ABSTRACT

STATISTICS-BASED ANOMALY DETECTION AND CORRECTION METHOD FOR AMAZON CUSTOMER REVIEWS

by
Ishani Chatterjee

People nowadays use the Internet to project their assessments, impressions, ideas, and observations about various subjects or products on numerous social networking sites. These sites serve as a great source of gathering information for data analytics, sentiment analysis, natural language processing, etc. The most critical challenge is interpreting this data and capturing the sentiment behind these expressions. Sentiment analysis is analyzing, processing, concluding, and inferencing subjective texts with the views. Companies use sentiment analysis to understand public opinions, perform market research, analyze brand reputation, recognize customer experiences, and study social media influence. According to the different needs for aspect granularity, it can be divided into document, sentence, and aspect-based sentiment analysis.

Conventionally, the true sentiment of a customer review matches its corresponding star rating. There are exceptions when the star rating of a review is opposite to its true nature. These are labeled as the outliers in a dataset for this work. The state-of-the-art methods for anomaly detection involve manual search, predefined rules, or machine learning techniques to detect such instances. This dissertation work proposes a statistics-based anomaly detection and correction method (SADCM), which helps identify such reviews and rectify their star ratings to enhance the performance of a sentiment analysis algorithm without any data loss. This data analysis pipeline preserves these outliers to correct them and prevents any information loss.

This research work focuses on performing SADCM in datasets containing customer reviews of various products, which are a) scraped from Amazon.com and b) publicly available. The scraped dataset includes 35,000 Amazon customer reviews while the publicly available dataset includes 100,000 Amazon customer reviews for multiple products reviewed this year. The research work also analyzes these datasets and concludes the effect of SADCM on the performances of several sentiment analysis algorithms. The results exhibit that SADCM outperforms other state-of-the-art anomaly detection algorithms with a higher accuracy and recall percentage for all the datasets. The proposed method should thus help businesses that rely on public reviews to enhance their performances in better decision-making.

STATISTICS-BASED ANOMALY DETECTION AND CORRECTION METHOD FOR AMAZON CUSTOMER REVIEWS

by
Ishani Chatterjee

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Engineering

Helen and John C. Hartmann Department of
Electrical and Computer Engineering

December 2021

APPROVAL PAGE

STATISTICS-BASED ANOMALY DETECTION AND CORRECTION METHOD FOR AMAZON CUSTOMER REVIEWS

Ishani Chatterjee

Dr. MengChu Zhou, Dissertation Advisor — Date
Distinguished Professor of Electrical and Computer Engineering, NJIT

Dr. Nirwan Ansari, Committee Member — Date
Distinguished Professor of Electrical and Computer Engineering, NJIT

Dr. Hieu Nguyen, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, NJIT

Dr. Qing Liu, Committee Member — Date
Assistant Professor of Electrical and Computer Engineering, NJIT

Dr. Zhipeng Yan, Committee Member — Date
Professor of Finance, Associate Dean of MT School of Management, NJIT

BIOGRAPHICAL SKETCH

Author: Ishani Chatterjee

Degree: Doctor of Philosophy

Date: December 2021

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2021

- Master of Science in Computer Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2017

- Bachelor of Science in Electronics and Communication Engineering,
  West Bengal University of Technology, Kolkata, West Bengal, India, 2015

Major: Computer Engineering

Publications:

I. Chatterjee, M. Zhou, A. Abusorrah, K. Sedraoui, and A. Alabdulwahab, "Statistics-Based Outlier Detection and Correction Method for Amazon Customer Reviews," Entropy, vol. 23, 1645. https://doi.org/10.3390/e23121645 (24 pages), Dec. 2021.

H. Liu, I. Chatterjee, M. C. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 6, pp. 1358–1375, 2020.

I. Chatterjee, "Multi-population-based differential evolution algorithm for optimization problems," *Thesis*, Newark College of Engineering, New Jersey Institute of Technology, USA, 2017. [Online].
Available: https://digitalcommons.njit.edu/theses/17

I. Chatterjee and M. C. Zhou "Differential evolution algorithms under multi-population strategy*", 26th Wireless and Optical Communications Conference (WOCC),* Newark, USA, April 2017

I. Chatterjee and A. Mukhopadhyay "Design of a microstrip line fed rectangular slot antenna and response under slit loading," *4th International conference of Technical and Managerial innovation in computing and communication in Industry and academia*, Kolkata, India, August 2013

Presentations:

I. Chatterjee and M. C. Zhou “A review of electric vehicles, charging stations and their impact on consumers and manufacturers”, *16th IEEE International Conference on Networking, Sensing and Control (ICNSC)*, Banff, Canada, May 2019

This dissertation is dedicated to
Dadu and Daidai for inspiring me,
Mom and Dad for believing in me.

ACKNOWLEDGMENT

For this dissertation research I have received a great deal of support and assistance throughout the process. First and foremost, I am thankful to my advisor Dr. MengChu Zhou, Distinguished Professor, for his valuable guidance and assistance, without which the accomplishment of this research work would have never been possible. Dr. Zhou has directed me with great enthusiasm and interest and has allowed me to have the freedom to exercise a thoughtful and scientific approach towards the problem.

I would also like to thank Professors Nirwan Ansari, Hieu Nguyen, Qing Liu, and Alan Zhipeng Yan for agreeing to join the committee for my dissertation defense. I would also like to thank all the professors who have guided me throughout these years, particularly Dr. Durgamadhab Misra, for his patient support and encouragement.

