in this case) as the company can use the extracted information or the feedback at the granular level to strengthen the advantage aspect (built quality in this case) and work on the improvement areas (in aspects like design, software, processor, price, display, camera, and sustainability) to grow in the market. Further, the information thus generated can be used in other business functions like advertising and marketing to appeal to the customers properly.

The sentiment distribution shows that most user comments are negative sentiments towards different aspects of Nokia Mobile phones. Eleven out of 14 aspects received more negative sentiment from users than other sentiment classes. The aspects, "price", "accessories", "battery", "camera", "design", "display", "processor", "service and support", "software", "storage", and "sustainability" received highly negative feedback from the users. The aspect "price" received the most negative sentiment from the people. Altogether, 305 instances were found in the dataset where the aspect price was mentioned, and 57% of the time, negative sentiment was shown toward the phone's price. When reviewing the comments that mentioned the price aspect, users largely mentioned the price is high compared to other mobile phones with similar specifications. Some of the comments that mentioned the aspect "price' in the comment are:

Comment 1: *"they are selling overpriced smartphones compared to other brand."*
Comment 2: *"overpriced compared to other brands in same specifications."*

Comment 3: *"its an overpriced phone when market is full of many more options in lower to this range"*

Similarly, out of 270 comments related to the aspect "accessories" category, in 34% of instances, users showed negative comments. The aspect "accessories" covers different phone functions like headphones, earbuds, Bluetooth, USB, and network. The users have shown negative sentiment towards this aspect for the reason that the phone does not have that aspect, or they are being compared to other phones, which provides a better solution. For instance, the comment, *"nokia earbuds third class",* highlights the disappointment of the user regarding the earbud from Nokia.

The negative feedback shown towards the aspect "battery" relates to the issues that Nokia Mobile phones have about the charging port. Also, the battery capacity and the power draining issues have been highlighted in the comments section. Similarly, for the camera, users have largely pointed out their concern for camera bumps, camera housing, and overall camera quality of the nokia phones. For the "design" aspect, users have shown dissatisfaction about the phone's design, the waterdrop notch, the bezel design, and the phone's size mostly. Further, for the "display" aspect, users have raised the issue of Nokia mobile phones not upgrading their technology compared to other phones. The use of HD displays has gathered a lot of negative attention. Also, the display size is one of the significant concerns mentioned.

Users have explicitly highlighted that Nokia mobile phones haven't been able to catch up with other mobile phone manufacturers regarding the "processor" of the phone. The use of outdated chipsets and their low performance has been discussed mainly under this aspect. For the "software" aspects, people have complained significantly about the slow release of software security updates from the company. Further, under the "storage" aspect, low ram capacity and no digital security card support have been highly criticized.

The only aspect that received the highest positive feedback is for the "built quality" of the phone. People have praised and shown a positive attitude towards the phone's build quality and durability aspect. Out of 196 mentions of aspects related to the build quality of the phone, in 64% of instances, users showed a positive reaction to the quality of the phone in terms of its durability. Some of the comments where people have highlighted the build quality of the phone are:

Comment 1: *"quality is an introduction to the name of nokia my favorite brand is nokia"*
Comment 2: *"nokia devices are strong and reliable, and the cameras are of excellent quality"*
Comment 3: *"bought this online, and i was shocked how well it is built for the money i paid"*

The second highest sentiment class after "negative" sentiment is the imperative sentiment class. From the sentiment distribution, it is clear that people have a lot of expectations (imperative class) with regard to different aspects of the phone, such as "software", "availability", "media", "display", "design", "storage" and "processor".

For the aspect "software", users have mainly discussed the need to bring "Symbian and Kai operating system" back. Also, they have requested to provide timely software updates and have shown expectations towards the company to have their own software. Similarly, for the aspect "availability", most comments have requested the company to make the phone available or release it in their country, in their location. Finally, under the aspect "media", most users have requested the company to have inbuilt music streaming services and dual loud stereo speakers.

