CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Defect Identification for Cell Phones Using Product Reviews

by

Muhammad Zeeshan Younas

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2021

My dissertation work is devoted to My Family and My Teachers. I have a special feeling of gratitude for My beloved parents, whose prayers and support enabled me to succeed in all spheres of life. Special thanks to my supervisor, whose uncountable confidence helped me to reach this milestone.

CERTIFICATE OF APPROVAL

Defect Identification for Cell Phones Using Product Reviews

by

Muhammad Zeeshan Younas

(MCS193026)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ayyaz Hussain	QAU, Islamabad
(b)	Internal Examiner	Dr. Abdul Basit Siddiqui	CUST, Islamabad
(c)	Supervisor	Dr. M. Shahid Iqbal Malik	CUST, Islamabad

Dr. M. Shahid Iqbal Malik
Thesis Supervisor
December, 2021

Dr. Nayyer Masood
Head
Dept. of Computer Science
December, 2021

Dr. M. Abdul Qadir
Dean
Faculty of Computing
December, 2021

Author's Declaration

I, **Muhammad Zeeshan Younas** hereby state that my MS thesis titled "**Defect Identification for Cell Phones Using Product Reviews**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

Muhammad Zeeshan Younas
(MCS193026)

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Defect Identification for Cell Phones Using Product Reviews**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

Muhammad Zeeshan Younas
(MCS193026)

Acknowledgement

Thanks to Allah Almighty for blessing me with wisdom and strength to complete the dissertation. Being an MS graduate at Capital University of Science and Technology has been a magnificent and challenging experience. During the degree, I have found clear guidelines in shaping my academic career. Here is a humble tribute to all those people. I would like to express my sincerest appreciation to my enthusiastic supervisor, **Dr. M. Shahid Iqbal Malik** for his supervision, assistance, and immense knowledge. I am sincerely thankful to him for his constant support, motivation, and patience. His invaluable help of constructive comments and suggestions throughout the thesis work has contributed to the success of this research. It has been an amazing experience, and I thank him wholeheartedly, not only for his tremendous support. I would like to thank my family, who had motivated me continuously to achieve this milestone. A word of applause for my friends and classmates who had assisted me in sharing knowledge and other resources required to conduct research. Thank you all.

Muhammad Zeeshan Younas

Abstract

Product defects can negatively impact product revenue and global image, particularly in the social media environment. Immediate and correct detection of product defects will help manufacturers perform quality control and increase the competitive advantage of products. We created, implemented, and assessed a novel industry-specific smoke word list for defects identification in the cell phone manufacturing industry to provide cell phone manufacturers with business intelligence to continuously develop their quality. Previous work in automated defect identification has had success in the medicine, automobile household appliances industries, and toy industry. It is a different nature of the product that has another industry-specific smoke word list for defects identification in the cell phone manufacturing industry to provide cell phone manufacturers with business intelligence to develop their quality continuously.

In this study, we proposed a framework for defect identifications using online product reviews on cell phones. We conducted a set of experiments to identify defects using amazon products reviews. We used Part-of-Speech (POS), Word2Vec, BERT, Smoke Words (Unigram, Bigram, Trigram), Domain-specific word, and Sentiment analysis features for the experimental setup. We implement and assess a novel industry-specific smoke word list for defects identification in cell phone reviews. We used the three following sentiment analysis approaches in this study as a baseline. Namely, ANEW, Harvard GI Negative, and AFINN to compare our proposed smoke words approach. This study demonstrates that our proposed smoke words list (unigram, bigram, and trigram) outperformed as compared to Sentiment Analysis. Smoke word lists are more effective than sentiment analysis in defect identification. Logistic regression outperformed among all other classifiers with higher accuracy than others. BERT presented superior results in training data among all other features with 89.55% accuracy, and Word2Vec presented superior results in Validation data among all other features with 83.83% accuracy. This study shows that smoke lists for cell phone products can be more successful than sentiment analysis for detecting performance defects.

Contents

List of Figures

List of Tables

Abbreviations

ANEW	Affective Norms for English Words
BERT	Bidirectional Encoder Representations from Transformers
CC	Correlation Coefficient
ML	Machine Learning
NLP	Natural Language Processing
POS	Part-of-speech
SVM	Support Vector Machine
SGD	Stochastic Gradient Descent
W2V	Word2Vec
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1

Introduction

The spread of Internet availability lately has achieved monstrous changes in business insight. IHS [1] According to estimates, humans now run over 20 billion Internet-connected devices worldwide, which is expected to rise to over 75 billion by 2025. Online word-of-mouth (WOM), or the informal information flow from person-to-person, has grown dramatically in parallel with the expansion of Internet connectivity. Online word-of-mouth (WOM) communication has been acknowledged as a significant indication of customer opinion for items and a window into product sales and marketing quality, as the accessibility of the Internet has increased globally. Product features, attractiveness, and use are discussed informally by consumers through word-of-mouth communication (WOM). The importance of word of mouth in product sales has long been acknowledged. It can enhance customer awareness and might be one of the only trustworthy sources of information regarding the quality of experiential goods. With the advent of the Internet, word of mouth has expanded from local groups and communities to large-scale consumer networks [2, 3].

Product defects can negatively impact product revenue and global image, particularly in the social media environment. Immediate and correct detection of product defects will help manufacturers perform quality control and increase the competitive advantage of products. Product defects have such a significant negative influence on a company's competitive edge. Product defects can be discovered

quickly and efficiently, which can help firms in performance control and increase product competitiveness [4]. Manufacturers are concerned about product defects, and they perform recalls to avoid the spreading of safety-related and performance-related defects, which impose enormous financial expenses. Non-performance and safety-related defects cannot result in returns of products, but they can affect consumer satisfaction and repurchase desires. Customer-oriented techniques enable manufacturing to respond to global competitiveness problems by producing high-quality, highly dependable goods with short lead times and cheap costs. Auto companies have been observed to spend a significant portion of their sales returns on fixing faults that arise during the warranty term. Automakers spend a huge portion of their sales income (2.5 % to 3.0 %) on vehicle repairs during the warranty term. Recalls and customer complaints cost the car industry between 45 and 50 billion dollars each year [5].

It is tough to find and analyze product defects, including associated knowledge from a huge amount of user reviews. Traditional aspect opinion mining algorithms that attempt to find product features and opinions but that are insufficient to extract product defect-related information from user postings, their opinions heavily influence customer's behavior. Quality testing and comments from the after-sales service centers have traditionally been the primary sources of product defects information. The high cost, incomprehensibility, and hysteresis of product defect information collecting modalities based on such traditional sources for information are all drawbacks. Consumers are increasingly using social media to report product problems and express personal opinions [6, 7]. But Nowadays, consumers may openly share their views without exposing their real identity without fear of negative social media repercussions. For knowledge about product efficiency and functionality, customers depend heavily on the internet, like cell phones. Consumers provide information about product value, durability, and reliability to their suppliers, retailers, and their fellow consumers by outlets such as product reviews on numerous retailer websites. Product defects are more common, particularly in independent enterprises in underdeveloped countries, due to the lack of existing infrastructure. Companies need to pay closer attention to product quality control and the marketing of advanced analytics.

There are different types of product defects, and they mostly occur by the poor testing or designing of the product. Thus the product will not perform the desired function. The manufacturing defect doesn't involve the design but how the product was made, like materials used. If the product lacks the warning or instructions required for its safe usage, it is rendered defective. It is a marketing defect. The product that has any defect is not dangerous, so this can be sold by selling it at a discounted cost, and it must mention the defect. Any product feature that becomes the cause of the unexpected problem the product usability for a reason it was manufactured and designed. Legally these defects come in the context of the safety of the product. The product liability addresses the wounds/injuries caused by the defects in the product.The rise of social media has recently brought with it several useful product sources of information. Manufacturers have acquired a strong Involvement in social data sources and use them to discover product defects due to the benefits of comprehensiveness and practicality. Successful firms must acquire product-relevant knowledge both inside and outside to understand better the challenges influencing their goods. Consumer complaints are universally acknowledged as a significant source of product intelligence. Outsiders' or user communities' expertise is a valuable source of product-related business insight. Many companies have traditionally spent a significant amount of time gathering specific product usage information from practitioners to diagnose or explain problems or assign them to technicians who can resolve them. These so-called communities of practice are particularly essential for businesses that offer mechanical consumer products. They provide a library of prior increased tolerance that may be drawn upon for operational problem solving, product development, and other reasons [8].

The dominant aspect of this study is defect identification from online social media reviews. A massive sample of online reviews is sourced from amazon. In recent years, the cell phone industry has attracted great exposure for performance defects, including a large variety of hardware and software defects in various devices. A significant stream of product defect discovery research emphasizes this problem, producing industry-specific lists of "smoke-words" intended to recognize defects. Smoke words list: "list of words that are substantially more prevalent in defects

than in non-defects" [9]. Different types of features can be used to identify product defects, with some interdependence. In this work, we create, implement, and assess a novel industry-specific smoke word list for defects identification in the cell phone manufacturing industry to provide cell phone manufacturers with business intelligence to develop their quality continuously. Previous work in automated defect identification has had success in the medicine [10], automobile [8, 9, 11–13] household appliances industries [14–16], and the toy industry [17], but there has been no application to the cell phone. There is no single study on cell phone devices for defect identification in the prior studies. This platform was revised as minor in prior work using online reviews but not in this specific defect identification for the cell phone product category.

1.1 Background Knowledge

As social media platforms continually grow and gain more popularity, they become essential for manufacturers to gather consumer information about product defects. Researchers have created automated algorithms to identify product defect occurrences in social media, such as online reviews and discussion forums. Product defects are a key source of concern for both producers and buyers. Product defect detection is critical for manufacturers to avoid huge unnecessary product expenses. With the popularity of social media, social media data has become an essential source of information for manufacturers to collect defective details. Product defects severely harm the competitive advantage of a product.

Early identifying product defects can help manufacturers enhance product high quality and high advantage. Text mining algorithms must identify reviews that indicate defective items for companies and regulators to benefit from defect identification. Manufacturers and regulatory authorities cannot read every product review due to time and resource restrictions and the volume of data that come online every day [8]. Through social media websites such as online reviews and forums and online social networks, it is now easy to learn about other people's opinions on a wide range of topics. According to research, 81% of Internet users

have researched products at least once. Studies have shown that customer evaluations may influence others' perceptions and, as a result, purchases [18].

The automated computation may identify frequent user safety issues for particular product categories using online review sources such as Amazon.com. In key product categories such as "Toys and Games," Amazon.com has a huge database of online reviews. Over two million user evaluations have been posted in this category. It's a goldmine of possible product safety information [17]. Particularly in the age of social media, product defects may cost firms millions in litigation and negatively affect their sales, reputation, and goodwill. When it comes to detecting these product issues early on, online customer reviews may be useful. Regulatory agencies have many mass-produced items to monitor and examine. Still, the sheer amount of online product evaluations makes it difficult to sort through them and discover the defective outliers, which may be a major problem for consumers. To understand the information obtained, the quality of the reviews is just as important as the quantity [10].The process of discovering safety defects and responding to them is complicated from the manufacturer's standpoint. Before items reach customers, manufacturers may test their products in quality control departments to prevent safety problems. Determining the causes of problems can also be accomplished by reviewing warranty claims. Due to the difficulty in reproducing consumer usage circumstances in quality control testing and product recalls cost the United States over $1 trillion each year, it is critical to discover safety problems once items reach the mass market [15].

Online product reviews provide in-depth information about consumers' issues and allow manufacturers to have a broad understanding of rivals, which may help them better their products.However, it is typically impossible to manually interpret all evaluations on various websites for competing items and gain helpful information. Over the last decade, several researchers, particularly in computer science, have focused on effectively and efficiently evaluating such large amounts of consumer data. Several researchers on opinion mining for online reviews have claimed to derive sentiment polarity from online reviews at various levels. Nevertheless, most researchers in this discipline fail to consider how to make their results usable by

designers. Recently, a limited number of studies have been identified that use the most recent advancements in data mining and artificial intelligence in the design community [19].

1.1.1 Social Media Platforms

Social network has made our lives very much facile since we do not have to wait days or months to hear about someone we care about, for we can call them directly in Whatsapp, Skype, Facebook messenger, etc., within a second, which was once a blue moon upon a time. However, when there was no Social networking, parents were closer to their children, and children were closer to their parents. After sending off a letter, the parents were fully booked for their kids, and they were fully available. Nonetheless, now all family members are busy in their social networks, even though they live together physically. Still, mentally they are connected with their friends in their so-called social networks (laptops, iPad, cellphones).

Users of social media businesses expanded to hundreds of millions, and business uses of Facebook, Twitter, and other platforms began to take shape. User monitoring data was among the most extensive ever compiled by social media firms. Social networks are now growing rapidly, particularly in the context of the continued expansion of web-based services such as facebook.com and Amazon.com. Identifying significant people inside a social network is a major problem for social network analysis [20].

Many people use Internet-based social media networks sites to keep connected with current social updates and rely entirely on social networking for whatever they want to do. For instance, before buying goods from web-store, they go through the reviews of other purchasers, which is good, since before paying for something off, it is good to research. However, it cannot be 100% true since everyone is different, and thus their experiences and choices are different. Alternatively, many tourists book hotels after reading the reviews of the consumers, which can be beneficial for those hotels whose feedback is high, even-though if it is fake or wrong. On the other hand, it would be a loss for other hotels whose feedback is low. Still,

services are good, resulting in dismissing the employees in the hotels, which will lead to joblessness and poverty in society. Nowadays, people, businesses, and knowledge are linked via social networks, and they would need to be examined, not as isolated entities but rather as a part of everyday life. Because of the growth of computer networks, group solidarity at work and in the community has been diminished, and networked societies that are loosely connected and sparsely knit have taken over in their place. People's social capital improves due to the Internet since they can keep in touch with friends and family who live locally and far away. New tools are needed to assist individuals in traversing complicated, fractured, and networked societies and locating the knowledge.

The rise of social media on a global scale has fundamentally altered the way customers express their thoughts to businesses. With the use of social media, customers may openly share their emotions and experiences. We live in a period where individuals may become increasingly linked through formal and informal networks. Technology is making these social networks more frequent and accessible [21]. Social networks, specifically Facebook, Whats app, and Instagram, have brought individuals from very different places together who would not have met otherwise. It has made networks more frequent in our modern society. Customer-generated social media data thus becomes a crucial source of information for gaining a thorough grasp of items. In light of this, academics have focused their attention on social media data and have produced several essential study findings. Product rivals and consumer perceptions of such competitors were discovered using ensemble learning [22].

Product and competitor data from social media may be used in competitive analysis to assist firms in making better managerial decisions. He and his colleagues identified product attributes customers preferred and then compared other goods based on customer feelings about these qualities [23]. Aside from competition analysis, researchers were particularly interested in extracting consumer needs and subsequently enhancing product design. The estimated probability that particular words corresponded to specific features automated text assessment and proactive quality policies then found enhanced technological characteristics [24].

1.1.2 Product Reviews

The social network is a great advantage of science, but we must use it moderately. We should not 100% rely on what other people have addressed. Since we all are different, we should look for the reviews, but when it comes to making a decision, we have to decide at the end of the day. Thus, do not go under the negative influence of the social network, do not let others decide for you when you are paying for your trip or online purchase. Think twice before opening up about your problems with someone who seems nice to you on social networks. We can still purchase whatever we want with or without product reviews on the internet. Online Product reviews are conceivably the most helpful way to eliminate user's concerns regarding a product.

Product reviews influence a majority of the people in their purchases. Product reviews offer a vast and underused opportunity. Although many practitioners undoubtedly utilize internet reviews in their product development processes, the volume of online reviews is so large that it is almost difficult for practitioners to examine them systematically. The usage and impact of online reviews show that the review's rating has a minor impact on a customer's purchasing choice. Still, the number of reviews severely influences users' decision to buy those products [25]. Users may freely provide feedback on product and service defects due to online communities and other kinds of social media. Feedback is valuable to other customers in making decisions and to industry professionals in increasing the quality of their product or service [13]. In order to enhance their goods, manufacturers must understand their customers' emotional preferences and responses to product characteristics [26].

1.1.3 Defects Identifications

It is difficult to detect and assess product defects and the associated knowledge from many customer reviews. Traditional aspect opinion mining methods seek to discover product features and opinions. Still, they are insufficient to extract

product defect-related information from user reviews, despite their views substantially affecting consumer behavior. Traditionally, the major sources of product defect information have been quality testing and feedback from after-sales support centers. The health and safety of the people are closely related to the safety of different kind of products, as the vast majority relies on these products. The use of unsafe products led to the mortality of 22k people. It resulted in the injury of approximately 29.5 million people. According to statistics, it annually resulted in the loss of 700 billion people approximately in the United States of America. Thus the problem identification is very important [27].

The defects in manufacturing arise during the process of production. It can result from the use of low-standard materials or ignorance by the manufacturer and endangers the goal to achieve the specific product. The manufacturer has set the standard up, and he can conclude whether there is a defect by comparing. The identification of the manufacturing defect is easy. If the product deviates from the intended product design, the defect may be in manufacturing [28, 29]. The manufacturer is strictly liable even if he pays proper attention to manufacturing the product. During the process of production, the defects that are produced are usually inevitable. It is reasonable to hold the producer responsible, as they have more resources than the consumers to bear the loss.This defect can be overcome by changing the design. The design defect can be typical or obvious in the design. The obvious defect involves judgment based on specific standards of safety. These standards could be external or internal. In the case of the typical defect, it may or may not involve particular safety standards. History tells us that the judgment is based on the reasonable standard of expectation of the consumer. The determination of the reasonable expectation is difficult for the consumer due to technical complexity. It replaces the risk-utility standard because it cannot identify/judge the defect independently. The risk-utility standard considers various factors like the level of development of the technology and the product's time of circulation.

The defect of warning is not present in the product itself. If the consumer can avoid the predictable risk even if the producer does not give the warning or instruction, the product has no warning and instruction defects. If the producer knows, the

producer must specify the warning [19]. They must indicate how to safely use this product, as the consumer is unaware of the risk associated with the product. If the producer does not provide a reasonable warning, they are fully responsible for any hazard. It reduces the cost required for remediation and fix and increases the productivity of developers and staff. The risk of business is reduced. It results in improving the security of the application and the overall quality of code. Problem prevention is a good practice rather than waiting for the problem on its own. The early detection of the defect will cost less than the flaw identified later [16]. The static technique is used to ensure that the product produced has minimum or almost no defects. It helps in detecting any safety risks, so if there are defects that threaten the safety of the people, then it is removed. The new designs are created that are far better than the previous ones. The consumer is satisfied, and the product they purchase is up to their expectation. The use of automatic defect detection technologies offers clear advantages over manual detection. It not only adjusts to an inappropriate environment, but it also operates with great accuracy and efficiency in the long term. Defect-detection technology research may decrease production costs, increase efficiency and productivity, and quality of products while also laying the groundwork for the intelligent transformation of the manufacturing industry [9, 10, 14].

1.2 Problem Statement

According to our knowledge, few studies have addressed the problem of defect identification using customers reviews in the literature.

According to the best of our knowledge there exist no study that addressed product defect identification in cell phones.

In addition, prior approaches are not directly applicable to identify product defects in cell phone devices due to the different nature of product types and out of domain context.

1.3 Scope

The scope of this study is restricted to a huge volume of Amazon product specific reviews. Because online reviews are linked with particular products, it's easy to tell which reviews are relevant to the particular product. Furthermore, Social media sites like Facebook, Instagram, and Twitter were omitted due to the inaccessibility of many posts due to confidentiality restrictions and the widespread removal of specific product identification from social media postings. It's also challenging to acquire meaningful and relevant data for research because of the dispersed nature of social media postings. Moreover, this research considers only cell phone products reviews, and we categorized the cell phone devices into two primary divisions using the Amazon.com product scheme.