I have been given a unique opportunity as Teaching Assistant by the Department of Electrical and Computer Engineering. Through this position, I was able to touch the lives of so many students and motivate and guide them in both professional and personal ways. I would also like to thank Ms. Montera Bass for giving me these duties and helping and supporting me throughout my doctoral study at New Jersey Institute of Technology that ranks among the Top 50 National Public University by U.S. News & World Report and #1 Nationally in Student Economic Upward Mobility by Forbes.

I would like to acknowledge my colleagues for their wonderful company. I would particularly like to thank Kaustav, Deepan, Craig, Sean and Fatemeh for their continuous encouragement and assistance with anything and everything. I would like to thank my parents for their wise counsel and sympathetic ear and Dadu and Daidai, who are always there with me. I would like to thank Arpit Bhaiya and Sanchi Bhabhi for their constant support, which helped me stay focussed and grounded. Finally, I could not have completed this dissertation without the support of my friends, Smarth, Aman, and Amol, who provided stimulating discussions and happy distractions to rest my mind outside of my research.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Q_1	First quartile
Q_2	Median
Q_3	Third quartile
F_L	Lower fence
$\boldsymbol{F_U}$	Upper fence
I_{QR}	Interquartile range
O_P	Potential outlier
O_D	Definite outlier
r_i	Customer review comment
r_i^*	Star rating of review, r_i
S^+	Positive star-rated review
S^-	Negative star-rated review
T_V	Target value
C_V	Compound sentiment score
V_D	Value difference

LIST OF DEFINITIONS

Interquartile range	The difference between the first and third quartile.
First quartile	The intermediate point of the first half of an ordered dataset.
Median	The value which separates the higher half from the lower half of an ordered dataset.
Third quartile	The intermediate point of the second half of an ordered dataset.
Lower fence	The cut-off values for lower outliers.
Upper fence	The cut-off values for upper outliers.
Box plot	A graph form to display quantitative data.
Accuracy	The measurement of closeness to a specific target value.
Recall	The capacity of a model to find all the applicable cases within a dataset.
p-value	The measure of the likelihood that a noticeable difference might have occurred by random chance.
T-score	The value that represents the number of standard deviations.
CI	The range of plausible values of an unknown parameter.

CHAPTER 1

INTRODUCTION

Disney's primary revenue sources are theme parks, cruise lines, Broadway shows, and movies. With the lockdowns during the pandemic, all were closed, and their stocks fell drastically. Through market research and customer comments, it was noted that the popularity of streaming services was increasing drastically. Disney launched Disney+ with the release of Frozen 2 movie in the streaming service. It immediately caught the attention and buzz amongst consumers. The company then realized a series of movies and recorded Broadway shows, and within the first five months, they had more than 50 million subscribers and their stocks were at an all-time high. This study shows how vital customer opinion is to a company and its growth.

Sentiment analysis is identifying the attitude or emotion towards a specific product or topic in a text, sentence, or for that matter, a comment. It can be performed on product reviews or surveys by an individual or an organization on a specific topic, issue, or event online or in social media in the form of video, graphics, or images also. Sentiment analysis plays an essential role in marketing. It can be used to fulfill several tasks. A few of such examples where sentiment analysis has helped the industry grow is discussed as follows:

1: **Marketing research**: Coca-Cola performed thorough market research to figure out that different cultures have different preferences of sugar levels. Hence it is sweeter in some countries than the others. For example, coke in some Asian countries uses sugar while countries like the USA use corn syrup. There are also region-specific Coca-Cola flavors to cater to the customer's needs.

2: **Tracking product success**: Tracking product success helps understand the response the product receives after its launch. For instance, when Tropicana changed its packaging, the sales decreased, and through customer reviews, it was proved that people could not connect to the product. Once they changed the packaging back to the original label, the sales increased automatically.

3: **Customer tracking**: Amazon follows data-driven marketing. It looks at products a consumer has been browsing and recommends similar products of different shapes, sizes, and brands. Under labels like frequently bought together, best-selling, related to items you've viewed, or recommended for you.

4: **Improving customer support**: It is crucial every time a customer mentions a brand name. Because each time they mention the brand, it gives the company a chance to glimpse their sentiment towards the brand and the products. To introduce new flavors, Gatorade actively tracks the overall customer satisfaction, participates in social media, engages with customers, and identifies critical emotional triggers to understand consumer sentiment and demand.

1.1 Sentiment Analysis

Sentiment analysis, emotion artificial intelligence, intent analysis is often used to describe the same concept, i.e., opinion mining. Sentiment analysis uses a combination of Natural

Language Processing (NLP), computational linguistics, and text mining to analyze, derive, calibrate, and evaluate textual information in the form of sentences, phrases, documents, etc. (Wikipedia, 2020). NLP has earned a lot of attention in recent times. People have started to rely on consumer reviews and sentiments shared over social media sites, blogs, and consumer feedback websites on the Internet before purchasing or opting for a particular product or service. It has also become a vital tool for decision-makers who plan to improve, modify, or perform necessary actions based on public opinions.

Sentiment analysis has become a rapidly growing research area since (Pang & Lee, 2004) created a comprehensive study to determine the sentiment polarity of movie reviews. It has received attention from academia and the industry because it can provide feedback information of customers through online reviews, help decide marketing policies, and detect changes in customers' opinions. Its primary task is to classify the expressed opinion of a given text into various classes, such as positive, negative, and neutral.