Furthermore, for the aspect "display", people expect to improve different aspects such as bezels, screen size, refresh rate, and aspect ratio. Most comments have requested the company to have an Amoled display, 1080p display, and pure view display technology. Also, under the "design" aspect, people have requested the company to remove the notch from the Nokia mobile phones, bring back Lumia designs, circular camera designs, and remove the bezels from the phone. Some of the comments with the "design" aspect discussed, and imperative sentiment shown are:

Comment 1: *"bring back the lumia designs please"*
Comment 2: *"nokia please bring back the circular camera design"*
Comment 3: *"please remove the ugly notch bezel punch hole camera is better"*

For the "storage" aspect, people have requested or wished to have an expandable secure digital card, better ram, and increased internal memory of the phone. Further, under the "processor" aspect, users expect to use better chipsets in their devices, do not use outdated specs, and launch their own chip for their phones.

The “availability” aspect received many questions regarding the release and the availability of the phone. Out of 279 availability aspect mentioning comments, in 70% of instances, people have raised some queries and asked questions. Some of the comments that mentioned the aspect “availability” are:

Comment 1: *“nokia mobile, when will you launch the phone in Philippines weve been waiting”*
Comment 2: *“will this model come to india i already have the nokia g20 i am trying to get a new mobile but if the g50 is coming to india then i am willing to wait for some time”*
Comment 3: *“when will it be available in malaysia any official launch”*

From the distribution of the five-sentiment class, it is clear that users have shown more negative sentiment towards different phone aspects. Also, users have put forward a lot of expectations and queries for the company. The information from the three sentiment classes, negative, imperative, and interrogative sentiment class has provided very useful information for the HMD Global company highlighting the areas of improvement.

The aspects of product improvement have been clearly outlined in this study for Nokia Mobile phones. The result shows the company has tremendous work to do in addressing the current complaints from the users and also a great possibility for improvement. The problematic areas of Nokia mobile phones have been highlighted in this study. The satisfaction level of Nokia Mobile users concerning different aspects can be articulated from the result of negative sentimental comments that have been identified. There is a business risk for the company to lose existing customers as well as potential customers, if necessary actions are not taken.

5.3. RQ 3 and 4: Determining the best performing Feature Extraction Technique and Machine Learning Model.

The thesis further studied how different machine learning models would perform with the labeled data to predict the aspect category and sentiment discussed in the comments. The study used two feature extraction techniques, and each model’s performance was evaluated with metrics such as accuracy, precision, recall, and F1 score. The results from the single train-test dataset show that the TF-IDF feature extraction technique performs better than the BoW technique for both aspect category and aspect sentiment classification. The best-performing model for aspect category and aspect sentiment with

BoW is Logistic Regression, with accuracy of 80.15%, and Random Forrest, with the accuracy of 70.48%, respectively. Similarly, Logistic Regression performed best with 89.69% accuracy for aspect category classification and Random Forrest with 69.85% accuracy for aspect sentiment classification.

Further, the 5-fold cross-validation test ensured that the TF-IDF feature extraction technique performed better on this data set as for both aspect category and aspect sentiment prediction, the result of models with the TF-IDF technique produced the best result. With the BoW approach, Logistic Regression produced the best result of 78% accuracy, while the same model had 88% accuracy with TF-IDF. Similarly, with the TF-IDF approach, Logistic Regression and Support Vector Machine produced the best result for aspect sentiment classification. Both the models had 57% accuracy in prediction. The result with the BoW approach shows the highest performance from Logistic Regression with 55% accuracy.

Hence, from the results obtained, it can be inferred that the TF-IDF feature extraction technique could extract the features from the dataset better than the BoW approach. The better performance of TF-IDF can be attributed to the fact that it captures the word importance in the sentence by assigning weight to the words. BoW, on the other hand, only considers the frequency of words. The TF-IDF, thus can better capture the semantic meaning of the text. Further, the Logistic Regression Model among the five chosen models performed best for this thesis's manually labeled comment dataset. The relatively small size of the data and the model's ability to handle noisy data could be one of the reasons for the Logistic Model to perform better than the other four models. The best-performing models can be referenced for further study or used to perform aspect and sentiment classification for new data.