1.4 Research Questions

There are following two research questions addressed that will be answered in this study.

RQ1: Among applied machine learning methods, which one provide more robust performance in defect discovery of cell phone ?

RQ2: Which category of features will demonstrate the best performance in defect identification of cell phones using product reviews ?

1.5 Research Objectives

Our objective is to propose a methodology to identify software and hardware defects in cell phone devices using product reviews.

There are the following main objectives of this research work:

- The first objective of this research work is to identify defects in the cell phones domain using product reviews.

- The second objective of this research is to investigate various kinds of defects by which affected product industries can easily identify defects and improve quality assurance in cell phones.

Chapter 2

Literature Review

This chapter presents an overview of the literature review and highlights some significant key problems that led to the suggested solution. This section evaluates relevant studies on defects identification using online product reviews and threads. We go through the areas of coverage for prior work and the limitations and unaddressed problems. We end the section, in particular, by highlighting the manual smoke term curation in a subjective approach happens and the possibilities of refining this technique.

2.1 Defects Identification using Product Reviews

In the literature, few studies have addressed the problem of defect identification using product reviews. Previous work on defect identification used social media surveillance, text classification, and sentiment analysis approaches in Medicine [10], Automobile [8, 9, 11–13], Countertop Appliances and Dishwasher [14–16], and the toy industry [17]. Box-office prediction and influential user discovery [30–32]. Particularly in the field of defect or accident prediction [9, 16, 33, 34]. However, Cell phone devices, on the other hand, were not investigated in the literature. Additionally, due to the distinct nature of product types and the lack of domain context, earlier techniques are not directly relevant to identifying product defects

in cell phone devices. Data from social media has also been used to evaluate service quality [35–38]. Researchers used machine learning methods to develop automated product defect-identification models that can assist manufacturers in drastically lowering labor expenses [9, 13] Previous work on discovering product defects using social media data is summarized in the Table 2.2.

David Z. Adams et al. in [10] As a result of their linguistic content, texts are classified into preset groups. In the medical profession, this approach has been used to uncover defects in products review. Online feedback can assist manufacturers in improving their quality assurance and discovering SE problems early on by using a text categorization system with a significant volume of text. As a result of their research, automated detection of safety and efficacy (SE) issues in online pain treatment product reviews may be possible using sentiment analysis techniques and manual smoke-word dictionaries. According to their findings, the AFINN sentiment analysis was statistically inaccurate in forecasting SE concerns. However, the other sentiment analysis approaches fared poorly when contrasted to the unigrams, bigrams, and trigrams custom smoke-word lexicon.

In terms of identifying general SE issues, the smoke trigram word dictionary scored best, while the safety-specific smoke word dictionary fared best [14]. When it comes to sales, electronic word-of-mouth (e-NWOM) may have a significant influence. Previous fault discovery research has confirmed their findings in the automobile, users electronics, toy, and appliance industries. As well as the custom smoke word dictionaries produced in this study, the medical and industrial pharmaceutical medication and pharmaceutical company sectors can profit from an automated defect finding approach for joint and muscle pain alleviation products, specifically those sold over-the-counter. Their research developed a system for detecting SE issues in pain-relief product assessments concealed in huge reviews. The approach is efficient and automated.

Mat Winkler et al. in [17], Compared to standard sentiment analysis, the smoke-word list is superior in word overlap and usefulness. According to research, [39], buyers trust review language more than review summary statistics for individual

goods. This study examines the effectiveness of text mining in finding potentially hazardous toys for children. "Smoke words" were produced based on injury and memory text narratives. They are then applied to more than one million Amazon.com reviews, with higher scores indicating possible safety issues. In terms of word overlap and efficacy, they compare the smoke-word list with traditional sentiment analysis approaches. According to the researchers, they found that smoke-word lists differed significantly from conventional sentiment dictionaries. They could use them to detect safety issues in children's product assessments with statistical significance. Their research found that text mining is a great way to monitor safety concerns in children's toys and help avoid accidents caused by toys.

David M. Goldberg et al. [15], A Tabu search method was proposed for smoke-word curation, which outperformed the human-curated smoke-word list by a statistically significant margin when it came to identifying flaws, the researchers discovered. Having the capacity to detect and respond quickly when safety concerns arise benefits businesses and regulatory authorities. They used text mining to create "smoke words" in the countertop appliance and over-the-counter medication industries to discover defects. Based on previous research, they suggest several scientific modifications to increase the accuracy of industry-specific language. First, they substitute the personal manual curation of these terms with an automatic Tabu search technique, which statistically outperforms a sample of human-curated lists by a significant margin.

According to prior studies. Darren Law et al. [14], Consider using a previously created text evaluation framework to detect underperformance in huge home appliances, notably dishwashing machines. When used in conjunction with typical cross-domain sentiment techniques, we found domain-specific smoke and sparkle word lists highly correlated with probable faults. This study enhances the text analytic approach used in previous research by looking at significant home appliance performance issues. They discover that generic cross-domain sentiment methods may improve using domain-specific smoke and sparkle word lists that are significantly linked with possible faults. Dishwasher appliance quality management can

dramatically benefit from these results.

W.M Want et al. [26], The designed system helps consumers make buying choices, but it also helps companies better identify their products and rivals, offering insights into product growth. Recognizing how goods influence consumers help them make purchasing decisions and helps manufacturing companies create new, quality services. Traditional approaches based on handwritten Kansei questionnaires, on the other hand, are insufficient for attaining organizational, vast, and forever evolving environments. This work provides an unconstrained Emotion text mining technique for extracting and analyzing emotional data from web consumer evaluations for affective manufacturing. They present a semi-automated approach for obtaining a list of Kansei terms and characteristics using publically existing knowledge. Kansei terms and elements are general and may be used for a variety of goods and businesses. We gather product characteristics from customer reviews found online. They identify interface opinions from online customer research suggesting the Kansei words gathered and the retrieved product characteristics by categorizing the opinions as a collection of affective qualities and connecting those only with existing products.

H.Almagrabi et al. [18], The subject of review quality is connected to identifying opinion spam, making it a significant study area. It varies from defects identification in that spam feedback may or may not be of bad quality. Fake reviews can be high quality, specifically well-published, making them difficult to find. Product reviews have become increasingly essential as e-commerce platforms expand in popularity. In sentiment classification, researchers are interested in obtaining and summarising the vast amount of information included in product information and analyzing it. Review sites are increasing, but their trustworthiness and quality are being questioned. Even though many merchant platforms now analyze the usefulness of reviews individually, it is necessary to automate the process for at least half-time purposes. Whenever people assessments are absent, the primary reason is to give a helpful estimate. Secondly, it is necessary to rectify the skew in individual helpfulness judgments, as described in [40].

2.2 Defects Identification using Threads

Furthermore, several kinds of research [8, 9, 14, 16, 17] have focused on automatically detecting product defects using online discussion forums. This research followed a classification method to product defect identification, classifying a discussion thread as defect-related and otherwise based on a set of criteria that identify the thread. Previous studies have indicated various characteristics, including social features, linguistic features, and distinguishing words [16]. Prior research applied standard single classifier approaches to identify product defects [8, 16]. Several researchers have created various techniques for detecting defects based on smoke words. Machine learning approaches have been used in certain research to discover defect-related data [8, 12, 30, 31, 41]. In other research, probabilistic-graphic models (PGMs) are used to find concealed product defects. These studies demonstrate the use of social websites data to detect product defects [6, 13, 14, 42].

Identify Defects in consumer review data [11]. They introduced a robust probabilistic graphic method by selecting useful information using three filters: sentiment, component-symptom, and similarity. And on the other hand, they proposed a probabilistic graphical model to analyze the remainder data and find data related to defects. Using social media platforms, they offered a novel probabilistic graphic model for detecting defects. As a preliminary step, they choose informative data using three filters: sentiment filter, element filter, and similarities filter. Second, they use the developed probabilistic graphic model to assess the rest of the data related to defects. Defect kinds, faulty elements, and side effects are all included in their technique. Prior studies have excluded this information. Research studies in the automotive sector have confirmed the usefulness of our strategy and its improved performance over previous approaches.

According to Alan S. Abraham et al. [8], Many people depend on current knowledge because of the vast amount of information that is increasingly accessible. As demonstrated by the prevalence of news-monitoring and digital library program subscriptions, people remain on top of the newest trends in their fields of interest. Distinguishing what seems relevant to a company amid the storm of social media

postings is a tremendous problem. Producers and suppliers in the consumption production line, for example, are confronted with an unending snowfall of thousands of discussion forum comments. This article discusses and analyses text mining techniques for classifying user-generated material and extracting meaningful knowledge from the mass of messages. We use the automobile sector as a case study to implement a text-mining algorithm to remove element information from social networking sites. It is possible to automatically and precisely extract the automobile element that is the topic of a user's conversation using the models. Manufacturers, resellers, centers, and vendors benefit from this process since it quickly determines the unique words for each element segment.

Finding and summarizing product defects and associated information from vast numbers of user posts, on the other hand, is a challenging challenge. This research considers the challenge of identifying product defects. Using a Product Defect Latent Dirichlet Allocation model "PDLDA" model, we extract critical information about product faults from UGC by extracting interdependent themes (part, complaint, and remedy). Unlike other aspect summarization methods, this one recognizes the essential elements of product defects as interconnected three-dimensional pieces rather than as separate entities [6]. Domain-oriented features for extracting and summarizing critical and correct defect data The findings of their studies provide preliminary insight into the relative effects of domain features on defect detection and discovery. This work contributes to the existing literature in several ways. For social identity, users like to engage with their peers and are affected by their ideas. A participant's ingroup behavior changes when they can't identify with a 695 participant's ingroup model. This new frame 696 work has made it easier and more accurate to identify influencers in a given domain. Using this approach, we've augmented current theories with domain-aware characteristics in 697 cases [13].

Gruss et al. have employed Naive Bayes to identify numerical characteristics from numerical phrases occurring in postings by using these numerical features to detect product defects. In recent years, automated text mining algorithms have

made tremendous progress by finding a range of novel approaches and characteristics. This paper contributes to this body of knowledge by providing a unique text mining technique based on numerical expressions found in text documents. The occurrence and magnitude of numerical phrases are saved as document characteristics in this technique, which extracts, categorizes, and bins them [43]. Machine learning excels in defect-related text categorization, but it is unable to offer specific defect knowledge. When obtaining defect information, manual analysis is necessary.

Alan S. Abrahams et al. [9], Identify and analyze a new method and performance measurement framework for recognizing and prioritizing vehicle defects. There is much information about automobile defect presence and strategic importance on car enthusiast's online discussion forums [11], as they showed in the study. Despite its efficacy in identifying problems in other industries, traditional sentiment analysis failed to discriminate safety from performance defects and defects from non-defects. Another collection of automobile smoke terms has a greater relative incidence among faults vs. non-defects and among safety problems versus other posts on the site, which they have collected. In addition to Toyota and Honda, Chevrolet was utilized as a confirmation company for the smoke words identified in Honda and Toyota posts. In a different Vehicle Defect Discovery System (VDDS), they used our approach to determine a system that enables robust and generalizable defect detection and categorization. Social media postings may be used to improve vehicle quality management, as illustrated in this article.

Yao Lio et al. [12] Formulate a novel approach for detecting product defects from social network forums that addressing two flaws in the previous analysis, including the incomplete use of knowledge found in responses and the simple use of conventional single classifier methods. The detection of product defects via social networks, particularly internet discussion forums [11], has attracted academic interest. Nevertheless, previous techniques for identifying product defects have not thoroughly utilized the knowledge included in responses and still haven't effectively addressed the enormous complexity of feature vectors and dependencies between distinct types that exist.

Notably, they suggest a new technique to overcome those two limitations in this research. Contextual elements depending on replies are incorporated into this technique to make the most of the usefulness of replies in strengthening and enhancing the initial postings. In order to survive with the high heterogeneity and dependency issues, an inter ensemble learning approach is presented. Researchers found that our technique outperformed previous ways in detecting defects compared with existing methods, so both novelty in our method related to the enhancement. While its approach is designed to identify product defects, it is relatively generic and may be customized and widely used in various social networking text classification issues, such as predicting the value of online reviews[18, 29, 44–46].

Jian jin et al. [19], Many online reviews are created from time to time, providing a depth of information about consumer requirements. These evaluations aid designers in performing detailed competitor assessments. Many academics in information management and computer science have successfully derived and evaluating consumer needs from large amounts of opinions data during the last decades.

In [47], Usually, consumers write reviews when they have an exciting, positive, or negative emotion. From the bundle package buying paradigm perspective, this paper examined the role of users' sentiments in their online review placement period. The mystery box model's selling point is its vibrant advertising, appealing since it elicits pleasant feelings like the surprise.

They discovered that emotion does play a role in influencing users' online review behavior when it comes to services and products qualities. In particular, both emotional states boost consumers' feedback on product, service, and fulfillment aspects except for value. Organizations can effectively grasp customers' needs and preferences this way, especially when it comes to surprise box buying. The favorable eWOM impacts through online customer evaluations may also be used by businesses to attain emotional and intellectual empathy, which is required to draw a continuing number of buyers.

Product defect identification or incident detection, PGMs are an effective approach for identifying flaws. It was discovered that Latent Dirichlet Allocation (LDA) offers extra predictive power for software development. The objective here is whether utilizing themes may enhance predictive defect power over standard static and historical measures. They analyze the effectiveness of their topic-based metrics on statically and historic metrics independently since articles are generated from source code files and may also be coupled with pre-release defects to find defect-prone topics [48].

To detect real-time traffic events, Kinoshita and his colleagues developed a new probabilistic topic model. Data from probe-cars is used to keep track of current traffic, compared to the normal traffic, which is anticipated in advance using batch processing [49].

The Structural Topic Model was employed by Kuhn et al. in order to discover hidden aviation events. Using natural language processing (NLP) techniques, massive quantities of text data may be analyzed reasonably quickly and in large part automatically. Relevant findings may be obtained by interpreting the results by subject matter experts and through the additional study. Nature-language processing in aircraft safety reports has a wide range of commercial and academic uses. Nevertheless, topic modeling, a method that may uncover latent structure inside a document corpus, has been used in few published publications. [33].

The problem is that these models only extract key subjects from texts. It may or may not have anything to do with faults at all. On top of all that, none of the techniques listed above give specific defect information. According to Zhang et al., a new PGM that may absorb defect information through social media study comprised product models, years of manufacture, investigative elements, and indications [13].

Zhang et al. created a PGM called Product Defect Latent Dirichlet Allocation that considers defect resolutions [6]. PGMs do not necessitate extensive manual processing and tagging. Their results demonstrate that they have been excellent tools for processing data from social media while detecting product defects.

TABLE 2.1: Summary of various approaches investigated in the literature review

Researches	Research Approaches	Domain	Using Smoke Words	Medium	
				Online Reviews	Online Threads
Winkler et al., [17]	Smoke words	Toys and Games	✓	✓	
David Z et al., [10]	Smoke words	Medicines	✓	✓	
Goldberg et al., [15]	Smoke words	Home Appliances & Medicine	✓	✓	
Law et al., [14]	Smoke words	Home Appliances	✓	✓	
Abrahams et al., [9]	Smoke words	Automobile	✓		✓
Liu et al., [12]	Machine learning	Automobile	✓		✓
Zhang et al., [41]	Machine learning	Automobile	✓		✓
Gruss et al., [43]	Machine learning	Automobile			✓
Zhang et al., [13]	(PGM)	Automobile	✓		✓
Zhang et al., [6]	(PGM)	Automobile & Mackbook	✓		✓
Lu Zheng et al., [11]	(PGM)	Automobile	✓		✓

2.3 Research Gap

Previous work in automated defect identification has had success in medicine [10], automobile [8, 9, 11–13], household appliances industries [14–16], and the toy industry [17], but there has been no application to the cell phone. There is no single study on cell phone devices for defect identification in the literature. Additionally, due to the distinct nature of product types and the lack of domain context, earlier techniques are not directly relevant to identifying product defects in cell phone devices. This platform was revised as minor in prior work using online reviews but not in this specific defect identification for the cell phone product category. Prior studies in automated defect discovery have found industry-specific smoke words. Still, earlier smoke words are not directly relevant to identifying product defects in cell phone devices. It is a different nature of the product that has a different industry-specific smoke word list for defects identification in the cell phone manufacturing industry to provide cell phone manufacturers with business intelligence to develop their quality continuously.

Chapter 3

Proposed Methodology

It is critical for both companies and regulators that reviews clearly identifying defective products be evaluated by the text mining algorithm used in defect identification. Companies and regulatory agencies can't read every product's online review because of time and resource restrictions and the volume of data that comes every day on the internet. Moreover, the smoke term methodology aims to score each review in a corpus-based on how much defect-related jargon it seems to include. As a result, practitioners are only required to examine the top section of

assessments rather than all of them. As a result, subsequent research employing this methodology has assessed their methods' success by concentrating on the accuracy gained in the top N-ranked reviews as determined by the algorithms. Of course, the amount of evaluations that may be reviewed depends on the firm's or regulatory agency's demands and the industry and number of reviews. After sorting the reviews using a smoking word list from most relevant to least relevant. we use the number of defects discovered in the top 200 ranked reviews to measure

performance. We intend to demonstrate that our technique performs well regardless of the cutoff using this wide range of performance metrics. The graphical representation of the suggested approach is shown in Figure 3.1.