Sentiment analysis is used extensively in various domains like marketing, politics, sports, and stocks for information extraction, improvement of an automated chatbot response system, or product modification. Most companies use sentiment analysis to research consumer requirements and understand the market trends. Positive reviews of a product or service drive online marketing, while negative comments motivate companies to improve their products or services based on customer demands. Social media have become a robust platform that helps understand public opinions, acceptance, or issues regarding specific laws or lawmakers. Sentiment analysis helps one study the endorsement rate of these policies based on previous trends, which allows lawmakers to prepare and motivate the public accordingly. Similarly, this method aids in fan engagements and player/team reputation build-up in sports. It also helps one study a company's prominence in the market, which impacts its stock valuation. These are some of the applications of sentiment analysis, to name a few.

In (R. Wang et al., 2019), the author explains that a sentence that holds an opinion consists of quintuple parameters (*e*, *a*, *s*, *h*, *t*) where *e* is the target or entity, *a* is the aspect or feature of *e*, *s* is the nature of the opinion on *e* or *a*, *h* is the opinion holder, and *t* is the time when *h* expresses the sentiment. For instance, in this 5-star Amazon review for a hand sanitizer, "*With having to use hand sanitizers so much due to the Covid situation, this is the best one I have found. Love the residual effects and the fact that is doesn't dry out my skin. Would recommend over other brands*." *e* is the hand sanitizer, *a* is the residual effect, the nature of the opinion is positive, and the opinion holder is the Amazon reviewer while time is during COVID-19 pandemic. Sentiment analysis focuses explicitly on *s*, which is the nature of the opinion. The four most popular types of sentiment analysis are explained as follows:

1: **Fine-grained sentiment analysis**: In most situations, sentiment analysis usually classifies the nature of text into positive or negative. On the contrary, fine-grained sentiment analysis categorizes the given text into very negative, negative, neutral, positive, and very positive. This usually helps in identifying 5-star ratings in a review such that very negative can be represented by 1-star and very positive as 5-star. For instance, this Amazon 5-star apparel review, "*My son loved this jacket. It is comfortable and light weight. Great looking. Washes well.*" is a very positive comment. This Amazon 1-star TV review is a perfect example of a customer review carrying a negative opinion, "*Waste of money. Wireless stopped working after a week.*"

2: **Emotion detection**: Emotion detection detects emotions like happiness, anger, sadness, and frustration in a text or phrase by implementing complex machine learning algorithms or lexicon approaches. The latter can be challenging as some words, like bad or kill, can reflect anger (your product is so *bad*, or these shoes are *killing* me) and express happiness in a sentence (this is a *bad ass* product, or their styles are *killing* it). For instance, this Amazon 1-star perfume review "*Fake, fake, fake. This is a complete knockoff! It smells like alcohol and soap. So upsetting it smells nothing like the original. The bottle is supposed to be pink, but it's clear, and the disgusting perfume is rose-tinted to fool you. Such a rip-off!*" holds an 'angry' emotion while this Amazon 5-star perfume review "*I have been almost exclusively wearing this perfume for 13 years, and I am 28 now. It is great for everyday wear, but also for special occasions. The smell is not abrasive or overpowering, but soft and still noticeable. I get many compliments on Lucky You. I'm not sure why this one is so much cheaper than in stores, but I would defend that it is a complete knock off because the smell is the same and it lasts as long.*" holds a 'happy' emotion, although both the comments share similar words like *knock off*.

3: **Aspect-based sentiment analysis**: In a review or comment, customers usually express different emotions for different aspects of a product. Aspect-based sentiment analysis helps in identifying the diverse nature of opinion based on various factors of a product. For example, this Amazon 5-star review of furniture express assembling, packaging, sturdiness, and cost "*I really like these, a lot. They were easy to assemble. It came well packaged, and once out up, we're pretty sturdy. I'd say they're a great purchase for the price*."

4: **Multilingual sentiment analysis**: Multilingual sentiment analyses are usually challenging as they involve rigorous pre-processing and corpus. Most of the resources are available online, but some need to be created. This is tedious and requires coding to identify the different languages and then use the corresponding resources for sentiment analysis. These Amazon 1-star figurine reviews are a perfect example of multilingual sentiment analysis like "*Es demasiado pequeña y la luz apenas visible habían villas más bonitas que esta me arrepiento haber comprado esto en Amazon.*" Spanish, and "*Die Lieferung hat sich verzögert. Dann kam der Artikel nicht wie abgebildet mit schönen kräftigen gelben Blumen, sondern in einer matschigen und dreckig wirkenden Farbe. Den Artikel wollte ich ersetzen lassen. Anscheinend ist nur der Rückversand eine Option für mich. Es sollte ein Geschenk zu Weihnachten werden. Von diesem Kauf bin ich mehr als enttäuscht.*" German.

Aspect-based sentiment analysis aims to find the overall sentiment associated with an entity and the opinion for each aspect of the referred entity. When a popular product has many reviews, potential customers have difficulty reading all the reviews before buying a product. In this case, a fine-grained analysis becomes very important. At the same time, it can provide manufacturers with more detailed information to help them improve their products in a specific aspect. For example, "*Good for Covid-19 but mask material is uncomfortable*". This comment reflects that the customer has a positive sentiment for the 'utility' of the product but a very negative opinion for the 'quality' of product material. Authors in (M. Hu & Liu, 2004) believe that aspect-based sentiment analysis (ABSA) consists of two sub-tasks: 1) identifying aspects in which customers express their opinions; and 2) for each aspect, identifying sentences that give positive, negative, or neutral opinions.

With the expansion in data available through Internet, researchers have started focusing on both the academic and commercial applications of sentiment analysis. The boost in smartphone usage has increased the development of mobile games and apps. (Oyebode et al., 2020) use sentiment analysis to analyze the mental health apps in smartphones to classify their features as positive or negative. This analysis leads to some design modifications based upon the negative factors of the app, which helps the app increase its potency. (Afzaal et al., 2019) use aspect-based sentiment analysis to implement a smartphone tourism app to identify the most recommended restaurants and hotels in a city by extracting and classifying information from tourist reviews.