The result from the cross-validation test showed the issue with the data quality. Multiple reasons could attribute to the low performance of machine learning models for aspect sentiment classification. One of the notable reasons is the distribution of class sample frequencies. The low number of samples for neutral comments might have affected the model's performance. Further, the sentence structure's complexity could be a significant influencing factor for low performance. As stated earlier, the comments are noisy and unstructured and do not sometimes contain the necessary details for the analysis.

5.4.Contribution of research

This study attempted to understand aspect-based sentiment analysis for YouTube comments. The study used a manual annotation process to annotate aspect terms and sentiments mentioned in the text. The thesis adopted the sentence type classification used in the study by Pokharel & Bhatta (2021). However, in this thesis, the two classifications, “imperative” and “interrogative”, were added to the three existing sentiment classes to better understand and describe people’s prevailing attitudes through comments. The idea to incorporate the two-attitude class was to explain that YouTube comments do not just fall under positive, negative, or neutral sentimental aspects but also provide other information that can be equally or even more significant than the three-sentiment dimension. Thus, this thesis successfully categorized the sentiment aspect of YouTube comments into imperative, interrogative, positive, negative, and neutral.

The result of the sentiment categorization into five classes has provided valuable insights into what customers say about different aspects of the Nokia mobile phones. Though negative sentiments towards multiple aspects were mostly identified, the imperative and interrogative classes provided critical information. These classes explained the user’s attitude of expectations and the need to answer or clarify certain things on different mobile phone aspects.

Further, though the scale of the research was small (use of small data size), the study identified 14 different phone aspects from the YouTube comments. This study can be said to contribute to defining the important aspects that people mention in YouTube comments. The 14 aspects identified in the thesis can be used as a baseline for the future with YouTube comments (mobile phone domain) in understanding public perception towards different aspects of mobile phones.

Moreover, this study adds value to previously identified issues and challenges of performing aspect-based sentiment analysis with YouTube comments, as the results from machine learning models, particularly for aspect sentiment classification, were seen to be impacted by the data quality. The notion that the unstructured and noisy nature of YouTube comments has been identified in this thesis as well.

Furthermore, this thesis can be considered a foundational benchmark for aspect-based sentiment analysis on YouTube comments. Therefore, the approaches and methods adopted here can be studied and challenged further in future studies.

6. CONCLUSION

This thesis aimed to explore the aspect-based sentiment analysis approach on YouTube comments. The thesis aimed to identify the prevailing attitude of people towards Nokia Mobile phones by identifying different phone aspects and the sentiment associated with the aspects. Four research questions were constructed to identify the research state of aspect-based sentiment analysis about the use of YouTube comments, identify the most discussed aspects of Nokia mobile phones and the associated sentiment, and understand how machine learning models would perform on YouTube comments in classifying the aspects and the sentiments.

The thesis initially discussed sentiment analysis, and its types, with particular attention to aspect-based sentiment analysis. The related work chapter of the thesis provides an overview of the current practices and approaches to performing aspect-based sentiment analysis. Further, the chapter defines some of the previous studies using YouTube data to perform sentiment analysis. Moreover, the section builds on the importance of identifying the determinant mobile phone aspects that influence people in their buying behavior of mobile phones.

9737 comments were extracted from YouTube to perform the aspect-based sentiment analysis. The objective was to perform supervised learning; thus, the data were manually labeled before running the machine learning models. Fourteen different aspects were identified from the annotation process. These aspects are similar to those identified in the literature highlighted as the key factors people consider when making a mobile phone purchase decision. The comments were preprocessed, and several preprocessing techniques were used to clean the data before feeding the data into the models. Two feature extraction techniques: Bow and TF-IDF, were used for each model, and four different evaluation metrics were used to evaluate the performance of machine learning models.

The first research question for this thesis aimed at understanding the current research situation for aspect-based sentiment analysis with YouTube comments. The literature review showed that YouTube comments have largely been left out by academicians and researchers for aspect-based sentiment analysis. The construct, noise, unstructured, and complexity of texts on YouTube were identified as one of the major issues in performing aspect-based sentiment analysis.