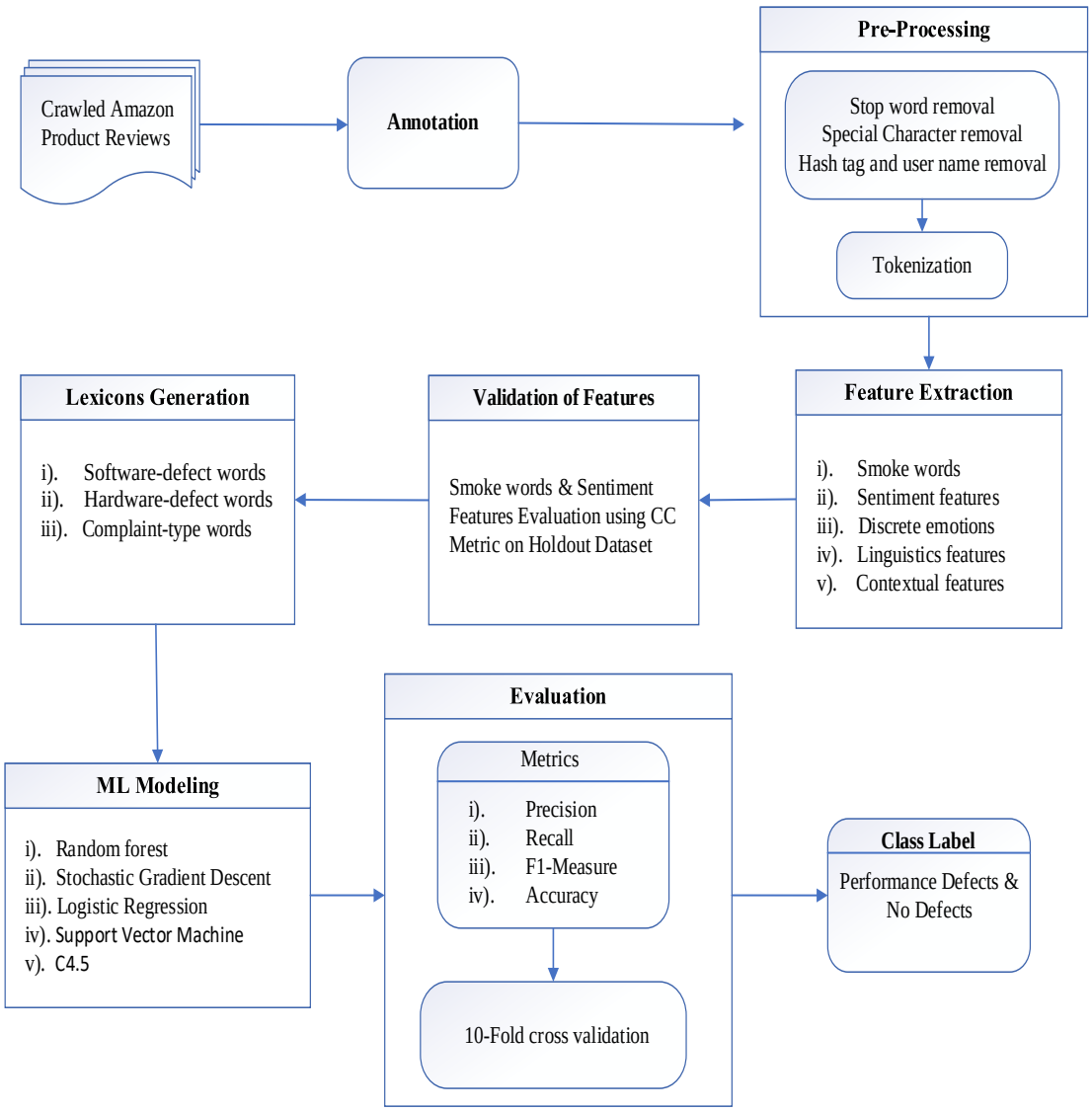

FIGURE 3.1: Block Diagram of proposed methodology

3.1 Dataset Description

The selection of datasets is a critical stage in a comprehensive assessment of the proposed system. A comprehensive dataset containing a lot of information, and we have to select data carefully to evaluate the suggested approach. In contrast to traditional data sources, online reviews are unstructured because the content does not match fields (tabular columns). Unless a customer leaves a review online, it may include a mix of different kinds of expression, such as expressive feedback that ranges from good to neutral to negative. These styles of communication might

even be used to represent non-emotive consumer feedback on product usage. According to a previous study, consumer-generated content is widely spread across 3rd party e-commerce platforms such as Alibaba, eBay, and others, making it harder to detect particular product-related concerns. Moreover, the language may be realistic when discussing domain-related issues like performance and safety concerns. The use of online datasets for safety concerns has lately become much more prevalent among companies and authorities. Since many customers share their daily product-related comments and experiences in the form of online reviews, organizations such as regulators may decipher these data to monitor potential safety problems. It is challenging to discover safety problems in internet evaluations, as only a tiny fraction of them do so. Involved parties find it difficult and time-consuming to manually go through millions of reviews to find only a tiny fraction of safety issue complaints in each product category. Several types of data sets contain distinct, interconnected types of information that may be accessed individually, combined, and controlled as a single entity.

3.1.1 Data Coding

"Performance defect" and "no defect" are mutually incompatible categories that we proposed by [9]. In the following, we'll explain how to distinguish between two different types of reviews.

3.1.1.1 Performance Defect

It refers to a non-serious failure of a product that is unlikely to cause harm if the product does not behave as the manufacturer expected or as the customer desires.. In most cases, these problems have to do with user satisfaction, such as how effectively the cell phone performs the hardware or software performance or how long the cell phone lasts before an issue occurs. So, here are some examples of a consumer complaining about cell phone issues. An example of a customer review marked as related to a "Performance Defect" is highlighted below.

"I hate this phone. The phone consistently drops calls. I tried to return it twice with a month of receiving the phone, but Amazon representatives coaxed me to try different fixes until the return deadline passed. It is now slightly better with calls but still fairly unreliable. Now the camera won't focus. Geeks at T-Mobile could not fix it after multiple resets, and I fear the long wait for service as the call will probably drop again, losing my place in line. Poor as a phone, pictures a blur, videos a blur as well. Poor processing speed and battery performance are third class. I should have bought the iPhone my family encouraged. Very very disappointed !! "

3.1.1.2 No Defect

Refers to reviews that include additional information but do not address a particular performance issue. Suppose the consumer doesn't identify any problems and only provides positive feedback with positive comments, which may or may not include vital information related to defects. In that case, this isn't considered a product failure. It may be Positive product reviews, advertisements, or general remarks. An example of a customer review marked as related to a "No Defect" is highlighted below.

"Best phone I've ever owned. I absolutely love it and am delighted with the fantastic phone quality for the low price. The camera is excellent. The style and size is perfect. Easily accessible for all applications and clear pictures. Fingerprint works very quickly. Charges fast and lasts 24hrs when I use my phone frequently for social media and phone interviews."

3.1.2 Crawling

We have crawled cell phone product reviews from Amazon.com, comprising over 52,000 cell phone product reviews, the world's largest e-commerce and massive online retailer. For this study, we utilized 30,000 reviews that spam from 2010 to

2020. We have specifically addressed two categories of cell phone reviews (Critical and Positive reviews). We subdivided the cell phones domain further into two major subcategories: Samsung and Apple iPhone. Table 3.1 presents the complete details of a dataset related to Critical and Positive reviews. And we can see in Table 3.2 and that two significant categories of reviews increased continuously in the period under review.

TABLE 3.1: Critical and positive reviews in amazon dataset

Product Name	**All Critical**	**All Positive**	**All Reviews**
Samsung	7,500	7,500	15,000
Apple iPhone	7,500	7,500	15,000
Grand Total	**15,000**	**15,000**	**30,000**

TABLE 3.2: Count of cell phone product reviews by year and category

	Categories		
Year	**Apple iPhone**	**Samsung**	**Grand Total**
2010	13	19	32
2011	79	90	169
2012	154	213	367
2013	271	345	616
2014	458	553	1011
2015	1,058	1,102	2160
2016	1,239	1,223	2462
2017	1,130	1,586	2716
2018	2,890	2,587	5477
2019	3,077	3,213	6290
2020	4,631	4,069	8700
Grand Total	**15,000**	**15,000**	**30,000**

3.1.3 Annotation

We randomly selected 2,000 unique reviews from the obtained dataset of "iPhone" and "Samsung" as a training set containing 1,500 critical and 500 positive reviews. All 2,000 records in the training set were individually tagged from the three annotators using the following cell phone tagging protocol described in Appendix A. At a significant public research university, teams of business undergraduates students were assigned the duty of labeling ("tagging") these online reviews, which means they had to examine each review and determine whether or not it included any product defects. Simultaneously, we randomly picked 200 reviews and tagged them by a mobile phone domain expert with the same tagging protocol for the results of three annotators' validation, who evaluated whether the defects were present or not. The second round of domain expert evaluations confirmed that 98 % of the results were linked to the three annotators labeled results. We collected the results from all three annotators in the form of excels files and categorized them by majority voting (best of three). The majority opinion was taken to avoid tied conditions. A training set was created using these labeled reviews. We marked a training set that classified reviews into two primary categories: "performance defect" and "no defect". Finally, they confirmed that 1503 out of 2000 review reports accurately pointed to Performance defects, and 497 out of 2000 review reports accurately referred to no defects as shown in Table 3.3. Table 3.4 presents the complete details of a training dataset. Defects are more common in low-star reviews (1, 2, and 3 stars) in our 2,000-review training set, as seen in the graph below in Figure 3.2, but defects are still be detected in high-star reviews (4 and 5) as well.

TABLE 3.3: Verified "Performance defect" and "No defect" by-product category in the training dataset.

Product Category	Performance defect	No Defect	Grand Total
Apple iPhone	754	246	1000
Samsung	749	251	1000
Grand Total	**1503**	**497**	**2000**

TABLE 3.4: "Performance Defect" and "No Defect" numbers per review star-rating

Star Rating	No Defect	Performance Defect	Grand Total
1	7	799	806
2	4	315	319
3	15	281	296
4	58	67	125
5	413	41	454
Grand Total	**497**	**1503**	**2000**

Defects are more common in low-star reviews (1, 2, and 3 stars) in our 2,000-review training set, as seen in the graph below in Figure 3.2., but defects are still be detected in high-star reviews (4 and 5) as well.

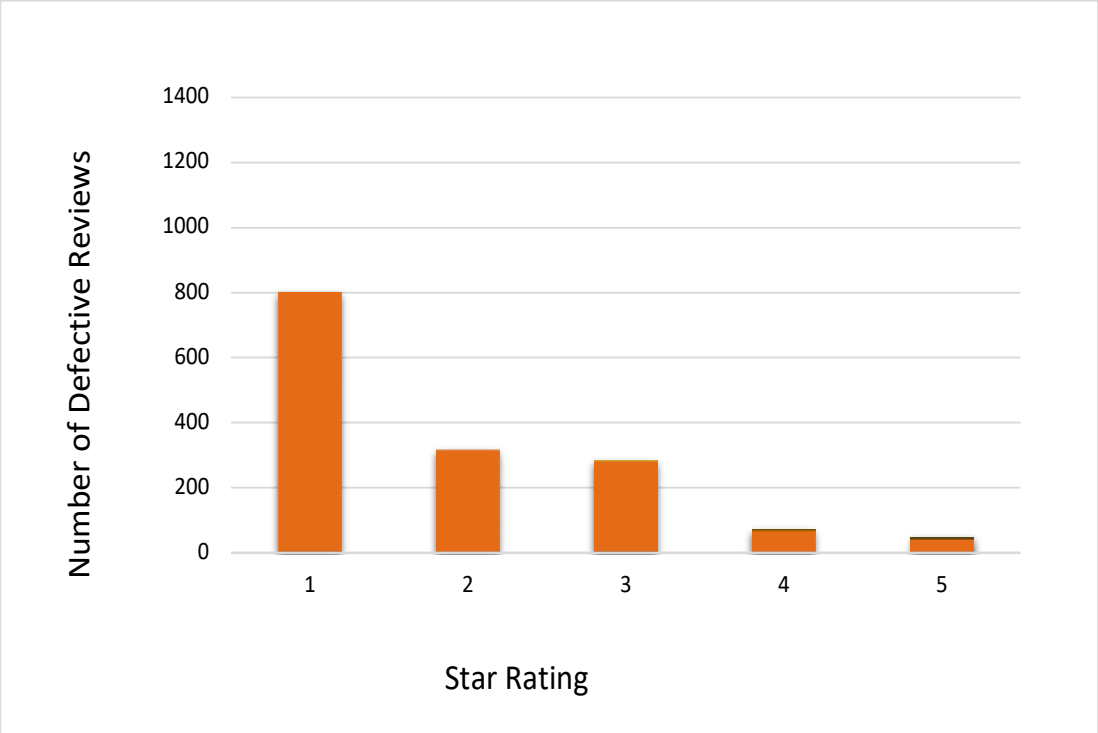

FIGURE 3.2: Number of Star Ratings In Performance Defects Reviews

3.1.4 Pre-processing

Preprocessing is a data mining approach that comprises converting a dataset into a format that can be understood. Datasets are frequently incomplete, with incorrect attribute values (Unknown Value), noisy data (meaningless data), and so on. The first procedure to be implemented in preparing data is data cleansing. This procedure focuses on deleting outliers that minimize duplication, handling incomplete information, noisy data, calculated biases, and information detection. Data mining is an extremely large number of data processing approaches. In such circumstances, analyses got more difficult while working with enormous volumes of data. We utilize data reduction techniques to get rid of this.
The objective is to enhance effectiveness in storage and to minimize data storage and analysis expenses. For machines to use the same to carry out activities like analysis, forecasts, etc., text pre-processing is preparing text data. There are many different stages in the preprocessing text, but in the article, you will only know about stopwords, why we remove them, and the libraries that can be used to delete them.

3.1.4.1 Stop words removal

The most popular terms in a language are stop words, such as top 20. The following mentioned some examples of English stop words are (is, am, are, we, this, as, he, she, on, that, the, it, at, be, by, these, from, a, to, in). Because these words have no significant significance, they must be removed from the text in order to obtain an accurate assessment. As a result, it is critical to eliminate these stopwords. We utilized the Natural Language Toolkit (NLTK) module to remove stop words from all data parameters in a dataset since it contains a list of stop words. NLTK compared its list of stop words to the tokenized list before removing stop words from the corpus. In every human language stop, words are accessible abundantly. When we delete these terms, we terminate low-level information from our text to focus on vital communication. In other words, the

removal of these phrases will not be detrimental to the model we train for our purpose [45].

3.1.4.2 Special Characters Removal

The process of removing letters, digits, and text fragments that may hinder our text analysis is known as noise reduction. Noise reduction is one of the most significant text preparation techniques. It's quite domain-specific. For example, in Amazon, noise can be any special character, except for hashtags, which denote concepts that characterize a Tweet. The noise issue is that it may produce consequences that are incompatible with your downstream operations. The removal of noise from data is critical since it can have a negative impact on accuracy. Null values, unnecessary punctuation, and other types of noise are common in datasets. There are several ways for removing noise, including ignoring the missing records, manually filling missing data, and filling with calculated values. Although there are a small number of occurrences with missing values, they are disregarded since it is the simplest and most efficient technique of managing missing data. Some punctuations may be considered tokens after tokenization, which may be superfluous (meaningless) and mislead us.

3.1.4.3 Lowercasing

All of our data sets, lowercasing is one of the simplest and most effective text preparation methods, despite being often overlooked. It can be used to tackle most text mining and natural language processing (NLP) problems, and it's especially beneficial if your dataset isn't too big. It also increases the consistency of expected production. This was most likely due to the dataset's mixed-case occurrences of the words, which provided insufficient evidence for the neural network to acquire the weights for the less common version properly. This problem is almost unavoidable when our data is little, and lowercasing is a great way to deal with sparsity.

3.1.4.4 Lemmatization

On the surface, lemmatization resembles stemming, in which the objective is to remove inflections and mapping a word to its root form. The only difference is that lemmatization makes an effort to do it correctly. It doesn't simply cut things off; it also converts words to their root. For instance, the term "better" would be mapped to "good." For mappings, it may utilize a dictionary-like WordNet or other specific rule-based methods. Using a WordNet-based method, here's an example of lemmatization in action. In our perspective, lemmatization does not give a substantial advantage over stemming in terms of search and text categorization. In reality, depending on the method you pick, it may be significantly slower than using a simple stemmer, and you may need to know the word's part-of-speech in order to produce a valid lemma.

3.1.4.5 Tokenization

Text can be split into several relevant bits in this procedure. Those parts are known as tokens. A piece of text, for example, is split into words or phrases. The input text can be divided into relevant tokens depending on the job involved. The words are split into words in this situation. The study utilized the most common and widely used library [50].

NLTK for this purpose. It is the most straightforward approach to tokenization. Even when white space is found, it is tokenized when a phrase or paragraph is divided. It's the most rapid tokenization approach, but it works with languages that divide the phrase into meaningful words. As in an ordinary language, there are unique sequences in formal languages. We typically name them as words, but people call them tokens in formal languages to prevent misunderstanding. But many techniques to combine tokens have emerged with a range of tokens throughout the time in NLP. But the idea behind tokenization remained the same to introduce certain finite concepts on the computers that it may combine to achieve the desired output.

3.2 Feature Extraction

We have extracted the following features for defect identification using amazon product reviews, and there are some terms from each cell phone domain-specific list. Table 3.9 shows the complete details of the proposed features.

3.2.1 Smoke Words

Use the Amazon.com data set to identify words (n-grams or smoke words) that have been present in each coding type, and the researchers conducted important words analyses. The machine learning (automatic method) algorithm was created by Fan Gordon and Pathak [51], and employs a prevalence measurement of correlation coefficient (CC) scoring. In contrast to more extensive sentimental analyses, industry-specific evidence has been disclosed that customers report product performance and safety defects using certain wording and sentences (n-grams) related to their product nature. For instance, although the word (airbag) is not connected to negative opinions in most emotional definitions, it probably implies a worry about safety in the context of online customers posting about a car. The main emphasis of a significant research stream in failure discovery is to generate lists of "smoke words" unique to industries, designed to identify language linked to defects [8–10, 14, 16, 17]. Using information retrieval algorithms to assess the relevance of words in a corpus is a crucial element in the smoke term curation process. Typically, researchers create a training sample to assess the relevance of words and a second holdout sample to assess the accuracy of these terms [9, 16].

The CC score has been extensively used in recent defect detection research, with outstanding results [10, 14, 17]. The weights these approaches allocate to each word, in the final analysis, are determined by the scores assigned by these techniques terms. Higher scores have more weight in marking a document's language relating to performance defects. Following recovery methodologies mentioned above, researchers must initially develop a Smoke Term List to gain a relevance score for each word in a corpus. Furthermore, smoke word lists are carefully

selected to keep just the phrases that are intuitively thought to give the most accuracy. The literature offers many justifications for removing words from the original list, but it also admits that the process is subjective. While the most pertinent phrases usually are considered most appropriate for smoking term lists, many of the relatively less pertinent terms need to be eliminated to guarantee that flaws are identified [8, 9, 14, 16, 17]. Indeed, after the top few hundred phrases, research has found a decrease in the quality of terms [8, 17].

The n-gram is a contiguous sequence of N elements from a given text or speech sequence in computer linguistics and probabilities. The objects might include the applications phonemes, phrases, letters, words, or base pairs. Typically the n-grams are taken from a word or language body. A size one n-gram is named a unigram, size two is a bigram, and size three is a trigram. And it depends on what size we require. The value of N is frequently used to refer to larger sizes, such as four-gram, five-gram, etc. Table 3.5 shows the example of unigrams, bigrams, and trigrams that have been used in this study.

TABLE 3.5: Unigrams, bigrams, and trigrams

Unigrams	**Bigrams**	**Trigrams**
disappointed	phone battery	phone not compatible
working	phone locked	button doesn't work
unlocked	not recommend	couple of months
stop	disappointed with	compatible with horizon
problem	phone call	lot of scratches

Performance defects: These are defects that impact consumer experience but are unlikely to cause damage or death. The following example is a list of terms that are much more common in defects than in non-defects.

3.2.2 Sentiment Analysis

In the subject of text mining, sentiment analysis is a growing topic of study. The computer handling of text's views, sentiments, and subjectivity is known as

sentiment analysis. Reviews are more likely to discuss these themes. Sentiment Analysis and Opinion Mining are two relative terms. They are expressing the same message. However, some academics say sentiment analysis and opinion mining have slightly distinct concepts [52]. It isn't easy to detect sentiments automatically, which are not represented in the lexicon but may also be found in news texts. However, they are less frequent than in product and film evaluations [53].

In filtering online reviews for the service characteristic of interest, we show that these approaches exceed basic techniques such as sentiment analysis. We also illustrate the performance of frequent sentiment dictionaries and randomness and defect detection rate while randomly sorting through all the reviews as baseline comparisons. Furthermore, we demonstrate the influence of incorporating star ratings in these assessments to increase the scores of reviews suspected of referring to safety flaws and reducing false positives. We have used the three following sentiment analysis approaches in this study. Namely, 1) Affective Norms for English Words (ANEW) [54], 2) Harvard General Inquirer 's Negative [55], And 3) AFINN [56]. We use sentiment analysis as a baseline for this study.

3.2.2.1 ANEW Method

Affective Norms for English Words (ANEW) is a project that aims to provide a set of normative emotional evaluations for a significant number of English words. The objective is to provide a set of linguistic materials assessed for enjoyment, arousal, and dominance to supplement the International Affective Picture System to make standardized resources available to researchers studying emotion and attention.

The presence of these emotional collections should aid in comparing results from various studies of emotion and permitting replication inside and between research laboratories examining fundamental or applied challenges in the study of emotion. ANEW utilizes a scale of 1 (most negative) to 9 (most positive); thus, valence levels of 5 are regarded neutral; values below "5" are considered to be positive, while values more than "5" are considered negative.