Online reviews play a vital role in the fashion industry as it helps designers understand a shopper's experience via the latter's feedback. (W. Li & Xu, 2020) propose an aspect-based fashion recommendation model with an attention mechanism. They use convolutional neural networks, long short-term memory networks, and attention mechanisms to process customer and product reviews simultaneously. They then combine them to apprehend both local and global reviews, which helps predict the customer rating.

There have been several investigations being conducted in this area. The survey paper (Harrag et al., 2019) discusses the improvements in predicting customer reviews and ratings. Many scholars have also compared different types of approaches for sentiment analysis and the evaluation of several algorithms to conclude the algorithm that best fits their respective datasets (Fan et al., 2021). Many statistical surveys and studies were conducted in this area (*The Role of Customer Product Reviews*, 2019) shows that 39% of future customers read around eight reviews while 12% read sixteen or more reviews. The customer reviews have become so vital that a study showed that if a buyer is confused between two products, the buyer will likely not choose the product with fewer or no reviews. 98% of buyers are resistant to buying a product with no reviews. At the same time, almost four out of five customers have changed their minds about buying a product recommended by their friends or family because of negative reviews (*20 Online Review Stats to Know in 2019*, 2019). Figure 1.1 shows the popularity percentage distribution of online retail stores in the USA.

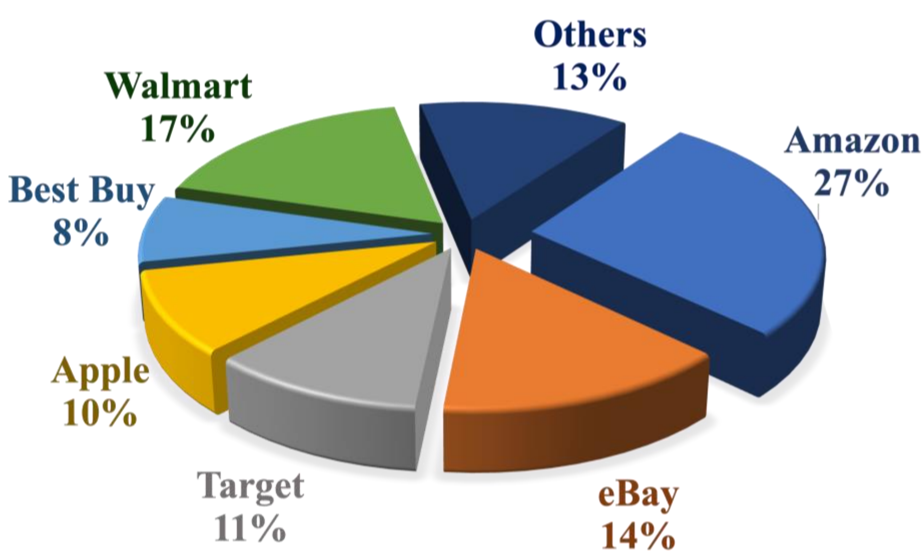


Figure 1.1 Popularity percentage distribution of online retail stores in the USA.

1.2 Objective of Dissertation

Outlier detection is a salient data analysis concern that focuses on identifying oddities in datasets. Outlier (a.k.a. anomaly, noise, and exception) detection helps recognize an entity that prominently differs from most samples in a dataset (Wikipedia, 2021). Such entities may represent bank frauds, spam emails, structural defects, and errors in a dataset. Anomaly detection faces many challenges due to a) the characteristic of input data or the nature of outliers, b) noise in a dataset that might mimic an outlier, c) inaccurate boundaries between standard data and outliers, and d) computational complexity. In (H. Wang et al., 2019), Wang et al. explain the importance of designing an efficient and scalable outlier detection algorithm because the probability of the number of outliers is directly proportional to the volume of a dataset. It is also critical to promptly identify and rectify the outliers in a dataset to have high-quality data.

The definition of an outlier may vary for various scenarios. For example, in this 5-star Amazon review for a hand sanitizer, "*Do not buy. Doesn't sanitize for ovid19. Does not contain alcohol. Fake description as a sanitizer*." the nature of the review is positive as opposed to the sentiment of the review comment. Much like the example, this paper defines novelty as the reviews with opinions opposite to their corresponding star ratings. Anomaly detection is an eminently researched topic in various domains (de la Torre-Abaitua et al., 2021). There is a preliminary study on outlier detection using sentiment analysis of a dataset. It is classified predominantly into supervised and unsupervised learning. The former is true when the dataset used is labeled, while the latter arises when the dataset is not marked. The techniques used to identify anomalies are based broadly on classification, clustering, distance, machine learning, and statistical approaches. Generally, the nature of a review or comment is related to the corresponding rating, but that's not true in some cases. Those are the exceptions or outliers in the dataset. Detecting and rectifying these outliers without information or knowledge loss is a challenge that researchers are studying now.