The second research question aimed to identify the most discussed aspects and the sentiment in the comments extracted from Nokia Mobile phones' YouTube videos. In the labeled data, 14 different aspects were identified, and five different sentiment classes were defined for the comments. The majority of identified aspects had negative sentiments. However, only the "built quality" aspect of the phone received positive user feedback. Further, the data showed that people have many expectations and questions about the aspect being discussed. Thus, the company must consider this feedback and adjust its product accordingly.

The third and fourth research questions were related to the classification tasks. The performance of two feature extraction techniques and five different machine learning models were evaluated with the labeled data. The results show that the TF-IDF feature extraction technique has performed better for the dataset. Also, Logistic Regression Model outperformed all four other machine learning models in its performance. The best performance for aspect category prediction was from the Logistic Regression model with the TF-IDF feature extraction method. The model's accuracy before and after the cross-validation test was 90% and 88%. Similarly, Logistic Regression and Support Vector Machine produced similar results for aspect sentiment classification with the TF-IDF method. The accuracy of the Logistic Regression and Support Vector Machine model before and after the cross-validation test was 66% and 57%, respectively.

The performance of models for aspect category classification can be considered good; however, the model performance for aspect sentiment classification could be improved. The models can predict the aspect category and aspect for a new study in the mobile phone domain for YouTube comments.

7. LIMITATIONS OF THE STUDY

This section of the thesis explains the study's limitations and the considerations for future studies for similar research.

Even though almost 10,000 comments were extracted from YouTube videos, only 25% of the comments were useful for the study, as other comments did not carry any useful information. To model the algorithms to produce good performance, they need huge data. Because the number of useful information-carrying comments is very less, there is always the possibility of the data size being undersized for the study. One of this thesis's limitations is that the data size is relatively small for models to produce good results. Further, the comments extracted were from a limited number of videos; hence, the data cannot be called a representative of the entire YouTube comments for Nokia Mobile phones. Thus, the results from the study can be considered to be a limited generalization of the population.

The emoticons used in the comments can provide an important message. However, in this study, the emoticons were not considered while labeling the data. Also, comments written in languages other than English were removed from the dataset. Thus, it can be said that some of the sources that provided valuable information were lost in the process. Similarly, the comments extracted were manually labeled by a single annotator in this study. Therefore, the data could have faced subjective bias. Furthermore, when a single person labels the data, there is no possibility of performing the inter-annotator agreement, which measures consistency in labeling the data. Hence, a single annotator labeling the data can limit the reliability and validity of the data.

Furthermore, in this thesis, the model was provided with aspect terms features for aspect category prediction. However, for aspect sentiment classification, no additional features were provided. Thus, this can be one of the reasons for the low performance of models for aspect sentiment classification.

8. CONSIDERATIONS FOR FUTURE STUDIES

Future studies can consider the following things to address this study's limitations and incorporate or widen the scope and practices of aspect-based sentiment analysis on YouTube comments.

In the future, studies similar to this thesis can be performed with more data (comments with useful information) to enhance the performance of the models. Further, the sentiment-bearing words can be fed to the model to produce better results for aspect sentiment classification. Increasing the input features might help the model learn better and thus enhance classification results.

There is a possibility of future studies to include emoticons and also comments in different languages to depict the true attitude of users toward the aspects mentioned. In addition, a dictionary or a corpus of mobile phone aspects and sentiment-bearing words could be developed for YouTube comments to help more studies perform aspect-based sentiment analysis of YouTube comments.

This study used only two feature extraction techniques or word embedding approaches (BoW and TF-IDF). It is possible to incorporate more methods like word2vec, where the probability of a word from its neighboring words is figured out. The general principle of this approach is to predict the neighboring word to understand the semantic relationship between the words in the text. Furthermore, as YouTube comments are unstructured and noisy, it would be a good idea to test different vectorization methods to see the difference in the performance of the machine learning algorithms.