According to [57] technique for adjusting for negative values, we inverted the polarity of any word in the three words before the word that signified negation not or no. This was accomplished by computing (valence – 5) as that of the new valence number [54]. Table 3.6 shows the example of ANEW sentiment analysis that has been used in this study.

TABLE 3.6: ANEW Sentiment analysis for unigram, bigram, and trigram.The scale of 1 (most negative) to 9 (most positive)

Unigram	**Valence**	**Bigram**	**Valence**	**Trigram**	**Valence**
average	4	Sim card	5.42	phone not working	4.45
disappointed	3.22	bought phone	4.67	sim card phone	5.42
work	4.07	buy phone	4.67	battery not charge	6.94
bought	4.67	phone work	4.07	phone works great	5.01
sad	2	stop working	4.45	phone buy phone	4.67
good	7.47	apple store	5.42	buy phone from	4.67

3.2.2.2 Harvard GI Negative Method

A pre-set list of positives and negative terms is taken from the Harvard Inquirer sentiment analysis program. The polarity of texts based on the occurrence of the term in each category is analyzed. It ranks the emotional content of online reviews in our datasets and contrasts this with our technology. Every approach evaluates the material on several emotional aspects, including positive, negative, strong, passive, pleasure, suffering, etc.

In this situation, we utilize positive feedback to measure positive feedback and negative feelings irritators [46]. Table 3.7 shows the example of Hardvard GI Negative sentiment analysis that has been used in this study. For the Harvard GI Negative scoring method, a polarity "-1" refers to negative sentiment, "+1" refers to positive sentiment, and "0" refers to Neutral.

TABLE 3.7: Harvard GI Negative Sentiment analysis for unigram, bigram, and trigram.

Unigram	Polarity	Bigram	Polarity	Trigram	Polarity
sad	-1	phone works	1	phone not working	-1
disappointed	-1	Bad phone	-1	button stop working	-1
sad	-1	not working	-1	months stop working	-1
good	1	waste money	-1	spend extra money	-1
nice	1	phone locked	-1	phone works great	1

3.2.2.3 AFINN Method

AFINN is a sentiment analysis technique. AFINN is a collection of "positive" and "negative" words with valences ranging from +1 to +5. The most strongly negative is -1, while the least strongly negative is -5, and "Neutral" is 0. On the other hand, the most strongly positive is 5, while the least strongly positive is 1 [56]. Table 3.8 shows the example of Harvard GI Negative sentiment analysis that has been used in this study.

TABLE 3.8: AFINN Sentiment Analysis for unigram, bigram, and trigram. Range from -5 (negative) to +5 (positive)

Unigram	scores	Bigram	scores	Trigram	scores
sad	-2	phone work	0	phone stopped working	-1
bad	-3	phone phone	0	phone works great	3
disappointed	-4	stopped working	-1	button stopped working	-1
good	3	apple store	0	months stopped working	-1
working	2	phone works	0	good battery life	3

3.2.3 Domain-Specific Words

We create and analyze a specific class of domain-specific words derived from performance defects in this study, addressed only unigram in domain-specific words. There are the following three different domain-specific words category are discuss in this study. Namely, 1) Software defects, 2) Hardware defects, and 3) Complaint-type defects.

3.2.3.1 Software-defect words

List of words that are usually more prevalent in software defects than in other-defect words. Software-defects words are much more common in performance defects than in non-defects such as few software defect words are (slow, working, glitches, app, connect, poor, reset, updates, restart, and startup).

3.2.3.2 Hardware-defect words

List of words that are usually more prevalent in hardware defects than in other-defect words. Hardware-defects words are much more common in performance defects than in non-defects such as few hardware defect words are (screen, fingerprints, life, lagging, repair, temperature, overheats, processor, draining, quickly).

3.2.3.3 Complaint-type words

List of words that are usually more prevalent in complaint-types than in other-defect words. Complaint-types words are much more common in performance defects than in non-defects, such as few complaint-types words are (broken, tray, charger, earphone, cable, ejector, box, earpods, cellula, USB, glass, lens).

3.2.4 Discrete Emotions

Positive emotions are those that are associated with pleasure. According to the Oxford Positive Psychology Handbook, they are "pleasant or desired situational replies distinct from a pleasant sensation and generic positives," according to Oxford Positive Psychology Handbook [58].

In general, this concept states that positive emotions are more sophisticated and focused than simple feelings and are pleasant reactions to our surroundings (or our internal dialogue). On the other side, we usually don't like to feel unpleasant emotions. Bad emotions may be defined as " a disagreeable or disappointing

emotion, that people elicit as a result of an event or individual's negative influence" [59]. It's most certainly a negative emotion if an emotion discourages and drags you down.There are the following two types of discrete emotions, Positive (Joy, Trust, Surprise, and Anticipation) and Negative (Angry, Anxiety, Sadness, and Disgust).

3.2.5 Linguistic Features

Linguistics is the scientific study of this extraordinary skill. Linguistics scholars and researchers try to figure out what permits us to communicate and, more crucially for historians, how tendencies within our languages have evolved through time and shaped the society we live in today. Because language and literacy are so common in today's society - you're reading this article using linguistics - the topics of study covered by linguistics are vast: one may research linguistic developments throughout history, language usage now, how language influences our brain.

3.2.5.1 Part-of-Speech Tagging

Part-of-speech (POS) tagging has attracted a lot of attention since it is an essential part of most NLP systems [60]. Tagging parts of speech (POS) assigns a word in a text according to the meaning and connection of a particular word to neighboring words, a sentence, or a paragraph. The POS tagging falls under two separate classes. A rule-based POS tagger for English is created. The tagger uses a few basic rules and a tiny dictionary to produce tokens patterns [61].

The POS is a category of language characterized by its behavior in syntactic or morphological terms. Speech Components The tagger tool analyses current text and identifies each section based on its role in a phrase (morphological characteristics). This contains nouns, verbs, adjectives, and other similar words. The following 36 POS tags are present in Figure 3.3 with tags and their descriptions.

Number	Term	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential *there*
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	*to*
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

FIGURE 3.3: Alphabetical list of part-of-speech tags

3.2.6 Contextual Features

For more than ten years, organizational contextual features have been acknowledged as key factors in implementing evidence-based approaches in health care

settings. Furthermore, there was no unanimity across application scientists on the main features for integrating evidence-based procedures. The objective of this research was to identify the more frequently reported contextual organization characteristics, which affect the adoption and application of evidence-based practices in healthcare settings [62].

The contextual features are frequently perceived as incompatible with physicians' primary devotion to individual patients, and they are. Some scholars believe that contextual circumstances have little or no bearing on an ethical choice concerning patient care and that the doctor's responsibilities should be solely focused on the patient.

This viewpoint is outdated and theoretically wrong, in our opinion. Several of the previously listed reasons place real obligations and duties on both patients and doctors. The ethical challenge is to figure out how to properly assess the relevance of these contextual factors in a given situation.

3.2.6.1 BERT Model

Google AI Language specialists have published new research called Bidirectional Encoder Representations from Transformers (BERT). Due to the revolution, various NLP-related tasks, including question answering and natural-linguistic inference, have been introduced due to the revolution [63]. Researchers utilize the trained neural network to modify a new purpose-specific model after creating a neural network model for a specific objective, such as Image Net.

Contrary to the directed patterns that scan the text, the BERT Trans-former encoder reads the entire word sequence at a time (left to right or right to left). It is therefore categorized rather than directed as two-way. It would be more accurate. The model can learn a word's context depending on the whole surroundings (left and right of the word).

Its purpose is to combine the left and right contexts before creating a deep bidirectional from unlabeled texts. The pre-trained BERT models may therefore be

used for different NLP applications [64]. The bi-directionality of a model plays an important part in understanding language. Figure 3.4 shows an example of BERT with two sentences that both utilize the term bank.

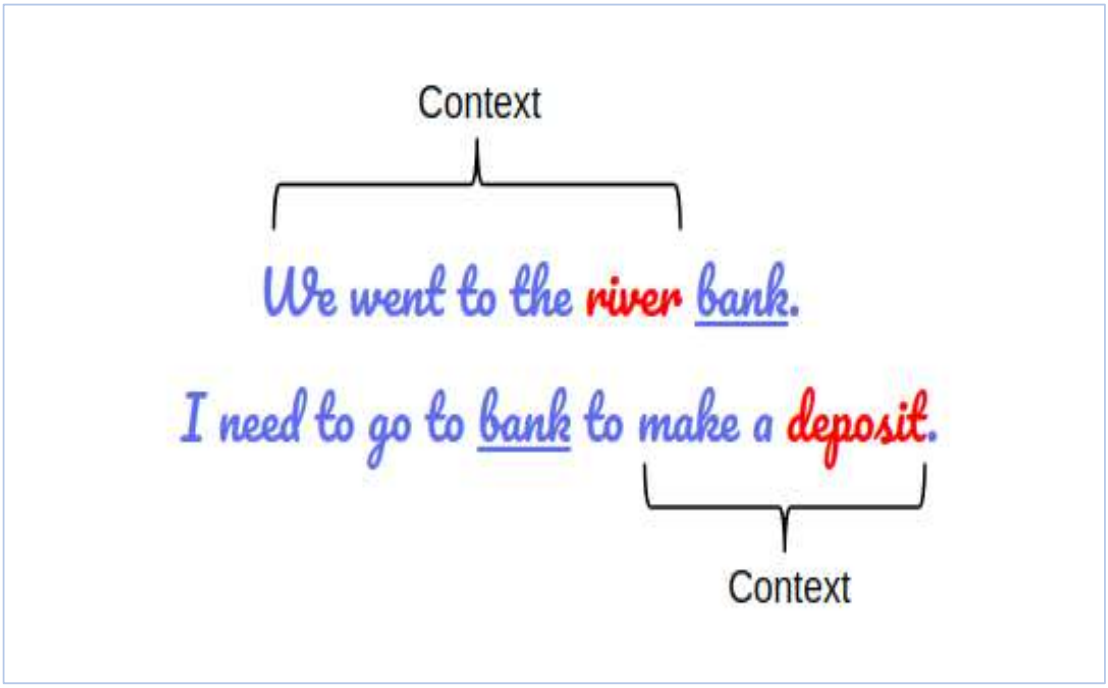

FIGURE 3.4: The BERT Example [63]

3.2.6.2 Word2Vec Model

Word2vec is a collection of models that describe distributed representations of words in a corpus. It is a method that takes text data as an input and generates word vectors representation as an output for each word. These word vectors can be represented as a large piece of text. The word2vec algorithm uses a neural network model to learn word associations.

Once it is trained, it can detect synonymous words or suggest additional words for a sentence. Word2vec is an advanced technique that is used for NLP. It is proposed by google [65] which is not an individual algorithm, but it comprises two different learning models: Continuous Bag of Words (CBOW) and Skip-gram.

TABLE 3.9: Features Extracted in this study

Types	**Features**
Linguistic features	Part-Of-Speech
Contextual features	BERT
	Word2Vec
Discrete Emotions	Positive
	Negative
Smoke Words	Unigram
	Bigram
	Trigram
Domain-Specific Words	Software-defects words
	Hardware-defects words
	Complaint-type words
Sentiment Analysis **(Baseline)**	AFINN
	ANEW
	Harvard GI Negative

3.3 Machine Learning Models

This section contains an abstract elaboration of machine learning models which we have implemented in our research.

The primary motivation for using more algorithms for classification is to examine the performance of various algorithms to detect performance defects. Another significant reason for selecting a machine learning algorithm is to do a comparison study of the algorithms that have been chosen. We have used five machine learning models in this study. Namely, 1) Random Forest, 2) Stochastic Gradient Descent, 3) logistic Regression, 4) Support Vector Machine, and 5) C4.5.

3.3.1 Random Forest

Random Forest is a strong tool that can provide results among the most reliable approaches available today [66]. A random forest is a classification group that works with the technique for bagging and boosting. The allocation of votes to the most accurate and appropriate input class is done using random forest classification models [67].

The random forest comprises several classifiers that help choose the best appropriate value for the input class in the voting process. In comparison with bagging or boosting, random forests provide numerous benefits, such as random forest handling outlying and proximity, in comparison with bagging and boosting. There are several forests, such as suggestion motors, image classification, and selection of function. Random forests It may classify faithful applicants for a loan, uncover fraud, and forecast illnesses. The Boruta algorithm is based on it and identifies significant aspects of a dataset[68].

3.3.2 Stochastic Gradient Descent

Numerous models have been shown to outperform gradient-enhanced decision trees. This is because boosting entails applying several models and combining their results. Stochastic Gradient improves the performance of a machine learning model by using a group of weak learners. Usually, decision trees are weak learners: their output, when combined, yields more accurate models. Unlike bagging, the boosting approach creates base models in sequential order[69, 70].

The accuracy of predictions is enhanced by creating several models in order and focusing on the difficult-to-estimate training instances. During the boosting phase, instances that are difficult to estimate using the prior base models emerge in the training data more frequently than successfully estimated examples. Each new basic model is intended to remedy the flaws in the prior base models [71].

3.3.3 Logistic Regression

Logistic regression is a commonly used standardized probabilistic statistical classification model for computing, technology, and social studies. The LR result of one sample, unlike the linear regression, is likely to be positive or negative, where the likelihood is dependent on aline sample measurement. Logistic regression is therefore extensively employed in the categorization [72]. A sigmoidal relationship between the probability of group membership and one or more predictor variables is a prerequisite for logistical regression.

Binary logistic regression is used for just two groups. However, three or more groups have to be decided between nominal and ordinal logistic regression. Nominal logistic regression must be applied if the irritating, neurotoxic, and embryotoxic categories are of no natural order [73].

3.3.4 Support Vector Machine

Support vector machine, commonly referred to as SVM, is a supervised machine learning algorithm that takes an input and produces a mapping function output using a labeled data instances [68]. SVM is considered one of the most successful and powerful algorithms for machine learning and has been widely utilized in several areas, such as text mining, image processing, and face recognition. The SVM method begins by converting the data set into high dimensions [74]. The method begins after the transformation to locate the most efficient hyper-planes, which divide the training set. The public-available Java-based implementation of LIB Support Vector Machinery (LIBSVM) can be incorporated with Weka [69] data mining tools.

3.3.5 C4.5

The advantage of the C4,5 is that models with continuous and discrete values may be readily interpreted and deployed. It also works with complex signals such as

ECG effectively. The C4.5 is an updated algorithm for ID3 [75]. Increase decision-making capabilities, and these trees can be implemented as a series of if-then rules. The tree is built from the top-down, beginning with the creation of the root node. The best-classified attribute is chosen as the test attribute at each node based on the largest information gain [76].

3.4 Statistical Analysis

This section contains a statistical analysis which we have implemented in our research.

3.4.1 Correlation Coefficient

Ng et al. [77] initially offered a correlation coefficient C as a substitute for X^2 as a technique for feature recuperation studies. Recall that the X^2 measurements of independence among two categorical variables are employed in classical statistics, R in their instance word t and relevance. Although X^2 was widely utilized in the machine learning process for feature selection, it has recently been shown that X^2 picks terms that show the importance of a text and those that show that a text is not relevant. Ng et al. empirical research [77].

Prior works [9, 10, 14–17] recommend the use of Fan et al. CC scores [51], a method for the retrieval of information that employs the X^2 distribution to give scoring for the applicant conditions of use (n-grams). Higher scores suggest that the word is often used in the positive class and seldom in the negative class. High-value words can therefore be strong positive class predictors. The technology is insufficient for the identification of the words for applicant smoke on its own. At the same time, it may not be easy to over-fit the pieces of training, and it does not take care of the interplay between multiple words. A contingency table for the word “j” in the training data set is defined in Table 3.10

TABLE 3.10: Contingency table for word "j" [51]

	Relevant (Performance Defect)	**Non-relevant (No Defect)**	**Sum**
Word j=1	A	B	A + B
Word j=0	C	D	C +D
Sum	A +C	B+D	N

A = number of relevant documents in which the word "j" appears.
B = number of non-relevant documents in which the word "j" appears.
C = number of relevant documents in which the word "j" does not appear.
D = number of non-relevant documents in which the word "j" does not appear.
N = is defined as = (A + B + C +D)

$$\mathbf{CC} = \frac{\sqrt{N} \times (AD - CB)}{\sqrt{(A+B) \times (C+D)}} \tag{3.1}$$

3.5 Evaluation Metrics

The performance metrics used in this study are well-known metrics to evaluate the results. We have used four performance metrics in our study, Accuracy, Precision, Recall, and F1-Measure.

3.5.1 Accuracy

Accuracy is our first evaluation metric which talks about how much proportion of our selected data have been identified correctly. It shows the correct proportion of predicted outcomes either true positive or true negative. The standard formula used for assessment of results is given below:

$$\mathbf{Accuracy} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP}+\mathbf{TN}+\mathbf{FP}+\mathbf{FN}} \tag{3.2}$$

3.5.2 Precision

Precision is our second evaluation metric which talks about how much our applied ML model produced accurate results. The standard formula used for assessment of results is given below:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}} \tag{3.3}$$

3.5.3 Recall

Recall is our third evaluation metric which talks about how much instances our applied ML model captured as actual positive (True Positive) .The standard formula used for assessment of results is given below:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \tag{3.4}$$

3.5.4 F1-Measure

F1-Measure is our fourth evaluation metric which is used when we need to seek relation between precision and recall an uneven class distribution. The standard formula used for assessment of results is given below:

$$\textbf{F}_\textbf{1} - \textbf{Measure} = 2 \times \frac{\textbf{precision} \times \textbf{recall}}{\textbf{precision+recall}} \tag{3.5}$$

3.6 Tools and Programming Languages

1. **Python** is used for the implementation of all algorithms.
2. **Microsoft Excel** is used to store all calculated results and dataset.
3. **Weka** is a well-known data mining tool is used for features selection.

4. **Jupyter Notebook** is an open-source online tool that allows us to create and share documents, including live code, mathematics, visualizations, and text.

Chapter 4

Results and Analysis

This chapter contains a comprehensive description of all results which are collected from our set of experiments. This section presents our findings and analyses the performance of the smoke word dictionaries and sentiment analysis. We conducted a set of experiments to identify defects using amazon products reviews. We have defined two class labels: i.e., Performance defect (Yes) and No defect (No). We used Part-of-Speech (POS), Word2Vec, BERT, Smoke Words (Unigram, Bigram, Trigram), Domain-specific word, and Sentiment analysis features for the experimental setup.

A significant stream of product defect discovery research emphasizes this problem, producing industry-specific lists of "smoke-words" intended to recognize defects. Smoke words list: "list of words that are substantially more prevalent in defects than in non-defects" [9]. Different types of features can be used to identify product defects, with some interdependence. We implement and assess a novel industry-specific smoke word list for defects identification in cell phone reviews. We have used the three following sentiment analysis approaches in this study as a baseline. Namely, ANEW, Harvard GI Negative, and AFINN to compare our proposed smoke words approach. All these selected features are evaluated using the following machine learning models:

1. Random Forest

2. Stochastic Gradient Descent
3. logistic Regression
4. Support Vector Machine
5. C4.5

Every machine learning model is evaluated with 10-fold cross-validation. For evaluating results collected from these machine learning models, we utilized precision, recall, f-measure, and accuracy as the evaluation metrics. After organizing all our selected machine learning models' results, we examine that Logistic Regression is the only ML model that performed best for all sets of features, so we chose it for further examinations. We evaluate our proposed Smoke Words performance with state-of-the-art baseline Sentiment Analysis using two provided class labels (Performance Defects vs. No Defects). This study shows that smoke lists for cell phone products can be more successful than sentiment analysis for detecting performance defects.

4.1 Experimental Setup

Following hardware and software are used for the analysis.

Hardware Requirements:

Following hardware is used for features selection.