1.3 Organization of Dissertation

This dissertation study proposes an outlier detection method named SADCM (Statistics-based Anomaly Detection and Correction Method) using a combination of statistical and distance-based techniques. Our concerned dataset can be broadly classified into two types a) scraped from the Amazon website, consisting of several Amazon products from various departments, and b) publicly available dataset consisting of Amazon customer reviews from seven different types of products. The rest of the dissertation structure is as follows: Chapter 2 reviews relevant sentiment analysis and outlier detection work. Chapter 3 discusses the problem statement, Chapter 4 analyses the datasets used, and Chapter 5 presents the proposed Statistics-based Anomaly Detection and correction method (SADCM). Chapter 6 summarizes various sentiment analysis models, and Chapter 7 discusses and projects the experiment results. Chapter 8 showcases the conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

Social media has become a powerful platform for people to share their opinions and concerns on topics ranging from socio-economic to political to technological advancements. Iglesias et al. in (Iglesias & Moreno, 2019) discuss advances in various approaches in sentiment analysis their contributions and applications in multiple domains. The work in (Chakraborty et al., 2020) compiles all the studies related to various limitations of sentiment analysis on social media datasets. It discusses problems as trivial as spelling and grammatical mistakes to situations as critical as rumor-mongering, community shaming, riots, and protests arising from posts or comments on the Internet. It also highlights the increasing impact of research conducted on sentiment analysis applied to social media datasets. The study (Hou et al., 2020) analyzes previous literature based on modern social media applications. Hou et al. also features its impact in healthcare, disaster management, and business.

Sentiments or emotions tenaciously drive a consumer's decisions and views regarding a product or service. The research in (T. Hu et al., 2020) focuses on social media's impact on people from a spatial and temporal vantage point. Using Alteryx, it filters the tweets based on residential users from the 2016 United States Geo-tweets dataset. The results show a higher impact of tweets, especially those with positive sentiments, based on several features like location, content, and time. Cosmetic brands apply sentiment analysis to obtain a clear and comprehensive insight into consumers' thoughts on product quality and desires. In (Park, 2020), Park implements Term Frequency-Inverse Document

Frequency to analyze the polarity of customer opinions and brand satisfaction for 26 different cosmetic companies. The research also focuses on the factors affecting the nature of consumers' views.

Understanding a consumer's buying choices is a challenging assignment for a machine learning algorithm. Hu et al. in (S. Hu et al., 2020) introduce credibility, interest, and sentiment enhanced recommendation model, which consists of five segments, namely, feature extraction of the review, interest mining on the aesthetic of the comment, candidate feature sentiment assignment based on the nature of their fastText sentiment, and a recommendation module that utilizes credibility weighted sentiment score of the feature selected by the buyer and reviewer credibility evaluation that helps in weighing the credibility of the reviewer to avoid fake reviewers. The reviews also depend on a reviewer's experience, which might differ from one customer to another. Li et al. focus on this problem (M. Li et al., 2020) by recommending an algorithm inspired by Dempster-Shafer's evidence theory. They use hotel customer reviews of four different properties as a case study and extract information from various travel websites to identify the practicability and capability of the algorithm. Their approach can help the managers develop strategies based on the customer reviews to outrun their competitors. Sentiment analysis can be investigated at three levels:

1: **Document-level sentiment analysis**: This task uses sentiment analysis to determine if the overall document expresses a positive or negative opinion. Document-based sentiment analysis is the simplest way of sentiment analysis. The task assumes that each document describes an opinion on only one entity.

However, this assumption does not hold for cases where multiple entities are evaluated in one document. Therefore, a more refined analysis is required.

2: **Sentence level sentiment analysis**: In this level, the task is to determine if a sentence expresses a positive, negative, or neutral opinion about an entity. This level has an unbiased opinion that does not exist in document-based sentiment analysis. An impartial opinion is one in which no opinion is expressed in a sentence. It has the same assumption as document-based sentiment analysis, i.e., only one entity is expressed in a sentence.

3: **Feature-level sentiment analysis**: The first two levels are very effective when an entire document or sentence points to a single entity. However, people tend to talk about entities with many different aspects (attributes). For each aspect, people tend to have different opinions. Feature-level sentiment analysis is the finest-grained analysis. This usually exists in product reviews from several online companies like Amazon, Yelp, and eBay regarding products such as cars, cameras, and mobile phones.

Aspect-Based Sentiment Analysis (ABSA) identifies the feature/aspect of an entity/target in an opinion/review and then performs sentiment analysis on each element analysed. In this 3-star Amazon review on gloves, "*Good value for the money, however, they do not hold up very well. They rip easily*", the two aspects the consumer discusses are a) 'affordability' whose sentiment is positive as they are cheap, and b) 'durability' which carries a negative polarity. In (Jerripothula et al., 2020), feature-focused sentiment analysis is applied to the customer comments, and the review votes of various mobile products are collected from Amazon. The result indicates that the method helps the manufacturers in product development and the buyers make a personalized decision based on multiple features of the product. Ali et al. (Ali et al., 2020) study the customer reviews and feedbacks for ridesharing services to modify and uplift several organizations for Kansei engineering in India-Pakistan. Since the languages used commonly are Urdu/Hindi and English, the work converts all the reviews into English and performs ABSA. They also extract the most frequently used aspect to improve further the services provided based on customer demands. ABSA also helps classify reviews or comments based on various product or service features related to the opinion. ABSA has several challenges, such as the attention-based models may sometimes a) lead to a given aspect to incorrectly target grammatically irrelevant words; b) fail to diagnose unique sentence structures such as double negatives; and c) weigh only one vector depicts context and target. In (B. Zhang et al., 2020), Zhang et al. propose a knowledge-guided capsule network to address the above limitations using Bi-LSTM and capsule attention network. In (S. Zhang et al., 2020), the authors propose a multiclassification model to perform sentiment analysis based on a directed weighted model that helps distribute the sentiments of review comments per various limited threshold rules, which helps in higher accuracy and improved efficiency. The study (Liu et al., 2020a) summarizes the state-of-the-art ABSA methods using lexicon-based, machine learning, and deep learning approaches.