Different aspects-based sentiment analysis approaches could be used to perform a comparative study. For instance, future studies can use the machine and deep learning approaches to compare the aspect and sentiment classification results for YouTube comments. It is reported that deep learning approaches (such as CNN, RNN, and LSTM) can achieve better results than machine learning algorithms on different NLP problems (Minaee et al., 2021; Wu et al., 2020). Also, the experiment with deep learning approaches can produce interesting results as these approaches are said to be able to handle noisy and better identify the relationships between the input and the output features of the models (Palanivinayagam et al., 2023).

Since the manual labeling of data is a time-consuming task, other machine learning approaches for data labeling can be experimented with the YouTube data. For instance, the active learning approach (Hu et al., 2016) or topic modeling methods (Pavlinek & Podgorelec, 2017) can be used to label the data to avoid manual labeling. Using other approaches eliminates the issues arising from manual labeling, like biases and inconsistency in the labeling process.

REFERENCES

Abubakar, H. D., & Umar, M. (2022). Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, *4*(1&2), 27–33. https://doi.org/10.56471/slujst.v4i.266

Alfaro, C., Cano-Montero, J., Gómez, J., Moguerza, J. M., & Ortega, F. (2013). A multi-stage method for content classification and opinion mining on weblog comments. *Annals of Operations Research*, *236*(1), 197–213. https://doi.org/10.1007/s10479-013-1449-6

Aslam, N., Xia, K., Rustam, F., Hameed, A., & Ashraf, I. (2022). Using Aspect-Level Sentiments for Calling App Recommendation with Hybrid Deep-Learning Models. *Applied Sciences*, *12*(17), 8522. https://doi.org/10.3390/app12178522

Bennett, K. P., & Campbell, C. (2000). Support vector machines. *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13. https://doi.org/10.1145/380995.380999

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10442-2

Briones, R., Nan, X., Madden, K., & Waks, L. (2012). When Vaccines Go Viral: An Analysis of HPV Vaccine Coverage on YouTube. *Health Communication*, 27(5), 478–485. https://doi.org/10.1080/10410236.2011.610258

Brody, S., & Elhadad, N. (2010, June). An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics,* 804-812.

Bruce, L. M., Koger, C. H., & Jiang Li. (2002). Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, *40*(10), 2331–2338. https://doi.org/10.1109/tgrs.2002.804721

Campbell-Meiklejohn, D., Bach, D., Roepstorff, A., Dolan, R., & Frith, C. (1970). How the opinion of others affects our valuation of objects. *Frontiers in Behavioral Neuroscience*, *20*(13). https://doi.org/10.3389/conf.neuro.08.2009.01.001

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chen, H. Y., & Chen, H. H. (2016, August). Implicit polarity and implicit aspect recognition in opinion mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 20-25.

Chen, Z., & Qian, T. (2020). Relation-aware collaborative learning for unified aspect-based sentiment analysis. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3685–3694. https://doi.org/10.18653/v1/2020.acl-main.340

Chiciudean, G. O., Harun, R., Muresan, I. C., Arion, F. H., Chiciudean, D. I., Ilies, G. L., & Dumitras, D. E. (2019). Assessing the Importance of Health in Choosing a Restaurant: An Empirical Study from Romania. *International Journal of Environmental Research and Public Health*, 16(12), 2224. https://doi.org/10.3390/ijerph16122224

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Cullen, F. (2005). Factors Influencing Restaurant Selection in Dublin. *Journal of Foodservice Business Research*, 7(2), 53–85. https://doi.org/10.1300/j369v07n02_05

Da'u, A., & Salim, N. (2019). Aspect extraction on user textual reviews using multi-channel convolutional neural network. *PeerJ Computer Science*, *5*, e191. https://doi.org/10.7717/peerj-cs.191

Desai, T., Shariff, A., Dhingra, V., Minhas, D., Eure, M., & Kats, M. (2013). Is Content Really King? An Objective Analysis of the Public's Response to Medical Videos on YouTube. *PLoS ONE*, *8*(12), e82469. https://doi.org/10.1371/journal.pone.0082469

Ding, X., Liu, B., and Yu, P. S. (2008). ‘A holistic lexicon-based approach to opinion mining’, In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 231-240. ACM.