1. Processor Intel(R), Core(TM) i5-3230M, CPU 2.60GHz.
2. 12 GB RAM
3. 750 GB Hard disk

Operating System and Development Software:

Following software is used for features selection.

1. Windows 10 Home
2. Python 3.9
3. Microsoft Excel 2019
4. Weka 3.8
5. Jupyter Notebook 6.4
6. Sublime Text

4.2 Experiment 1: Generation of Smoke words lexicons

Three lists of smoke-words have been generated, "a list of words that are substantially more prevalent in defects than in non-defects" [9] in the training group across all subcategories. We utilize the CC score to rank each of the top-ranked terms in the Unigram, bigrams, and trigrams training set. There were 5,324 distinctive unigrams, 33,026 bigrams, and 47,861 trigrams in the training set of 2000 reviews. The following three Smoke words dictionaries have been generated. Some 20 top-ranked unigram, bigram, and Trigram by relevance (CC score) in each performance defects review are listed in Table 4.1.

- Unigrams (single words), this lexicon comprised top-250 ranked words that are the more common in performance defects vs. no defects, as calculated by the CC metric.
- Bigrams (two sequential words), this lexicon comprised top-350 ranked words that are the more common in performance defects vs. no defects, as calculated by the CC metric.
- Trigrams (three sequential words), this lexicon comprised top-350 ranked trigrams that are the more common in performance defects vs. no defects, as calculated by the CC metric.

Table 4.1: Top-20 ranked Words Indicative of Defects for Unigram, bigram and Trigram from the training sample by CC score [51]

CC Score Rank	Unigram	CC Score	Bigram	CC Score	Trigram	CC Score
1	not	46600.33	it not	15950.37	it will not	8411.84
2	but	28074.08	will not	15921.64	not compatible with	7075.63
3	for	27920.25	not work	15207.18	but it not	7075.63
4	work	27258.73	not buy	14559.78	will not work	6911.80
5	turn	26024.60	one not	14517.81	phone doesn't work	6572.23
6	all	25382.43	phone not	14399.90	it doesn't work	6395.87
7	with	25061.54	doesn't work	14310.41	not buy phone	6214.63
8	use	24702.07	turn it	13091.44	not buy from	5835.67
9	buy	22911.31	use it	12749.45	one but it	5636.85
10	return	21798.89	could not	12168.71	had return it	5636.85
11	charge	21642.15	but it	11714.14	didn't come with	5636.85
12	battery	21274.39	return it	11456.54	phone will not	5636.85
13	out	21011.64	didn't work	11353.68	it not unlock	5430.90
14	doesn't	21000.43	it doesn't	11061.76	not buy it	5430.90
15	does	20847.94	not working	11039.76	it didn't work	5216.97
16	if	20157.85	one it	10333.19	send it back	5216.97
17	did	20088.64	it does	10312.53	phone but it	5216.97
18	they	19595.01	cause it	10078.77	it not unlocked	5216.97
19	get	19493.39	it will	10067.92	one of them	4994.03
20	didn't	19401.01	very disappointed	10043.55	new but it	4994.03

4.3 Experiment 2: Evaluation of Smoke words and Sentiment lexicons

Finally, when we created three smoke words dictionaries from 2000 reviews in the training set, the remaining 28,000 holdout sample reviews were evaluated by Correlation Coefficient (CC) score based on how frequently the smoke-words list of top-250 ranked Unigram, top-350 ranked bigram, and top-350 ranked Trigram used in the holdout set, however the relative weight rank of each unigrams, bigrams, or trigrams as determined by the Correlation Coefficient metric. A total score for each smoke word in the evaluation was calculated by adding the Correlation Coefficient (CC) score for the relevant words from its related smoke words lexicon. We also assessed the performance of our three Smoke Word Lists in a holdout set using three common sentiment analysis dictionaries: (ANEW) [54], Harvard GI Negative [55], and AFINN [56] to rank the online reviews in our dataset according to their emotional content and compare this ranking to our method. Once again, each review was assessed by adding up a total score based on the relative weight of each three sentiment analysis methods. After Applying the 6 scoring methods

to the unseen holdout dataset (3 sentiment analysis methods and 3 smoke-words analysis methods), it was used to rank and score the holdout set (28,000 reviews). Respectively each smoke word has its lexicon comprised of specific terms with a weight for each word, same as for each sentiment analysis method which has pre-defined dictionaries comprised of specific terms with a weight for each word. We scored a complete holdout sample for each review with the help of 6 scoring methods terms weights and then cumulated the total score for all matched terms in each review. We sorted the reviews for each scoring method in descending order, then selected top-200 ranked reviews, and bottom-200 ranked reviews from the holdout sample to generate a "Validation" set of 2400 reviews. After all this, we finalized our Validation set of 2400 reviews. From the Validation set, top-ranked reviews are predicted to be negative that have performance defects by the scoring method in each case according to ranking. In contrast, bottom-ranked reviews are predicted to be positive that have no defects.

Again, we tagged our validation set (2400 reviews) through cell phone domain experts to reduce bias using the cell phone tagging protocol described in Appendix A. We provided them text-only without any indication of review scores. This last and third tagging session aimed to evaluate the performance of various sentiment analysis and smoke word methods to determine the defects in the cell phone industry. The results of the "smoke word" and "sentimental analysis" were compared to evaluate their best performance.

Table 4.3 shows the defects identified in the validation set using sentiment analyses and smoke word dictionaries. Each smoke words list (unigram, bigram, and trigram) outperformed compared to Sentiment Analysis. Smoke word lists are more effective than sentiment analysis in defect identification. According to our findings, the AFINN sentiment analysis approach finds more defects than the other two, ANEW and Harvard GI Negative Sentiment analysis techniques. Smoke trigram dictionary outperformed than other two smoke dictionaries in terms of performance defects.

Finally, once all the reviews were tagged and our labeled dataset was obtained, we checked and removed duplicate reviews from the validation set that appears in the top-200 and bottom-200 for multiple methods; the validation set leaves 1840 unique reviews.

To find out if there is a relationship between the user's star rating and the occurrence of defects, we also compared the number of "performance defects" and "no defects" per star rating review to evaluate the results between them. The most common compliments were seen in reviews with high star ratings. Therefore, assessing reviews with extreme star ratings is risky because the only star-rating approach is incapable of identifying defects. Usually, the existence of defects is maximum in lower star ratings than the higher star rating reviews. Still, we also found evidence that the proportions of many 4-star and 5-star rating reviews apparently discussed defects – see Table 4.2. However, Even 1-star and 2-star rating reviews were well discussed with no defects.

TABLE 4.2: Number of Performance Defects and No Defects Per Star Rating Reviews

Star Rating	No Defect	Performance Defect	Grand Total
1	9	531	540
2	5	170	175
3	12	168	180
4	153	98	251
5	659	35	694
Grand Total	**840**	**1000**	**1840**

4.4 Experiment 3: Generation of Domain-Specific Lexicons

In this experiment, we have generated three domain-specific lexicons from performance defects reviews. We used here our 2,000 labeled dataset. We randomly selected 2,000 unique reviews from the obtained dataset and tagged them by three annotators. All 2,000 records in the training set were individually tagged from the three annotators using the following cell phone tagging protocol described in Appendix A, which means they had to examine each review and determine whether or not it included any product defects. Once they categorized the dataset into two class labels, "performance defects" and "No defect." after this step, they further tagged the labeled data set into three different domain-specific categories. Namely, 1) Software defects words, 2) Hardware defects words, and 3) Complaint-type words. We collected the labeled dataset from all annotators and organized the results for all three domain-specific categories by majority voting (best of three).

The majority opinion was taken to avoid tied conditions. Simultaneously, we randomly picked 200 reviews and tagged them by a mobile phone domain expert with the same tagging protocol for the results of three annotators validation, who evaluated whether the domain-specific label was valid or not. We evaluated three domain-specific categories where we see that 168 out of 1503 defective reviews

accurately pointed to software defects words, 610 out of 1503 defective reviews accurately pointed to hardware defects words, and 903 out of 1503 defective reviews accurately pointed to complaint-type words.

After we established three performance defects domain-specific categories, we used labeled reviews to construct domain-specific lexicons. We created domain-specific word lexicons from defective reviews using the correlation coefficient (CC) score [51]. The CC score was used as the particular measure to calculate the relative occurrence of each term (word or phrase) for each domain-specific attribute value. This method allows us to construct unigrams (single word) and bigrams (two words) for each domain-based category using the methodology as described in earlier studies [9, 10, 14, 17]. Three lists of domain-specific lexicons have been generated, a "list of words that are substantially more prevalent in defects than in non-defects" [9] in the domain-specific group across all subcategories.

We utilized the CC score and sorted all terms in descending order to rank each of the top-ranked terms in the Unigram and bigrams for each domain-specific lexicon. There were 2,758 distinctive unigrams in software-defect words, 3,772 in hard-defect words, and 4,281 in complaint-type words. The following three domain-specific Lexicons have been generated by relevance (CC score) in each performance defects review. Smoke word lists are more effective for all three domain-specific subcategories.

4.4.1 Software-defect words lexicon

Software defects are considered to have a detrimental influence on product quality attributes such as adaptability and manufacturability [78]. 56% percent of all mobile users have been generally troubled by mobile software or applications. In most situations, applications crash or launch slowly for many reasons, and test circumstances might create these issues. Modern gadgets are often accompanying the use of multipurpose software. Cellphone market growth shows significant smartphone software demand to make a profit, and the product must share high-quality and up-to-date bug-free solutions. A professional quality assurance team

and users feedback can help improve the quality of the product and decrease failure risks. There are 20 top-ranked software-defect words (unigrams) listed in Table 4.4.

4.4.2 Hardware-defect words lexicon

Hardware defects are unavoidable and unpredictable incidents throughout the usage of any smartphone. If such occurrences are not correctly diagnosed, they can be reduced product reputation [79]. Samsung's first Note 7 version was flopped due to overheating and exploding defective batteries. Over 2 million devices had to be remembered, and the product was discontinued. Samsung recall cost was estimated to reach $5.3 billion in 2017, and It was a vast recall cost that sent shockwaves to the industry [80]. Product defects can negatively impact product revenue and global image, particularly in the social media environment. The defects in manufacturing arise during production, and It can result from low-standard materials or ignorance by the manufacturer and endangers achieving the specific product. There are 20 top-ranked hardware-defects words (unigrams) listed in Table 4.4.

4.4.3 Complaint-type words lexicon

A complaint-type defect occurred when the user complained about damaged or missing parts upon arrival, like (scratches and dents on the body, cracked cell phone screen, faulty and used cell phone delivery), and likewise many more. It can be related to software and hardware defects. We have mentioned the list of complaints-type in Appendix A. Product, and competitor data from social media may be used in competitive analysis to assist firms in making better managerial decisions. Identified product attributes customers preferred and then compared other goods based on customer feelings about these qualities. There are 20 top-ranked hardware-defects words (unigrams) listed in Table 4.4.

TABLE 4.3: Top-20 ranked Words Indicative of Defects for Unigram, bigram and Trigram from the training sample by CC score [51]

Scoring Methods	No Defect	Performance Defect	Grand Total	Accuracy	Precison
AFINN					
Top 200	33	167	200		
Bottom 200	160	40	200	82%	83.5%
	193	**207**	**400**		
ANEW					
Top 200	52	148	200		
Bottom 200	179	21	200	79.5%	74%
	231	**169**	**400**		
Harvard GI Negative					
Top 200	35	165	200		
Bottom 200	141	59	200	76.5%	82%
	176	**224**	**400**		
Unigram					
Top 200	29	171	200		
Bottom 200	199	1	200	92.5%	85.5%
	228	**172**	**400**		
Bigram					
Top 200	13	187	200		
Bottom 200	191	9	200	94.5%	93.5%
	204	**196**	**400**		
Trigram					
Top 200	8	192	200		
Bottom 200	185	15	200	95%	96%
	193	**207**	**400**		
Grand Total	**1212**	**1188**	**2400**		

Table 4.4: Top-20 ranked Words Indicative of Defect for Unigram from the Domain-Specific Words Lexicons Sample by CC score [51]

CC Score Rank	Software-defect words lexicon (Unigram)	CC Score	Hardware-defect words lexicon (Unigram)	CC Score	Complaint-type words lexicon (Unigram)	CC Score
1	freeze	2816.70	battery	16636.74	lock	20876.47
2	turn	2557.05	screen	10009.72	return	10286.07
3	shut	2201.89	charge	6815.28	unlock	8876.89
4	call	2145.95	turn	6612.62	broken	8436.66
5	start	1904.47	phone	5997.05	cause	8019.68
6	off	1808.64	charging	5377.92	stopped	7697.16
7	calls	1796.28	issue	4975.97	charge	7667.07
8	started	1680.95	bad	4778.97	not	7585.46
9	error	1644.28	camera	4313.57	refurbished	7532.02
10	restart	1497.82	stopped	3933.14	carrier	7528.55
11	laggy	1497.82	poor	3883.16	damage	7492.45
12	slow	1454.61	times	3818.6	beware	7307.44
13	reboots	1422.35	stop	3818.6	accessories	7120.47
14	wifi	1422.35	touch	3800.39	horrible	7102.62
15	fix	1411.42	speaker	3798.53	stolen	7030.26
16	message	1411.42	disappointed	3709.23	scratches	6998.69
17	setting	1411.42	fingerprint	3608.05	glitches	6877.10
18	waste	1411.42	button	3553.56	empty	6865.03
19	connect	1337.74	waste	3504.72	disappointed	6864.60
20	starting	1319.59	take	3494.96	card	6782.60

4.5 Experiment 4: Training and Validating Classifiers

In our third experiment, we are interested in examining the influence of proposed features with their evaluated results for defect identification. Two types of the labeled dataset are used here for our experimental framework, the first is the training, and the second is the validation dataset. Our experimental setup consists of six types of features set that described in Table 3.9.

We implement five machine learning classifiers. Namely, 1) Random Forest, 2) Stochastic Gradient Descent, 3) logistic Regression, 4) Support Vector Machine, and 5) C4.5, with 10-fold cross-validation and consider accuracy, precision, recall, and F1 measure as evaluation metrics. We investigate the impact of each category of features for performance defects.

4.5.1 Classification Performance on Training and Validation Data

This portion was carried out to analyze the performance analysis of various classification methods, and extensive experiments were performed using training and validation datasets. We applied the five ML classifiers on training and validation data to evaluate the performance of classifiers. The outcomes of all these experiments are discussed here, including the accuracy, precision, f1-measure, and individual classifiers accuracies for each technique using 10-fold cross-validation. In this sub-experiment, we ran each ML classifier for each feature on training data, evaluated their performance based on accuracy, and selected the best one for further proceeding. From these results, we just gather accuracy and F1-measure.

Table 4.5 illustrates the results of all features for training data with respect to accuracy using the selected machine learning models. It is evident that logistic regression mostly outperformed among all other classifiers with higher accuracy than others. BERT presents superior results among all other features with 89.55%

accuracy, and Word2Vec shows second-best results with 89.10%. Whereas all remaining features, including Domain-Specific, Smoke Words, POS, Discrete Positive, Sentiment Analysis, and Discrete Negative show 83.50%, 82.55%, 81.90%, 81.65%, 80.86%, and 75.65% accuracy score respectively. Table 4.6 illustrates the results of all features for training data with respect to F1-measure with the selected machine learning models.

TABLE 4.5: Accuracy using Training Data

Features \ **Classifiers**	**Logistic Regression (%)**	**SVM (%)**	**C4.5 (%)**	**Random Forest (%)**	**Stochastic Gradient Descent (%)**
Sentiment Analysis	80.86	80.20	80.55	80.25	80.85
Discrete Negative	75.65	75.15	75.50	74.50	75.20
Discrete Positive	81.65	80.40	81.64	81.40	81.50
POS	81.90	80.70	80.20	81.95	80.80
Smoke Words	82.55	82.25	82.50	81.85	82.50
Domain-Specific	83.50	82.50	83.50	80.40	83.00
Word2vec	89.10	84.65	80.65	85.30	86.95
BERT	**89.55**	**89.45**	**82.6**	**87.8**	**89.50**

In this portion, we applied the same five machine learning classifiers on validation data to evaluate the performance of classifiers for each feature. The outcomes of all these experiments are discussed here, including the accuracy, precision, f1-measure, and individual classifiers accuracies for each technique using 10-fold cross-validation. In this sub-experiment, we ran each ML classifier for each feature on the validation set, evaluated their performance based on accuracy, and selected the best one for further proceeding. From these results, here we also gather accuracy and F1-measure for the validation set.

Table 4.7 illustrates the results of all features for validation data with respect to accuracy using the selected machine learning models. It is evident that logistic regression mostlyoutperformed among all other classifiers with higher accuracy

TABLE 4.6: F1-Measure using Training Data

Features \ Classifiers	**Logistic Regression (%)**	**SVM (%)**	**C4.5 (%)**	**Random Forest (%)**	**Stochastic Gradient Descent (%)**
Sentiment Analysis	79.40	77.30	82.40	79.30	77.60
Discrete Negative	70.23	73.20	72.60	65.50	73.50
Discrete Positive	78.70	74.60	78.80	79.80	78.00
POS	79.30	75.50	77.50	79.90	77.30
Smoke Words	81.60	82.20	81.60	81.60	81.70
Domain-Specific	82.50	80.90	81.90	80.10	82.40
Word2vec	87.50	82.60	80.60	84.50	86.40
BERT	**87.30**	**88.00**	**82.00**	**86.00**	**89.00**

than others. For the validation dataset, Word2Vec presents superior results among all other features with 83.83% accuracy, and BERT shows second-best results with 82.37%. Whereas all remaining features, including Smoke Words, POS, Sentiment Analysis, Domain-Specific, Discrete Negative, and Discrete Positive, show 75.52%, 74.50%, 69.68%, 69.34%, 66.34%, and 65.19% accuracy score respectively. Table 4.8 illustrates the results of all features for training data with respect to F1-measure with the selected machine learning models.

BERT outperformed all other selected features in training data with 83.83% accuracy and 82.10% f-measure scores, and Word2Vec outperformed all other selected features in validation data with 89.55% accuracy and 87.30% f-measure scores.

4.5.2 Feature-wise Analysis

As discussed earlier, we use only a single machine learning classifier for further proceeding based on the outperformed accuracy. It is evident that logistic regression mostly outperformed among all other classifiers with higher accuracy than others,

TABLE 4.7: Accuracy using Validation data

Features \ Classifiers	**Logistic Regression (%)**	**SVM (%)**	**C4.5 (%)**	**Random Forest (%)**	**Stochastic Gradient Descent (%)**
Sentiment Analysis	69.68	62.43	69.6	70.31	62.49
Discrete Positive	65.19	65.05	65.92	64.18	65.14
Discrete Negative	66.34	63.13	66.34	65.43	62.93
Domain Specific Words	69.34	61.84	69.59	69.34	64.34
POS	74.50	72.22	71.90	74.38	72.22
Smoke Words	75.52	73.12	75.32	75.50	73.50
BERT	82.37	79.82	70.92	82.20	79.23
Word2vec	**83.83**	**82.39**	**76.90**	**83.39**	**80.81**

and we investigated each category of features' impact on performance defects. We conducted this experiment using the machine learning model Logistic Regression of individual features from each category for training and validation set.

4.5.2.1 Smoke Words Performance

Figure 4.1 illustrates the results of all smoke word features set (Unigram, Bigram, and Trigram) for Training data with respect to the accuracy, precision, and f1-measure using Logistic Regression as a machine learning model.

It is evident from the following graphical representation that Unigram outperformed among all of them with 80.50% accuracy, 80.10% precision, and 80.00% f1-measure scores. Bigram showed the second-best performance with 79.45% accuracy, 79.35% precision, and 79.00% f1-measure scores. The remaining Trigram performance is lower than Unigram and Bigram, with 75.15% accuracy, 75.00% precision, and 75% f1-measure scores.