Generally, the ABSA methods can be categorized into a lexicon-based method (Zhang et al., 2019a), a machine learning method (Kiritchenko et al., 2014), and a deep learning method (Lakkaraju et al., 2014). Traditional ABSA methods mainly focus on using a group of feature engineering methods, such as bag-of-words (Y. Zhang et al., 2010) and part-of-speech (Rquez & Rodrfguez, 1998), to train traditional machine learning classifiers, e.g., Naïve Bayesian, Support Vector Machine, and Neural Network. Although the aforementioned methods can achieve comparable performance, their performance depends on hand-crafted features whose actuation is unfortunately labor-intensive.

Based on the observation of Twitter sentiment, Jiang et al. (Jiang et al., 2011) are the first to present the importance of targets. They also prove that 40% of classification errors in traditional classifiers are caused by their failure to consider the targets' information. They show that if traditional classifiers incorporate target-dependent features, they can gain better performance than target-independent classifiers.

In (Kiritchenko et al., 2014), the authors propose a feature-based Support Vector Machine for classification. It relies on the surface, lexicon, and parse features and obtains acceptable performance in terms of accuracy. However, they cannot achieve very high performance because the sparseness and discreteness of features restrict them.

As a probabilistic generative model, Sentence Latent Dirichlet Allocation is proposed (King et al., 2011). It is used to solve the problem caused by using Latent Dirichlet Allocation alone, i.e., Latent Dirichlet Allocation ignores the position information of words. An Aspect Sentiment Unification Model is proposed to extend Sentence Latent Dirichlet Allocation. It incorporates both aspects and sentiment. Both Sentence Latent Dirichlet Allocation and Aspect Sentiment Unification Model assume that words from one sentence are generated for a single topic.

The work (Kumar et al., 2015) introduce a feature selection technique for ABSA. The most relevant set of features for ABSA can be automatically extracted based on their proposed method. This method is based on Particle Swarm Optimization (Poli et al., 2007), which is a computational method whose particles or solutions evolve iteratively until a local or global optimum is found. After removing the irrelevant set of features, Conditional Random Field is used as a learning algorithm in (Lafferty et al., 2001) to catch the most relevant features on the benchmark datasets of SemEval-2014 Task-4 (Pontiki et al., 2016).

The work (Vo & Zhang, 2015) presents a novel context representation for ABSA, specifically for Twitter data. The traditional feature extraction methods are based on syntax. However, the accuracy of using syntactic analysis is considerably lower on Twitter than on conventional text. This work uses a distributed word embedding and neural pooling function to enrich features automatically and solve the low parsing accuracy problem. However, this method highly depends on the effectiveness of the laborious feature engineering work, and it can easily reach its performance bottleneck. Using a pooling function to capture syntactic and sentiment information for Twitter data is indeed over intuitive.

Since information is so readily available in this digital age, buyers tend to read customer reviews and comments before purchasing a product, which affects their purchasing decision. Researchers usually focus on the review body, but a review contains more information than that, which is generally not exploited, such as review time, number of helpful votes, review time, reviewer id, and review rating. In (Benlahbib & Nfaoui, 2020), Benlahbib and Nfaoui visualize the reputation of a product differently by considering all the parameters and projecting the reputation value, opinion category, top positive review, and top negative review. They implement the time of review and the number of helpful votes for each review from the Transformers model to Bidirectional Encoder Representations. This helps to predict the probability of the nature of review sentiment. They also propose equations that calculate the reputation value for a product. Extensive research is being conducted not only focusing on sentiment analysis in English but also several other languages like Arabic (Almaghrabi & Chetty, 2020), Persian (Basiri et al., 2020) (Jamshidi Nejad et al., 2020), Urdu (Younas et al., 2020), Hindi (Yadav et al., 2021), Russian (Yaqub et al., 2020), Chinese (G. Li et al., 2020), Indonesian (Saputra et al., 2020).

Several studies have been conducted on sentiment analysis (García-Mendoza et al., 2020) and its e-commerce application. With the increase in online consumption, e-commerce enhancement has become a hot topic for research. Many scholars have introduced methods focusing on deep neural networks (Y. Wang et al., 2021), probabilistic classifiers (Technocrats Institute of Technology et al., 2019), linear classifiers (Saranya et al., 2020), lexicon-based approaches (B. Zhang et al., 2019b), or decision trees (Singh & Tripathi, 2021) to increase accuracy and efficiency. In (L. Wang et al., 2020), Wang et al. propose an iterative sentiment analysis model called SentiDiff, which predicts polarities in Twitter messages by considering the interconnections between textual information of Twitter messages and sentiment diffusion patterns. The authors in (Shofiya & Abidi, 2021) use Support Vector Machine to identify the keywords and extricate the sentiment polarity of Twitter data specific to Canada on social distancing due to COVID-19. Zhang et al. (Y. Zhang et al., 2020) introduce a convolutional multi-head self-attention memory network to glean valuable and intricate semantic information from sequences and aspects of a sentence. This algorithm uses a convolutional network to capture n-gram grammatical knowledge and multi-head self-attention to acknowledge the linguistic information of the sequence by the memory network. Abdalgader et al. (Abdalgader & Shibli, 2020) apply a lexicon-based word polarity identification method by studying the semantic relatedness between the set of the target word and synonyms of words surrounding the target on several benchmark datasets. The result has outrun several existing methods that use pairwise relatedness between words at term-level around the target over a fixed size. The performance of various sentiment analysis methods differs due to such factors as datasets, feature representations, or classification processes. Liu et al. (Liu et al., 2020b) conducted a detailed survey on several deep learning approaches for aspect-based sentiment analysis using benchmark datasets evaluation metrics and the performance of the existing deep learning methods.