En, L. E., & Balakrishnan, B. (2022). Smartphone Purchase Intention by Nilai University Students Living in the City of Nilai Malaysia. *SRIWIJAYA INTERNATIONAL JOURNAL of DYNAMIC ECONOMICS and BUSINESS*, *6*(2), 131. https://doi.org/10.29259/sijdeb.v6i2.131-146

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of Language Resources and Evaluation Conference*, pages 417–422.

Filippova, K., & Hall, K. B. (2011). Improved video categorization from text metadata and user comments. In *Proceedings of the 34th international ACM SIGIR* (pp. 835–842).

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 211–220.

Hakim, A. F., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *International Conference on Information Technology and Electrical Engineering*. https://doi.org/10.1109/iciteed.2014.7007894

Harrington, R. J., Ottenbacher, M. C., & Kendall, K. W. (2011). Fine-Dining Restaurant Selection: Direct and Moderating Effects of Customer Attributes. *Journal of Foodservice Business Research*, 14(3), 272–289. https://doi.org/10.1080/15378020.2011.594388

Harrison, O. (2018, September 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Medium; Towards Data Science. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Hu, M., and Liu, B.(2004). 'Mining and summarising customer reviews', In R. Kohavi, J. Gehrke, W. DuMouchel, & J. Ghosh (Eds.), KDD-2004 - In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177.

Hu, R., Mac Namee, B., & Delany, S. J. (2016). Active learning for text classification with reusability. *Expert Systems with Applications*, *45*, 438–449. https://doi.org/10.1016/j.eswa.2015.10.003

Jin, W., Ho, H. H., and Srihari, R. K. (2009). A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 465–472.

Karjaluoto, H., Karvonen, J., Kesti, M., Koivumäki, T., Manninen, M., Pakola, J., Ristola, A., & Salo, J. (2005). Factors Affecting Consumer Choice of Mobile Phones: Two Studies from Finland. *Journal of Euromarketing*, 14(3), 59–82. https://doi.org/10.1300/j037v14n03_04

Kemp, S. (2023, January 26). *Digital 2023: Global Overview Report*. DataReportal. https://datareportal.com/reports/digital-2023-global-overview-report

Kim, T.-T., Nguyen, Huynh, S., Pham, P., Phan, L., Nguyen, D.-V., & Van Nguyen, K. (2021). *Span Detection for Aspect-Based Sentiment Analysis in Vietnamese*.

Kumar, A., & Jaiswal, A. (2019). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, *31*(1), e5107. https://doi.org/10.1002/cpe.5107

Landy, S. H., & Szalay, A. S. (1993). Bias and variance of angular correlation functions. *The Astrophysical Journal*, 412, 64. https://doi.org/10.1086/172900

Lee, B. D. (2013, September 3). Nokia: The rise and fall of a mobile giant. *BBC News*. https://www.bbc.com/news/technology-23947212

Lewis, S. P., Heath, N. L., Sornberger, M. J., & Arbuthnott, A. E. (2012). Helpful or Harmful? An Examination of Viewers' Responses to Nonsuicidal Self-Injury Videos on YouTube. *Journal of Adolescent Health*, 51(4), 380–385. https://doi.org/10.1016/j.jadohealth.2012.01.013

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., and Yu, H. (2010). Structureaware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics*, pages 653–661. Association for Computational Linguistics.

Li, G., Law, R., Vu, H. Q., Rong, J., and Zhao, X. R. (2015). 'Identifying emerging hotel preferences using Emerging Pattern Mining technique', *Tourism Management*, 46, pp.311–321.

Li, X., Bing, L., Li, P., & Lam, W. (2019). A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 6714–6721. https://doi.org/10.1609/aaai.v33i01.33016714

Lin, C., He, Y., Everson, R., & Ruger, S. (2012). Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 1134–1145. https://doi.org/10.1109/tkde.2011.48

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/s00416ed1v01y201204hlt016

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops* (pp. 81-88). IEEE.