TABLE 4.8: F1-Measure using Validation data

Features \ **Classifiers**	**Logistic Regression (%)**	**SVM (%)**	**C4.5 (%)**	**Random Forest (%)**	**Stochastic Gradient Descent (%)**
Sentiment Analysis	59.00	59.80	74.80	69.70	60.20
Discrete Positive	63.80	44.50	64.30	63.40	44.60
Discrete Negative	64.00	56.40	66.10	65.20	61.80
Domain Specific Words	69.30	59.80	70.40	69.00	63.10
POS	73.70	71.30	71.55	74.00	71.60
Smoke Words	73.80	73.10	75.80	73.50	73.60
BERT	79.60	79.30	70.90	81.22	79.00
Word2vec	**82.10**	**82.00**	**76.80**	**82.30**	**80.80**

Figure 4.2 illustrates the results of all smoke word features set (Unigram, Bigram, and Trigram) for Validation data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Trigram outperformed among all of them with 71.82% accuracy, 71.80% precision, and 71.50% f1-measure scores. Bigram showed the second-best performance with 71.23% accuracy, 71.00% precision, and 70% f1-measure scores. The remaining Unigram performance is lower than Trigram and Bigram, with 66.75% accuracy, 66.00% precision, and 65.50% f1-measure scores.

4.5.2.2 Domain-Specific Word Performance

Figure 4.3 illustrates the results of all Domain-Specific word features set (Complaint-type defect words, Hardware defect words, and Software defect words) for Training data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Complaint-type

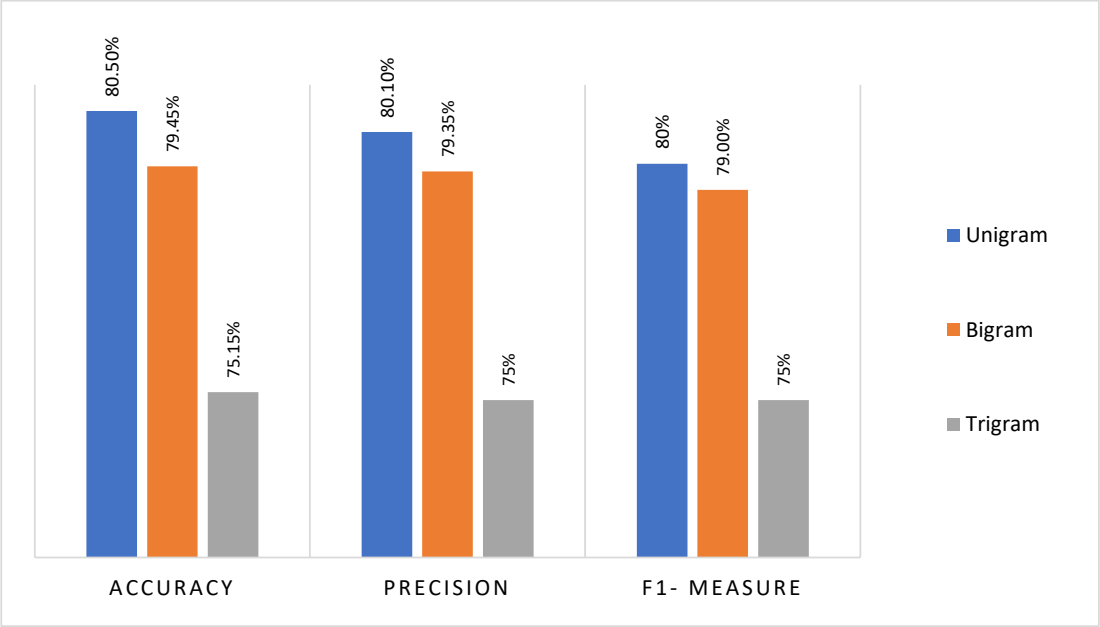

FIGURE 4.1: Smoke Word Features Performance Using Training Data

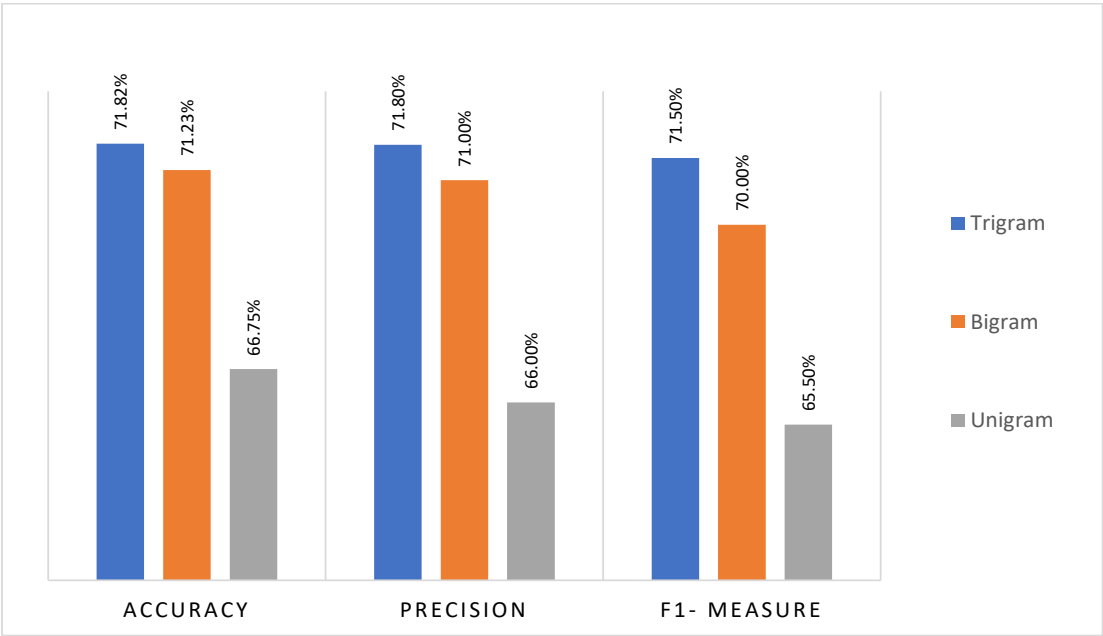

FIGURE 4.2: Smoke Word Features Performance Using Validation Data

defect words outperformed among all of them with 81.65% accuracy, 80.00% precision, and 79.40% f1-measure scores. Hardware defect words showed the second-best performance with 77.50% accuracy, 77.00% precision, and 76.% f1-measure scores. The remaining Software defect words performance is lower than Complaint-type defect words and Hardware defect words, with 77.00% accuracy, 76.50% precision, and 75.50% f1-measure scores.

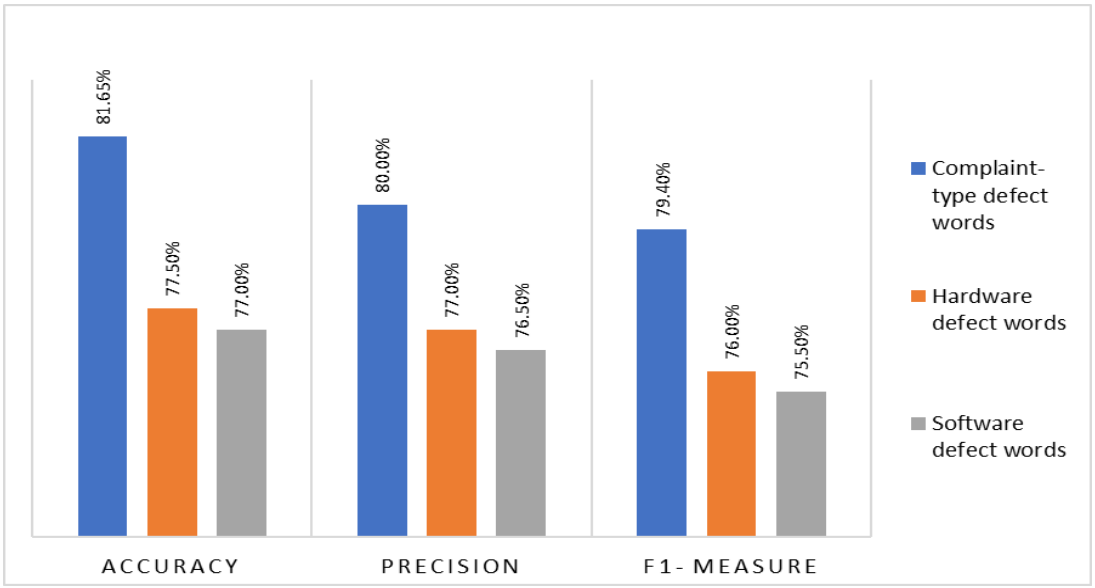

FIGURE 4.3: Domain-Specific Word Features Performance Using Training Data

Figure 4.4 illustrates the results of all Domain-Specific word features set (Complaint-type defect words, Hardware defect words, and Software defect words) for Validation data with respect to the accuracy, precision, and f1-measure using Logistic. It is evident from the following graphical representation that Software defect words outperformed among all of them with 68.81% accuracy, 67.90% precision, and 65.00% f1-measure scores. Complaint-type defect words showed the second-best performance with 68.20% accuracy, 68.00% precision, and 67.00% f1-measure scores. The remaining Hardware defect words performance is lower than Software defect words and Complaint-type defect words, with 65.36% accuracy, 64.00% precision, and 63.00% f1-measure scores.

4.5.2.3 Sentiment Features Performance

Figure 4.5 illustrates the results of all Sentiment Analysis features set (AFINN, ANEW, and Harvard GI Negative) for Training data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that AFINN outperformed among all of them with 78.00% accuracy, 77.00% precision, and 76.50% f1-measure scores. ANEW showed the second-best performance with 77.50% accuracy, 76.50% precision, and

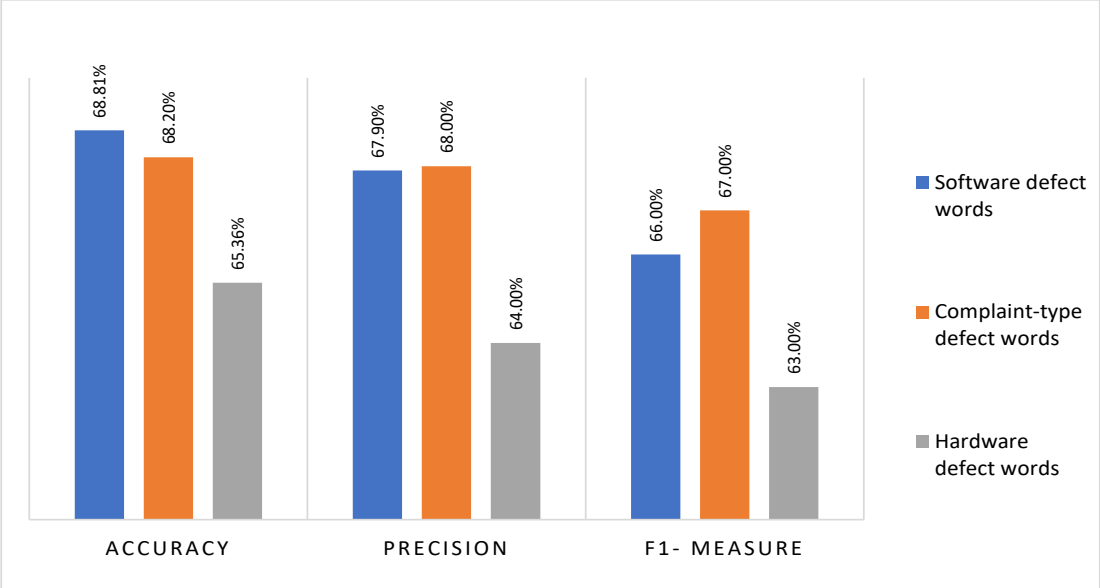

FIGURE 4.4: Domain-Specific Word Features Performance Using Validation Data

76.00% f1-measure scores. The remaining Harvard GI Negative performance is lower than AFINN and ANEW, with 75.00% accuracy, 75.00% precision, and 74.00% f1-measure scores.

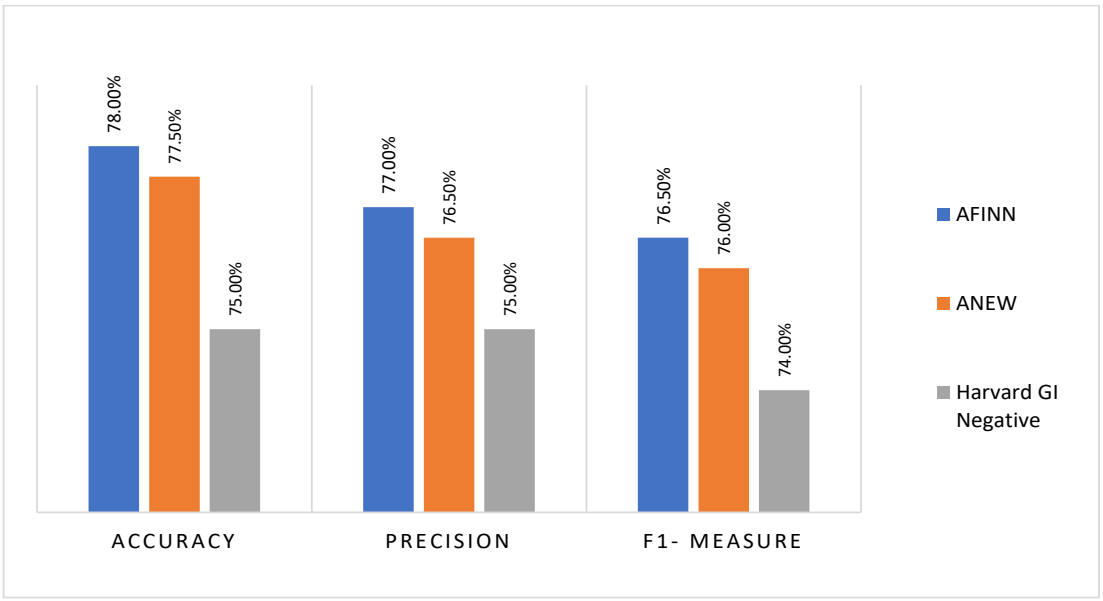

FIGURE 4.5: Sentiment Analysis Features Performance Using Training Data

Figure 4.6 illustrates the results of all Sentiment Analysis features set (AFINN, ANEW, and Harvard GI Negative) for Validation data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the

following graphical representation that AFINN outperformed among all of them with 61.45% accuracy, 61.00% precision, and 58.00% f1-measure scores. ANEW showed the second-best performance with 57.85% accuracy, 57.00% precision, and 46.00% f1-measure scores. The remaining Harvard GI Negative performance is lower than AFINN and ANEW, with 56.82% accuracy, 56.00% precision, and 53.00% f1-measure scores.

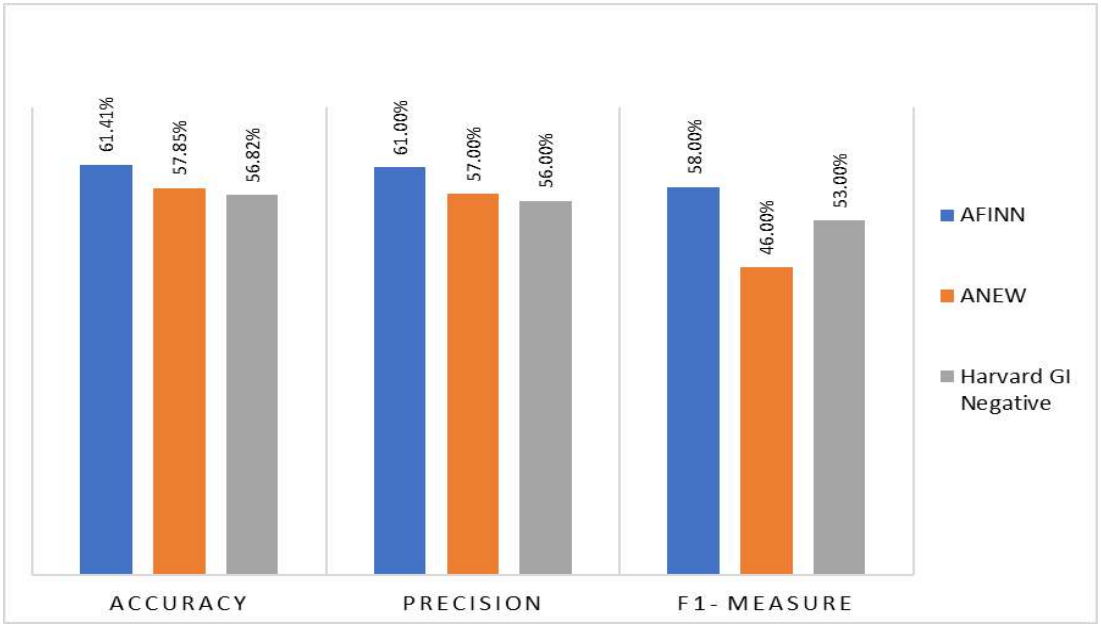

FIGURE 4.6: Sentiment Analysis Features Performance Using Validation Data

4.5.2.4 Discrete Emotions Features Performance

Figure 4.7 illustrates the results of all Discrete Positive Emotions features set (Anticipation, Joy, Surprise, and Trust) for Training data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Joy outperformed among all of them with 81.70% accuracy, 81.00% precision, and 80.00% f1-measure scores. Trust showed the second-best performance with 78.75% accuracy, 78.00% precision, and 77.00% f1-measure scores. Anticipation showed the third-best performance with 78.70% accuracy, 78.20% precision, and 77.00% f1-measure scores. The remaining Surprise performance is lower than Joy, Trust, and Anticipation, with 77.00% accuracy, 76.80% precision, and 76.00% f1-measure scores.

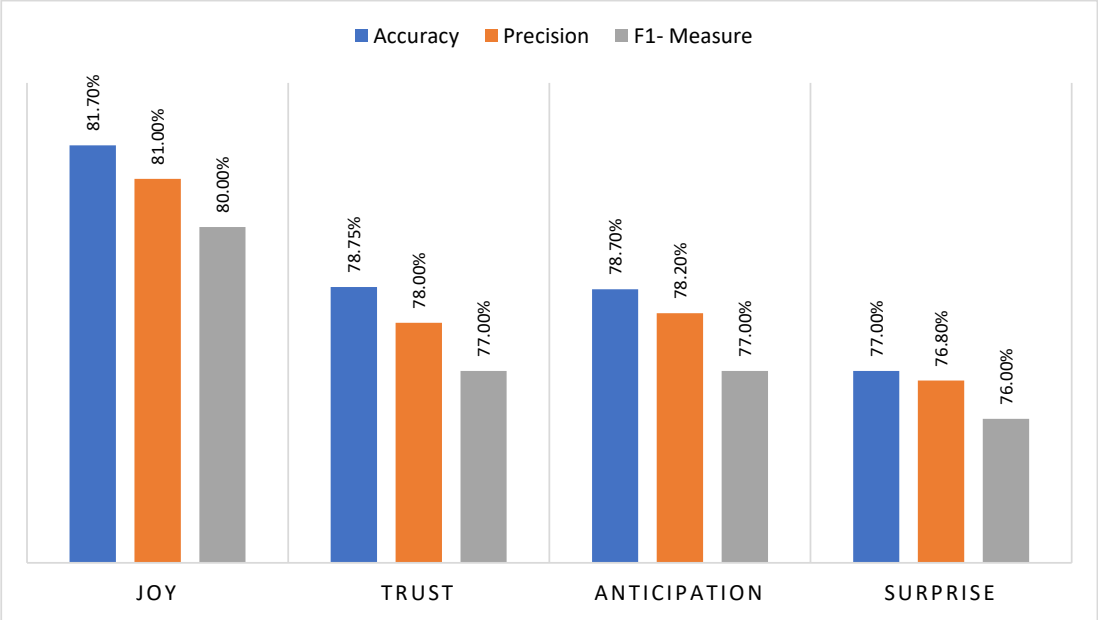

FIGURE 4.7: Discrete Positive Emotions Features Performance Using Training Data

Figure 4.8 illustrates the results of all Discrete Positive Emotions features set (Anticipation, Joy, Surprise, and Trust) for Validation data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Joy outperformed among all of them with 55.60% accuracy, 54.00% precision, and 53.50% f1-measure scores. Surprise showed the second-best performance with 55.30% accuracy, 53.00% precision, and 52.00% f1-measure scores. Trust showed the third-best performance with 55.00% accuracy, 53.50% precision, and 53.00% f1-measure scores. The remaining Anticipation performance is lower than Joy, Trust, and Surprise, with 53.59% accuracy, 52.00% precision, and 51.00% f1-measure scores.