Sentiment analysis is one of the most challenging tasks of NLP because human emotions are complicated and understanding them is not simple. There might be a stark difference between what they mean and what they indicate; even humans struggle to analyze sentiments accurately. Over the years, data scientists have been trying to improve the sentiment analysis classifiers for better and more accurate results. While doing so, there are several challenges they usually come across, amongst which some of them are discussed as follows:

1: **Subjectivity & Tone**: A sentence can either be subjective or objective. Objective sentences are informative, while subjective ones contain explicit sentiments. For example, in this 5-star bedding Amazon review, "*This duvet is lightweight, fluffy, and doesn't make my daughter sweat- even in our warm climate. It comes with corner and side loops to secure it to the duvet cover. The duvet actually has more loops available than the specific cover I purchased, but* *it still stays perfectly in place inside the cover. Very happy with this purchase.*", the start of the comment carries product information, while the sentence that concludes the review is a subjective sentence as it consists of the sentiment of the reviewer.

2: **Context & Polarity**: Sentiment analysis without context is tricky. However, machines cannot learn about contexts automatically if they are not mentioned explicitly. One of the problems that arise from context is changes in polarity. For example, consider these two hypothetical responses a) "Everything ok about it"; and b) "Absolutely nothing", were for the question "What did you like about the product?", The reply a) "Everything ok about it"; would be positive and the reply b) "Absolutely nothing", would be negative. But, if the responses come from answers to the question "What did you dislike about the event?", the nature of opinions changes altogether.

3: I**rony & Sarcasm**: Irony and sarcasm are two techniques of stating one's opinion where people use negative words to express positive sentiments and vice-versa. This is challenging for a machine to understand irony and sarcasm without having a thorough understanding of the context of the situation in which a feeling was expressed. For instance, the following Amazon review of car accessories is such a perfect example "*This is so COOL!!! I balance a ball on mine and try to keep it from rolling off the edge while driving by tilting the* *wheel back and forth and using the gas and brake. I must do this well as everyone around me honks with encouragement.*"

4: **Comparisons**: Treating comparisons in sentiment analysis is another challenge worth tackling. Amongst the given three texts, a) "This product is second to none.", b) "This is better than older tools.", and c) "This is better than nothing." The first comparison doesn't need any contextual clues to be classified correctly. It carries a positive sentiment, while the second and third texts are a little more challenging to categorize. Once again, context can make a difference. For example, if the 'older tools' in the second text were considered useless, then the second text is pretty like the third text.

5: **Emojis**: There are two types of emojis according to (Guibon et al., 2016) a) Western emojis (:D), which are encoded in only one or two characters, and b) Eastern emojis (¯ \ (ツ) / ¯), which are a longer combination of characters of a vertical nature. Emojis play an important role in the expression of sentiment in a text, particularly in tweets. For example, in this 3-star Amazon review of baby apparel, "*The Disney onesies are very cute 💝 Fast shipping. However, they came in different sizes. *2. newborn onesies *2. 0-3 months & * 1. 3-6 months onesies. However, since I have different sized reborn I decided not to return them and make the best of it 💝 I took two stars off they weren't all the right sizes & upon reading reviews, it seems as though this is a recurring issue with other customers and the seller has not managed to resolve this issue by* *giving the actual sizes customers actually want.*", to achieve an improved sentiment analysis performance, data pre-processing of the content is a vital step which also includes transforming of both Western and Eastern emojis into tokens and whitelist them, i.e., always take them as a feature for classification purposes.

6: **Defining Neutral**: Classifying a text into positive or negative classes based on the sentiment of the text is a more straightforward task compared to classifying the text into a neutral class. Defining what is meant by neutral is another challenge to tackle to perform accurate sentiment analysis. The neutral, positive, or negative definition is essential for training a text in sentiment analysis models. Since tagging data requires that tagging criteria be consistent, a good description of the problem is necessary. One of the ways to define neutral text is when the sentence is an objective text, which means the sentence does not contain explicit sentiment, or any text that might contain irrelevant information can be considered a neutral text. The neutral class should be defined carefully as it might add noise to the classifier, negatively impacting the performance.

7: **Human Annotator Accuracy**: Sentiment analysis is a tremendously difficult task even for humans. On average, inter-annotator agreement (a measure of how well two (or more) human labelers can make the same annotation decision) is pretty low when it comes to sentiment analysis. And since machines learn from labeled data, sentiment analysis classifiers might not be trained as precisely as expected.

Outliers are extreme values that diverge from the rest of the data samples. It might occur due to an imbalanced dataset or experimental error, or novelty. The research (Shanmugam et al., 2020) defines an outlier in its experiment as any tweet in a Twitter dataset that is not relevant to the topic in consideration. Once the outliers are detected and eliminated, it is noticed that the algorithm's accuracy improves significantly. Similarly, in (Schmitt & Spinosa, 2018), it is observed that before implementing a convolutional neural network to the document to be classified, the efficiency increases if outliers are identified and erased by using a density-based clustering algorithm, and the computational cost decreases. Kim et al. (Kim et al., 2019) applied a combination of four outlier detection methods, namely a) Gaussian density estimation, b) Parzen window density estimation, (c) Principal component analysis, and (d) K-means clustering to identify malicious activities in an institution using a user log database. The outlier identification methods can be broadly categorized into statistical-based (Neagu et al., 2017), distance-based (Ahmed et al., 2019), graph-based (Cui et al., 2019), clustering-based (Verma et al., 2021), density-based (Corain et al., 2021), and ensemble-based (Sapegin & Meinel, 2020). Once the outliers are detected, it is crucial to delete, consider, or modify the outlier. This usually depends on an outlier's effect on the dataset if it is deleted or tampered with. The condition of an outlier can vary for different applications and datasets; for instance, if in a population estimation survey the number of people with height over 7ft is very few, then this data can be verified and kept as they are natural outliers. In contrast, if in a dataset with various brands of shoes, the price of one or two are extraordinarily high, then those outliers can be deleted before calculating the average cost of a pair of shoes. Table 2.1 compiles some of the current research work done in the field of outlier detection.