Mack, Z., & Sharples, S. (2009). The importance of usability in product choice: A mobile phone case study. *Ergonomics*, 52(12), 1514–1528. https://doi.org/10.1080/00140130903197446

Madden, A., Ruthven, I., & McMenemy, D. (2013). A classification scheme for content analyses of YouTube video comments. *Journal of Documentation*, *69*(5), 693–714. https://doi.org/10.1108/jd-06-2012-0078

Mai, L., & Le, B. (2020). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03534-7

Marrese-Taylor, E., Balazs, J. A., & Matsuo, Y. (2017). Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis,* 102–111.

Mate, C. (2015). Product aspect ranking using sentiment analysis: a survey. *International Research Journal of Engineering and Technology*, *3*(01), 126-127.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Mishra, N., & Singh, A. (2018). Use of twitter data for waste minimisation in beef supply chain. *Annals of Operations Research*, 270(1–2), 337–359.

Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019, October). Sentiment analysis of positive and negative of youtube comments using naïve bayes–support vector machine (nbsvm) classifier. In *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*,199-205. IEEE.

Mukherjee, A., & Liu, B. (2012). *Aspect Extraction through Semi-Supervised Modeling*. Meeting of the Association for Computational Linguistics, 1, 339–348. https://www.cs.uic.edu/~liub/publications/ACL-2012-aspect-extraction.pdf

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 1-5.* https://doi.org/10.1109/icccnt.2013.6726818

Nguyen, B. T., Morell, C., & De Baets, B. (2016). Large-scale distance metric learning for k-nearest neighbors regression. *Neurocomputing*, 214, 805–814. https://doi.org/10.1016/j.neucom.2016.07.005

Palanivinayagam, A., El-Bayeh, C. Z., & Damaševičius, R. (2023). Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms*, *16*(5), 236. https://doi.org/10.3390/a16050236

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, *80*, 83–93. https://doi.org/10.1016/j.eswa.2017.03.020

Pedraja, M., & Yagüe, J. (2001). What information do customers use when choosing a restaurant? *International Journal of Contemporary Hospitality Management*, 13(6), 316–318. https://doi.org/10.1108/eum0000000005966

Perikos, I., & Hatzilygeroudis, I. (2017). Aspect based sentiment analysis in social media with classifier ensembles. In *Annual ACIS International Conference on Computer and Information Science*. https://doi.org/10.1109/icis.2017.7960005

Phan, L., Pham, H., Nguyen, T., Huynh, K., Nguyen, T., Nguyen, T., Huynh, V., & Nguyen, V. (2021). SA2SL: From AspectBased Sentiment Analysis to Social Listening System for Business Intelligence. In H. Qiu, C. Zhang, Z. Fei, M. Qiu, & S. Kung (Eds.), *Knowledge Science, Engineering and Management* (pp. 647–658). Springer International Publishing.

Poché, E., Jha, N., Williams, G., Staten, J., Vesper, M., & Mahmoud, A. (2017, May). Analyzing user comments on YouTube coding tutorial videos. In *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*, 196-206. IEEE.

Pokharel, R., & Bhatta, D. (2021). Classifying YouTube Comments Based on Sentiment and Type of Sentence. *arXiv preprint arXiv:2111.01908.*

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, Orphée, Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, Salud María, & Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. https://doi.org/10.18653/v1/S16-1002

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. *In Proceedings of SemEval*.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Qader, W. A., Ameen, M. M., & Ahmed, B. (2019). An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. In *2019 International Engineering Conference (IEC).* https://doi.org/10.1109/iec47844.2019.8950616

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1), 9-27.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 77(2):257–286.