Figure 4.9 illustrates the results of all Discrete Negative Emotions features set (Sad, Anger, Anxiety, and Disgust) for Training data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Sad outperformed among all of them with 77.00% accuracy, 76.00% precision, and 75.50% f1-measure scores. Anger showed the second-best performance with 75.15% accuracy, 75.00% precision, and 74.00% f1-measure scores. Anxiety showed the third-best performance with 75.00% accuracy, 74.00% precision, and 74.00% f1-measure scores. The remaining Disgust

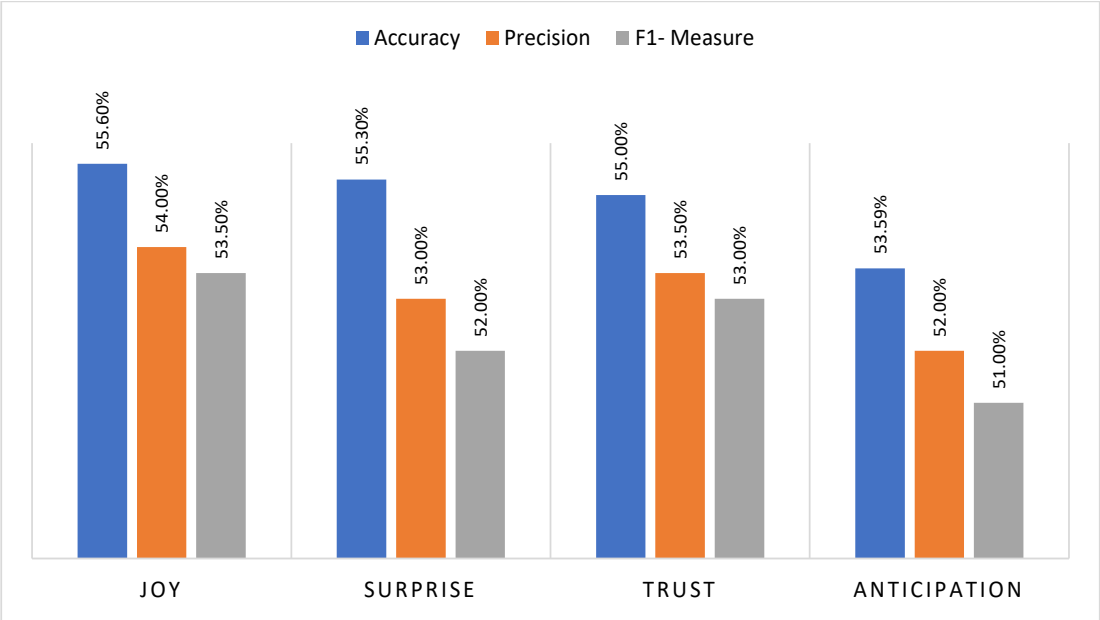

FIGURE 4.8: Discrete Positive Emotions Features Performance Using Validation Data

performance is lower than Sad, Anger, and Anxiety, with 74.90% accuracy, 74.50% precision, and 74.00% f1-measure scores.

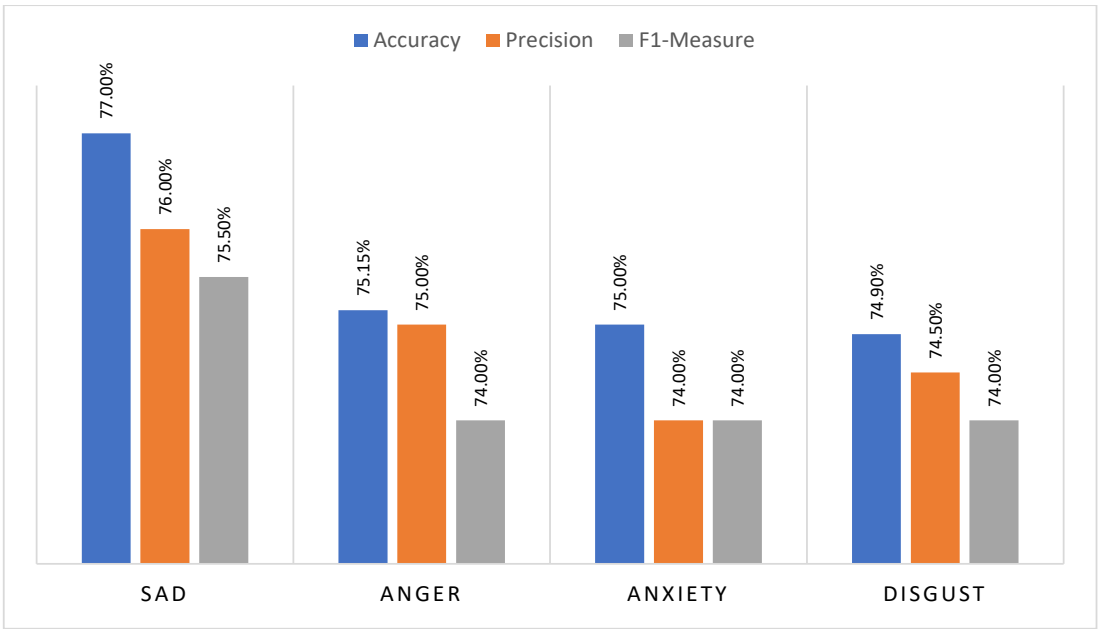

FIGURE 4.9: Discrete Negative Emotions Features Performance Using Training Data

Figure 4.10 illustrates the results of all Discrete Negative Emotions features set

(Sad, Anger, Anxiety, and Disgust) for Validation data with respect to the accuracy, precision, and f1-measure using Logistic Regression. It is evident from the following graphical representation that Disgust outperformed among all of them with 63.80% accuracy, 62.00% precision, and 59.00% f1-measure scores. Anxiety showed the second-best performance with 60.86% accuracy, 60.00% precision, and 59.90% f1-measure scores. Anger showed the third-best performance with 59.18% accuracy, 59.00% precision, and 56.00% f1-measure scores. The remaining Sad performance is lower than Disgust, Anxiety, and Anger with 55.00% accuracy, 54.00% precision, and 53.80% f1-measure scores.

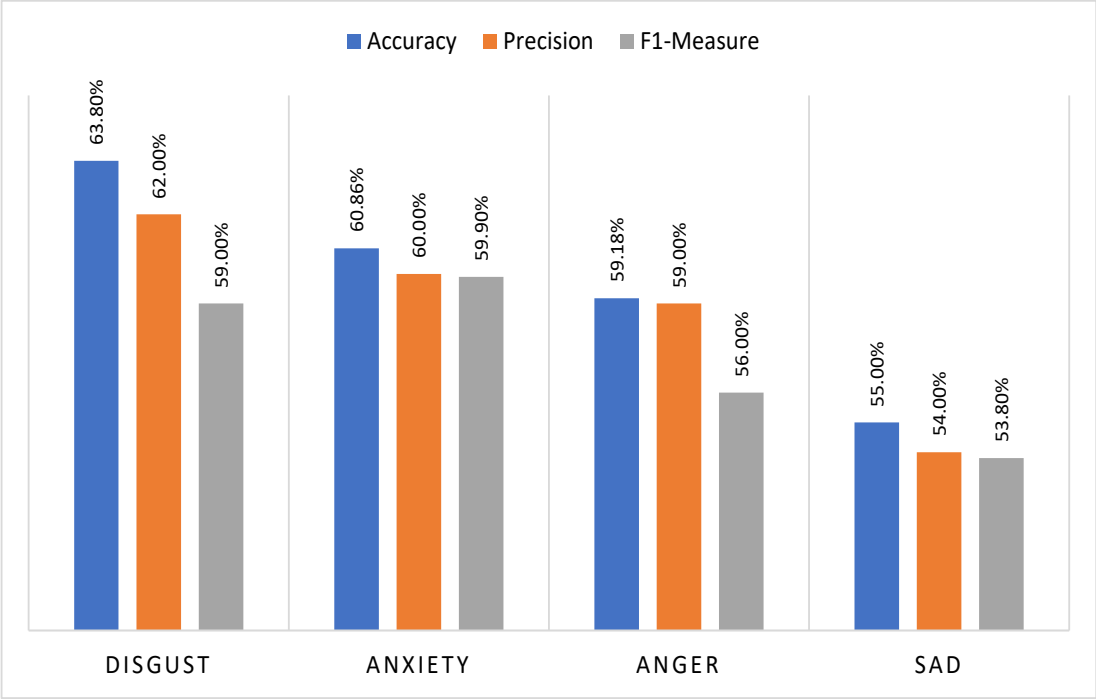

FIGURE 4.10: Discrete Negative Emotions Features Performance Using Validation Data

4.5.2.5 POS Features Performance

Figure 4.11 illustrates the result of Linguistic feature sets (Part-of-speech tagging) for Training data with respect to the Accuracy using Logistic Regression. It is evident from the following graphical representation that adjective (JJ) outperformed among all of them with 80.00% accuracy, verb base form (VB) showed the second-best performance with 77.00% accuracy, and Possessive pronoun (PRP$)

showed the third-best performance with 75.90% accuracy score. The remaining POS tags are lower performance than JJ, VB, and PRP$).

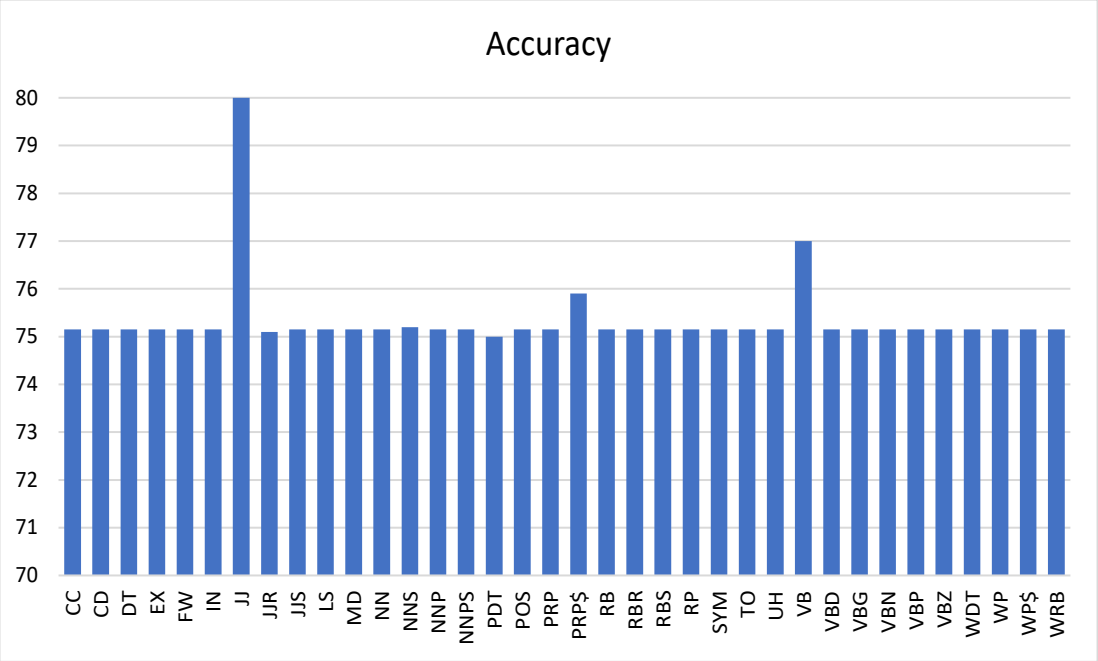

FIGURE 4.11: Standalone POS Features Performance Using Accuracy for Training Data

Figure 4.12 illustrates the result of Linguistic feature sets (Part-of-speech tagging) for Training data with respect to the Precision using Logistic Regression. It is evident from the following graphical representation that adjective (JJ) outperformed among all of them with 78.00% precision, verb base form (VB) showed the second-best performance with 76.20% Precision, and verb past participle (VBN) showed the third-best performance with 76.10% precision score.

Figure 4.13 illustrates the result of Linguistic feature sets (Part-of-speech tagging) for Validation data with respect to the Accuracy using Logistic Regression. It is evident from the following graphical representation that verb base form (VB) outperformed among all of them with 70.00% accuracy, adjective (JJ) showed the second-best performance with 69.40%.accuracy, and modal (MD) showed the third-best performance with 64.15% accuracy score. The remaining POS tags are lower performance than VB, JJ, and MD).

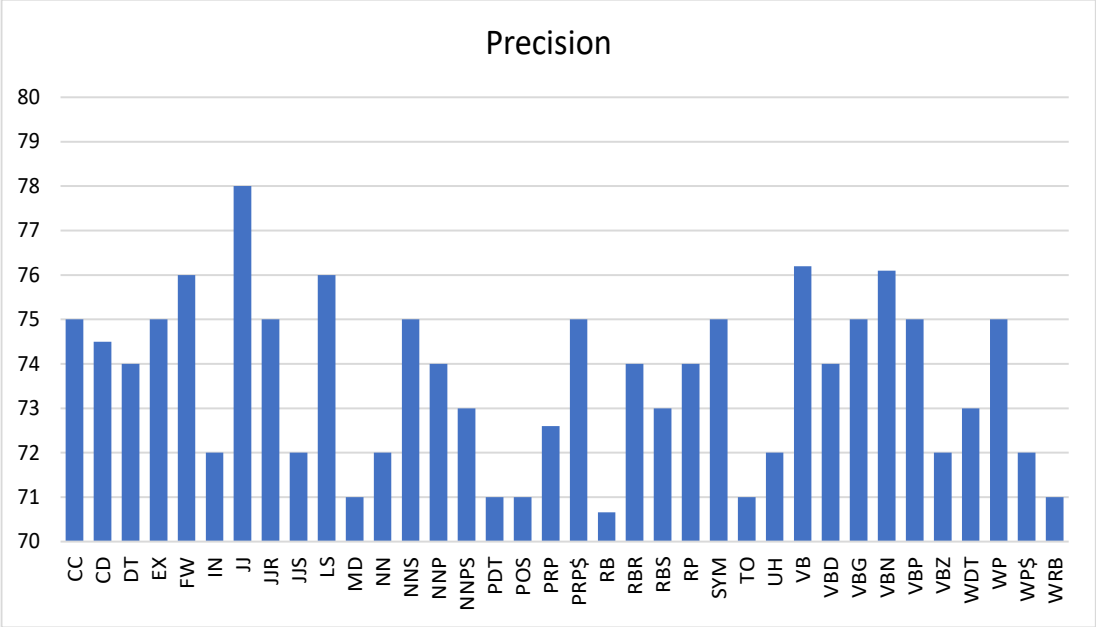

FIGURE 4.12: Standalone POS Features Performance Using Precision for Training Data

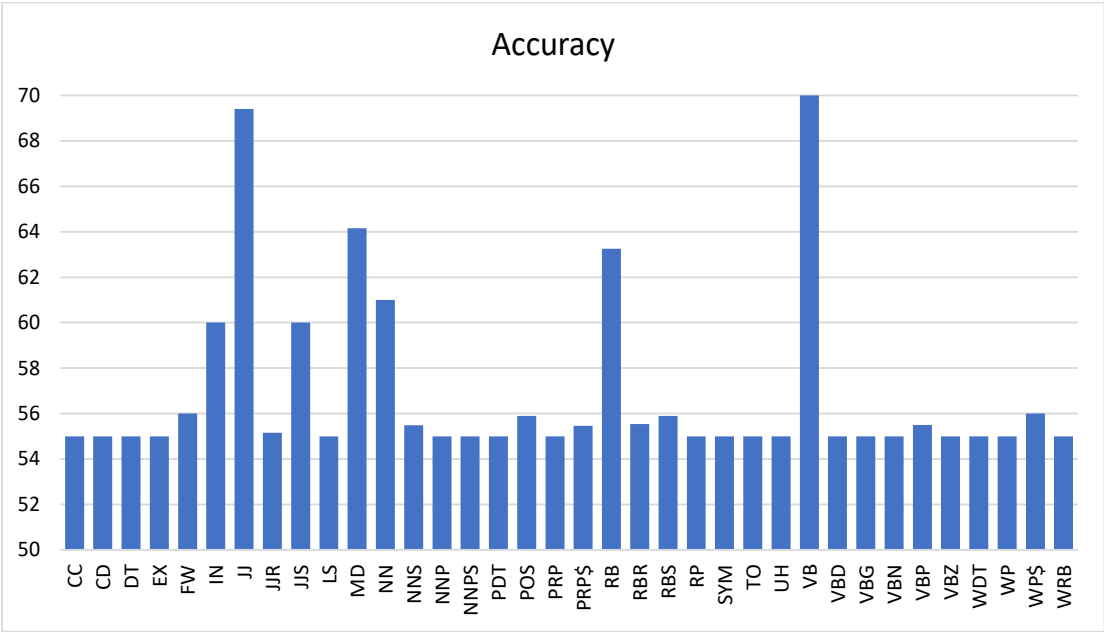

FIGURE 4.13: Standalone POS Features Performance Using Accuracy for Validation Data

Figure 4.14 illustrates the result of Linguistic feature sets (Part-of-speech tagging) for Validation data with respect to the Precision using Logistic Regression. It is evident from the following graphical representation that verb base form (VB) outperformed among all of them with 70.00% precision, adjective (JJ) showed the second-best performance with 69.00% precision, and adjective superlative (JJS)

showed the third-best performance with 66.10% precision score. The remaining POS tags are lower performance than JJ, VB, and VBN.

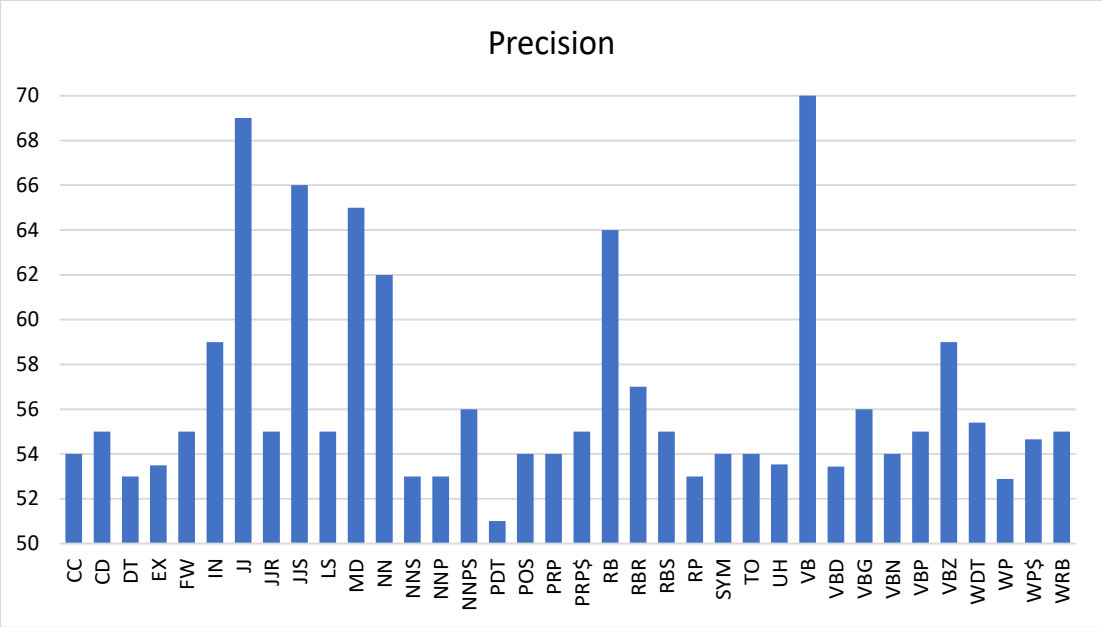

FIGURE 4.14: Standalone POS Features Performance Using Precision for Validation Data

Chapter 5

Conclusion and Future Work

This section will provide the conclusion of our research work and limitations for the future work.

5.1 Conclusion

In this work, we implemented and assessed a novel industry-specific smoke word list for defects identification in the cell phone manufacturing industry to provide cell phone manufacturers with business intelligence to continuously develop their quality. Product defects have such a significant negative influence on a company's competitive edge. Product defects can be discovered quickly and efficiently, which can help firms in performance control and increase product competitiveness. We have crawled cell phone product reviews from Amazon.com. We subdivided the cell phones domain further into two major subcategories: Samsung and Apple iPhone. We randomly selected unique reviews from the obtained dataset of "iPhone" and "Samsung" as a training set. All records in the training set were individually tagged from the three annotators using the cell phone tagging protocol. Three lists of smoke-words have been generated, "a list of words that are substantially more prevalent in defects than in non-defects" [9] in the training group across all subcategories.

This study demonstrates that our proposed smoke words list (unigram, bigram, and trigram) outperformed as compared to Sentiment Analysis. Smoke word lists are more effective than sentiment analysis in defect identification. Furthermore, Smoke word lists are more effective for all three domain-specific subcategories. Our findings show that the Smoke-Trigram dictionary outperformed than other two smoke word dictionaries (Unigram and Bigram) in performance defects. The AFINN sentiment analysis approach finds more defects than the other two, ANEW and Harvard GI Negative Sentiment analysis techniques. We have generated three domain-specific lexicons using performance defects which are Software defect words lexicon, Hardware defect words lexicon, and Complaint-type words lexicon. Complaint-type words lexicon outperformed than other two domain-specific lexicons (Software defects and Hardware defects) in performance defects. We adopted machine learning techniques and rechecked our framework performance. Our framework showed promising performance. Every proposed machine learning model in this study is evaluated with 10-fold cross-validation. For evaluating results collected from the machine learning models, we utilized precision, recall, f-measure, and accuracy as the evaluation metrics. After organizing all our machine learning models' results, our study showed that logistic regression outperformed among all other classifiers with higher performance in defects discovery when supplied with smoke words, sentiment analysis, POS, domain-specific, contextual, and linguist features.

5.2 Future Work

This research can be further extended on multiple levels. There is very limited work in defects identification, especially in electronics. This research could be enhanced by using this smoke word list in different contexts or exploring other methods of creating a cross-category list. More research is needed to determine smoke and domain-specific words for other companies and product categories. Future work needs to look at how risk assessment techniques may be adapted to

provide the outliers (products) a more acceptable risk score with only one or two or three review sets existing in Amazon.com.

Bibliography

[1] IHS Markit. The internet of things: a movement, not a market. *Critical IoT Insights*, pages 1–9, 2017.

[2] Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

[3] Xinxin Li and Lorin M Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, 2008.

[4] Yung-Ming Li, Hsuan-Ming Chen, Jyh-Hwa Liou, and Lien-Fa Lin. Creating social intelligence for product portfolio design. *Decision Support Systems*, 66: 123–134, 2014.

[5] Mohammed Alkahtani, Alok Choudhary, Arijit De, and Jennifer Anne Harding. A decision support system based on ontology and data mining to improve design using warranty data. *Computers & Industrial Engineering*, 128:1027–1039, 2019.

[6] Xuan Zhang, Zhilei Qiao, Aman Ahuja, Weiguo Fan, Edward A Fox, and Chandan K Reddy. Discovering product defects and solutions from online user generated contents. In *The World Wide Web Conference*, pages 3441–3447, 2019.

[7] Shixi Liu, Cuiqing Jiang, Zhangxi Lin, Yong Ding, Rui Duan, and Zhicai Xu. Identifying effective influencers based on trust for electronic word-of-mouth marketing: A domain-aware approach. *Information sciences*, 306:34–52, 2015.

[8] Alan S Abrahams, Jian Jiao, Weiguo Fan, G Alan Wang, and Zhongju Zhang. What's buzzing in the blizzard of buzz? automotive component isolation in social media postings. *Decision Support Systems*, 55(4):871–882, 2013.

[9] Alan S Abrahams, Jian Jiao, G Alan Wang, and Weiguo Fan. Vehicle defect discovery from social media. *Decision Support Systems*, 54(1):87–97, 2012.

[10] David Z Adams, Richard Gruss, and Alan S Abrahams. Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International journal of medical informatics*, 100:108–120, 2017.

[11] Lu Zheng, Zhen He, and Shuguang He. A novel probabilistic graphic model to detect product defects from social media data. *Decision Support Systems*, 137:113369, 2020.

[12] Yao Liu, Cuiqing Jiang, and Huimin Zhao. Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. *Decision Support Systems*, 105:1–12, 2018.

[13] Xuan Zhang, Zhilei Qiao, Lijie Tang, Patrick Weiguo Fan, Edward A Fox, and Alan Gang Wang. Identifying product defects from user complaints: A probabilistic defect model. Technical report, Department of Computer Science, Virginia Polytechnic Institute & State ..., 2016.

[14] Darren Law, Richard Gruss, and Alan S Abrahams. Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67:84–94, 2017.

[15] David M Goldberg and Alan S Abrahams. A tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*, 105:52–65, 2018.

[16] Alan S Abrahams, Weiguo Fan, G Alan Wang, Zhongju Zhang, and Jian Jiao. An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6):975–990, 2015.

[17] Matt Winkler, Alan S Abrahams, Richard Gruss, and Johnathan P Ehsani. Toy safety surveillance from online reviews. *Decision support systems*, 90: 23–32, 2016.

[18] H Almagrabi, A Malibari, and J McNaught. A survey of quality prediction of product reviews. *International Journal of Advanced Computer Science and Applications*, 6(11):49–58, 2015.

[19] Jian Jin, Ping Ji, and Rui Gu. Identifying comparative customer requirements from product online reviews for competitor analysis. *Engineering Applications of Artificial Intelligence*, 49:61–73, 2016.

[20] Andrea Landherr, Bettina Friedl, and Julia Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385, 2010.

[21] Adrianna Kezar. Higher education change and social networks: A review of research. *The journal of higher education*, 85(1):91–125, 2014.

[22] Yao Liu, Cuiqing Jiang, and Huimin Zhao. Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*, 123:113079, 2019.

[23] Wu He, Harris Wu, Gongjun Yan, Vasudeva Akula, and Jiancheng Shen. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7):801–812, 2015.

[24] Jian Jin, Ping Ji, Ying Liu, and SC Johnson Lim. Translating online customer opinions into engineering characteristics in qfd: A probabilistic language analysis approach. *Engineering Applications of Artificial Intelligence*, 41:115–127, 2015.

[25] Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45 (4):1007–1016, 2008.

[26] Wai Ming Wang, Zhi Li, ZG Tian, JW Wang, and MN Cheng. Extracting and summarizing affective features and responses from online product descriptions and reviews: A kansei text mining approach. *Engineering Applications of Artificial Intelligence*, 73:149–162, 2018.

[27] Jing Yang, Shaobo Li, Zheng Wang, Hao Dong, Jun Wang, and Shihao Tang. Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. *Materials*, 13(24):5755, 2020.

[28] Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan. Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion*, 36:149–161, 2017.

[29] Sangjae Lee and Joon Yeon Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.

[30] Minhoe Hur, Pilsung Kang, and Sungzoon Cho. Box-office forecasting based on sentiments of movie reviews and independent subspace method. *Information Sciences*, 372:608–624, 2016.

[31] Jingfei Du, Hua Xu, and Xiaoqiu Huang. Box office prediction based on microblog. *Expert Systems with Applications*, 41(4):1680–1689, 2014.

[32] Xiao Fang and Paul J Hu. Top persuader prediction for social networks. *MIS Quarterly, Forthcoming*, 2016.

[33] Kenneth D Kuhn. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122, 2018.

[34] Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni. A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580–596, 2018.

[35] Nikolaos Korfiatis, Panagiotis Stamolampros, Panos Kourouthanassis, and Vasileios Sagiadinos. Measuring service quality from unstructured data: A

topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, 116:472–486, 2019.

[36] Xun Xu and Yibai Li. The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International journal of hospitality management*, 55:57–69, 2016.

[37] Xun Xu, Xuequn Wang, Yibai Li, and Mohammad Haghighi. Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of information management*, 37(6):673–683, 2017.

[38] Yabing Zhao, Xun Xu, and Mingshu Wang. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76:111–121, 2019.

[39] Nan Hu, Ling Liu, and Jie Jennifer Zhang. Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and management*, 9(3):201–214, 2008.

[40] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 334–342, 2007.

[41] Xuan Zhang, Shuo Niu, Da Zhang, G Alan Wang, and Weiguo Fan. Predicting vehicle recalls with user-generated contents: A text mining approach. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 41–50. Springer, 2015.

[42] Qiao Liang and Kaibo Wang. Monitoring of user-generated reviews via a sequential reverse joint sentiment-topic model. *Quality and Reliability Engineering International*, 35(4):1180–1199, 2019.

[43] Richard Gruss, Alan S Abrahams, Weiguo Fan, and G Alan Wang. By the numbers: The magic of numerical intelligence in text analytic systems. *Decision Support Systems*, 113:86–98, 2018.

[44] Ying Liu, Jian Jin, Ping Ji, Jenny A Harding, and Richard YK Fung. Identifying helpful online reviews: a product designer's perspective. *Computer-Aided Design*, 45(2):180–194, 2013.

[45] Srikumar Krishnamoorthy. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759, 2015.

[46] Xiaolin Zheng, Shuai Zhu, and Zhangxi Lin. Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach. *Decision Support Systems*, 56:211–222, 2013.

[47] Xun Xu. Examining the role of emotion in online consumer reviews of various attributes in the surprise box shopping model. *Decision Support Systems*, 136: 113344, 2020.

[48] Tse-Hsun Chen, Weiyi Shang, Meiyappan Nagappan, Ahmed E Hassan, and Stephen W Thomas. Topic-based software defect explanation. *Journal of Systems and Software*, 129:79–106, 2017.

[49] Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi. Real-time traffic incident detection using probe-car data on the tokyo metropolitan expressway. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 43–45. IEEE, 2014.

[50] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[51] Weiguo Fan, Michael D Gordon, and Praveen Pathak. Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems*, 40(2):213–233, 2005.

[52] W Medhat, A Hassan, and H Korashy. Sentiment analysis algorithms and applications: a survey. ain shams eng. j. 5 (4), 1093–1113 (2014).

[53] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news, 2013.

[54] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ..., 1999.

[55] Edward F Kelly and Philip James Stone. *Computer recognition of English word senses*, volume 13. North-Holland, 1975.

[56] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[57] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[58] Michael A Cohn and Barbara L Fredrickson. Positive emotions. *Oxford handbook of positive psychology*, 2:13–24, 2009.

[59] Bonnie M Le and Emily A Impett. When holding back helps: Suppressing negative emotions during sacrifice feels authentic and is beneficial for highly interdependent people. *Psychological Science*, 24(9):1809–1815, 2013.

[60] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[61] Bao Pham. Parts of speech tagging: Rule-based. 2020.

[62] Matthew K Nock and Mitchell J Prinstein. Contextual features and behavioral functions of self-mutilation among adolescents. *Journal of abnormal psychology*, 114(1):140, 2005.

[63] MSZ Rizvi. Demystifying bert: A comprehensive guide to the groundbreaking nlp framework. *Link: https://www. analyticsvidhya. com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework*, 2019.

[64] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[65] Long Ma and Yanqing Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE, 2015.

[66] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[67] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[68] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[69] Ayon Dey. Machine learning algorithms : A review. 2016.

[70] Issam El Naqa and Martin J Murphy. What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer, 2015.

[71] Yanru Zhang and Ali Haghani. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58: 308–324, 2015.

[72] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. *Advances in neural information processing systems*, 27:253–261, 2014.

[73] Andrew P Worth and Mark TD Cronin. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM*, 622(1-2):97–111, 2003.

[74] Shuchuan Lo. Web service quality control based on text mining using support vector machine. *Expert Systems with Applications*, 34(1):603–610, 2008.

[75] Wen Mau Chong, Chien Le Goh, Yoon Teck Bau, and Kian Chin Lee. Fast numerical threshold search algorithm for c4. 5. In *2014 IIAI 3rd International Conference on Advanced Applied Informatics*, pages 930–935. IEEE, 2014.

[76] Monalisa Mohanty, Santanu Sahoo, Pradyut Biswal, and Sukanta Sabut. Efficient classification of ventricular arrhythmias using feature selection and c4. 5 classifier. *Biomedical Signal Processing and Control*, 44:200–208, 2018.

[77] Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–73, 1997.

[78] Marco D'Ambros, Alberto Bacchelli, and Michele Lanza. On the impact of design flaws on software defects. In *2010 10th International Conference on Quality Software*, pages 23–31. IEEE, 2010.

[79] Automata based test plans for fault diagnosis in batch processes. volume 37 of *Computer Aided Chemical Engineering*, pages 1781–1786. Elsevier, 2015.

[80] Maribel Lopez. Samsung explains note 7 battery explosions, and turns crisis into opportunity. *Retrieved April*, 7:254, 2017.

Appendix A

TAGGING PROTOCOL – CELL PHONES

This tagging protocol document contains all tagging information. It was provided to the human taggers, and it contains all of the necessary information, including tag type attributes and definitions, as well as examples of how to tag a team.

Tagging Types:

Defect Severity: Tag type; describes the type of defect.

⇒ *For the defect type column, tag as follows:*

- *No Defect* – If the user doesn't mention any problems or just offers positive feedback with good comments.
- *Performance Defect* – The product doesn't really behave as the manufacturer expected or as the customer desires.

Software Defect: Tag type; describes the software defects that occurred in the cell phone (labeled as defective by the users).

⇒ *For the Software Defects column, tag as follows:*

- *No Defect* – If the user doesn't mention any problems or just offers positive feedback
- *App Frozen and Slow User Interface* – It is an unexpected event wherein the operating system, or cell phone's applications don't work correctly.
- *Connectivity Issues* – If the users experience a issue connecting to Bluetooth, Wi-Fi, or cellular network.
- *App Crashes* – If the user experiences with an unexpected termination of a process.
- *Upgrade Failure* – If users face an experience with a cell phone system, a software update failed.
- *Apps Not Downloading* – User experiences cell phone applications not downloading due to some reasons.
- *Other Software Defect* – Defects not listed above
- *Multiple Software Defects* – The user is concerned about the many software defects outlined above.

Hardware Defect: Tag type; describes the hardware defects that occurred in the cell phone(labeled as defective by the users).

⇒ *For the Hardware Defects column, tag as follows:*

- *No Defect* – If the user doesn't mention any problems or just offers positive feedback with good comments.
- *Bad Battery Life Defect* – Cell phone's battery is draining too fast and doesn't last long.
- *Overheating Cell Phone Defect* – Cell phone gets hot and doesn't work correctly.
- *Micro-SD card does not work Defect* – SD card does not work properly and does not show up even testing SD card on another Cell phone.

- *Facial Recognition Reliability Issues Defect* – Face ID is not working, or the process of identifying a person's identity using their face is prolonged.
- *Fingerprint Scanner Reliability Issues Defect* – Cell phone's Fingerprint scanner doesn't work. Lousy accuracy or low recognition.
- *Poor Processing Speed Defect* – Cell phone runs very slowly/lags or takes a long response time.
- *Wi-Fi, NFC, and Bluetooth Not Working Defect* – Cell phone's Wi-Fi, NFC, and Bluetooth do not function properly due to faulty hardware components of wireless
- *Cell Phone Signal Antenna Low Coverage Area Defect* – Network Coverage does not function properly due to defective signal antenna. Weak signals, dropped calls automatically, and slow data speeds.
- *Power, Volume, and Home buttons Not working Defect* – Power, volume, and home buttons Not working due to faulty board.
- *Speaker or Microphone Obstruction Defect* – Cell phone's speaker or mic is not working properly.
- *Dead Pixel and Flickering Screen Defect* – Dead pixel is the stuck point or several matrix screen points that do not correctly reflect the color and pixels and the screen continually flickering.
- *Charging Port or Headphone Jack Defect* – Cell phone's Charging Port or Headphone jack is not working due to hardware defect.
- *Front or Rear Camera Not Working Defect* – Cell phone's Front or Rear Camera is not working due to low-quality hardware.
- *Touch Screen Less Sensitivity Defect* – Cell phone's touchscreen isn't responsive or not working due to less touch sensitivity.
- *Flashlight Not Working Defect* – Cell phone's Flashlight or torch is not functioning correctly for several reasons.
- *Waterproof Resistance Defect* – Cell phone's waterproof resistance is not worked properly due to low-quality resistance.

- *Other Hardware Defect* – Defects not listed above.
- *Multiple Hardware Defects* – – if the user is concerned about several of the hardware mentioned above defects.

Complaint Type: Tag type; User issues/complaints about damaged or missing parts upon arrival are described in this tag category.

⇒ *This column is for the Complaint types, tag as follows:*

- *No Complaint* – If there are no complaints from the user.
- *Scratches and Dent on the Body Complaints* – User complains that the cell phone was delivered with scratches and dented on the body.
- *Cracked Cell Phone Screen Complaint* – User complains that the cell phone was delivered with a cracked LCD.
- *Locked Cell Phone Delivery Complaint* – User complains that the locked cell phone delivered even they have bought fully unlocked cell phone.
- *Faulty and Used Cell Phone Delivery Complaint* – User complains that the defective and used cell phone delivered even they have bought a brand-new cell phone.
- *Cell Phone Carrier Locked Complaint* – User complains that the carrier locked cell phone delivered, a user can't switch or change to another carrier without "unlocking" the cell phone first.
- *IMEI invalid complaint* – User complains that a cell phone delivered with an invalid IMEI number.
- *Durability Compliant* – The cell phone has ceased working, according to the user. The cell phone has only operated for a few [Days, Months, Years], and its performance and functioning are inconsistent
- *Cost Complaint* – The cost of a mobile phone is high, and replacement and repair are prohibitively expensive, according to the user.

- *Design Complaint* – The user claims that the cell phone falls short of expectations owing to a design fault.
- *Accessories/Parts Missing or Broken Complaint* – User complains that the cell phone delivered with broken or missing accessories/parts.
- *Blacklisted Cell Phone Delivery Complaint* – User Complains that the blacklisted or stolen cell phone delivered.
- *Black Blemish/Blotch on the Screen Complaint* – User complains that a cell phone delivered with the black blemish/Blotch on the LCD.
- *Other Complaint* – If complaints that were not addressed above.
- *Multiple Complaints* – If the user has numerous issues and concerns from the preceding list.

Important Notes:

- Use the comments if any of the given tags aren't appropriate for this tagging protocol.
- Use the comments if additional information is required.
- Please, be careful when labeling reviews like "I am really disappointed by this mobile phone because its screen is not large enough to watch movies on this cell phone." This isn't a screen defect. It would be considered a designed complaint.