Rakib, Md. R. H. K., Pramanik, S. A. K., Amran, Md. A., Islam, Md. N., & Sarker, Md. O. F. (2022). Factors affecting young customers' smartphone purchase intention during Covid-19 pandemic. *Heliyon*, 8(9), e10599. https://doi.org/10.1016/j.heliyon.2022.e10599

Rokach, L., & Maimon, O. (2006). Clustering Methods. Springer eBooks, 321–352. https://doi.org/10.1007/0-387-25465-x_15

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE*, 16(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Sata, M. (2013). Factors Affecting Consumer Buying Behavior of Mobile Phone Devices. *Mediterranean Journal of Social Sciences*, 4(12). https://doi.org/10.5901/mjss.2013.v4n12p103

Savigny, J., & Purwarianti, A. (2017). Emotion classification on youtube comments using word embedding. 2017 In *International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. https://doi.org/10.1109/icaicta.2017.8090986

Schultes, P., Dorner, V., & Lehner, F. (2013). Leave a comment! An in-depth analysis of user comments on YouTube. *Wirtschaftsinformatik*, 42, 659–673.

Schwartz, S. H. (1977). Normative Influences on Altruism. In *Advances in Experimental Social Psychology*, 10, 221-279.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on YouTube. *Information Processing and Management*, 52(1), 46–60.

Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. (2010, April). How useful are your comments? Analysing and predicting YouTube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web* (pp. 891-900).

Singh, S. M., & Mishra, N. (2016). Aspect based opinion mining for mobile phones. *2016 2nd International Conference on next Generation Computing Technologies (NGCT)*, 540–546. https://doi.org/10.1109/NGCT.2016.7877474

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing*, 26(2), 102–113. https://doi.org/10.1016/j.intmar.2012.01.002

Steinberger, J., Brychcín, T., & Konkol, M. (2014, June). Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 24-30.

Tan, P., Steinbach, M., & Kumar, V. (2016). *Introduction to Data Mining*. Pearson Education India.

Trinto, N. I., & Ali, M. K. (2018). Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. https://doi.org/10.1109/icbslp.2018.8554875

Trivedi , R., & Raval , R. (2016). Consumer buying intentions towards smartphones: A conceptual framework. *International Journal of Applied Research* , 2(12), 736–742.

Tubishat, M., Idris, N., and Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing & Management*, 54(4):545–563.

Türkmen, H., Ekinci, E., & Omurca, S. İ. (2016). A novel method for extracting feature opinion pairs for Turkish. In *Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7-10, 2016, Proceedings 17*, 162-171. Springer International Publishing.

Twenefour, F. B. K. (2017). Major Determinants that Influence the Choice of Brand of Mobile Phone. *Open Journal of Statistics*, 07(04), 663–675. https://doi.org/10.4236/ojs.2017.74046

Uryupina, O., Plank, B., Severyn, A., Rotondi, A., & Moschitti, A. (2014). Sentube: A corpus for sentiment analysis on YouTube social media. In *Proceedings of the ninth international conference on language resources and evaluation* (pp. 4244–4249).

Wang, H., Lu, Y. and Zhai, C. (2011). 'Latent aspect rating analysis without aspect keyword supervision', In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 618-626. ACM.

Wang, T., Zhou, J., Hu, Q. V., & He, L. (2019, July). Aspect-level sentiment classification with reinforcement learning. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, 3316–3322.

Wu, H., Liu, Y., & Wang, J. (2020). Review of Text Classification Methods on Deep Learning. *Computers, Materials & Continua*, *63*(3), 1309–1321. https://doi.org/10.32604/cmc.2020.010172

Xu, Z., Chen, M., Weinberger, K. Q., & Sha, F. (2013). An alternative text representation to TF-IDF and Bag-of-Words. *ArXiv Preprint ArXiv:1301.6770*.

Xue, W., & Li, T. (2018). Aspect-based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043.*

Yiran, Y., & Srivastava, S. (2019). Aspect-based sentiment analysis on mobile phone reviews with LDA. *In Proceedings of the 2019 4th International Conference on Machine Learning Technologies (ICMLT 2019)*, 101–105. https://doi.org/10.1145/3340997.3341012

Zahoor, K., Bawany, N. Z., & Hamid, S. (2020). Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning. *In 2020 21st International Arab*

Conference on Information Technology (ACIT), 1-6.
https://doi.org/10.1109/acit50332.2020.9300098