Table 2.1 Recent Research Work on Anomaly Detection

Study	**Dataset**	**Outlier detection method**	**Outlier correction method**
(Riahi-Madvar et al., 2021)	UCI repository	PCA-based subspaces (SODEP)	N/A
(Studiawan et al., 2020)	Operating system logs	pylogsentiment	N/A
(X. Wang et al., 2021)	UCI repository	k-nearest neighbor medoid-based method	N/A
(Du et al., 2021)	Wisconsin breast cancer dataset	FAST-ODT based on outlier detection tree	N/A
(Z. Wu et al., 2020)	Review streams	Anomaly clearing cumulative SUM (ACCSUM)	Elimination
(Shanmugam et al., 2020)	Twitter dataset	k-means clustering algorithm	Elimination
(Ding et al., 2019)	Driver’s emotional state dataset	Gaussian mixed model (GMM)	N/A

CHAPTER 3

PROBLEM STATEMENT

Our buying habits have been vastly revolutionized with the growth of the Internet. Amazon is one of the biggest online retail stores. Majority of the Amazon customers check the product reviews before making a purchase. Thus, these customer reviews often emerge as the deciding factor for many buyers. So, it is essential to have an efficient system for analyzing and monitoring them. Increasing product inventory increases the number of reviews; hence the review analysis must adapt to changing trends dynamically. Manually analyzing these reviews is a traditional approach, but it could be time-consuming and a tedious job. Similarly, searching and then comparing customer reviews for a product can be frustrating for a potential buyer. For these reasons, artificial intelligence can be potentially used to study and analyze these customer reviews.

Amazon customer reviews are laden with information; hence they are significantly crucial for the market. These reviews usually consist of information about the product, its most popular features among customers, and the features that need improvement, along with the market demand and the trending features of a product. Efficient monitoring of these reviews helps the sellers to improve their products based on the market trend.

Amazon reviewers can be broadly classified as verified purchasers or non-verified purchasers. Verified purchasers have purchased the product from the Amazon marketplace, and the latter are those who use other sites to buy the product. Customers with a verified purchase are generally highly affected by the product to such an extent that they invest time from their busy schedule into writing this review. On the contrary, individuals who compose the non-verified purchase reviews often have never actually purchased the product. Instead, they have heard about it from friends and family or came across an advertisement for it online such as Facebook or Google AdWords. Unfortunately, non-verified purchases include entirely fake reviews as well.

An anomaly is any data point in a dataset that deviates from the usual pattern in a dataset. Abnormalities are sometimes referred to as novelties or outliers, as these points typically "lie outside" the data patterns in a dataset. There are several types of outliers, but the common trait they share is that they behave much differently than other data points of that dataset. The three major types of anomalies are discussed below:

1: **Point Anomalies**: Point anomalies occur when data points fall outside the anticipated range, norm, or pattern, resulting in an unexpected data point. Such anomalies usually help identify various problems, such as providing insight into potential issues in the manufacturing process.

2: **Collective Anomalies**: Collective anomalies occur when the data points lie between the expected range, but examining the data points reflects a remarkable result, atypical pattern, or unexpected behavior.

3: **Contextual Anomalies**: Contextual Anomalies don't focus on specific data points or data groups but instead look at anomalies in the overall data context.

Outlier detection has been an actively researched area for several decades due to its broad applications in many critical domains such as risk management, compliance, security, financial surveillance, health and medical risk, and AI safety. Although it is a problem widely studied in various scientific communities, including data mining, machine learning, computer vision, and statistics, some unique problem complexities and challenges require advanced approaches. Some of the challenges are as follows:

1: **The difficulty of obtaining high anomaly detection recall rate**: Since the nature of outliers is of rare occurrence; it becomes challenging to identify all the anomalies. In this case, many standard occurring instances are wrongly reported as outliers while authentic, yet sophisticated abnormalities are missed.

2: **Anomaly detection in the high-dimensional dataset**: Novelties often exhibit evident abnormal characteristics in a low-dimensional space and become hidden and unnoticeable in a high-dimensional area.

3: **Noise-resilient outlier detection**: Some supervised anomaly detection models assume that the given labeled training dataset is clean, which can be highly vulnerable to noisy instances labeled as an opposite class label.

An Amazon customer review consists of two significant aspects that help convey the nature of a customer's opinion a) star rating and b) review comment. Based on the reviewer's experience, a star rating indicates the intensity of the former's happiness or satisfaction for a product. A 1-star rating indicates a terrible experience, and a 5-star rating represents a strong product recommendation. The review comment carries a more in-depth experience and opinion of a buyer towards the product. Some of these reviews consist of the general opinion a person holds towards a product and critically evaluate each aspect of a product. Although star rating and review comments have a strong correlation, they still carry significantly different information. The star rating helps give potential buyers the intensity of recommendation to purchase the same by previous shoppers. Review comment makes the potential buyers aware of various experiences regarding the product with critical evaluation of the same.

This Amazon review of a gym bench "*Bought this bench Wednesday, it arrived Saturday. I could tell from the packaging that it was of great quality. Well packaged, all parts packaged neatly & tightly not to move during shipping. I needed a bench that inclines & declines for the versatility of chest workouts. Searched many benches and many sites. Along with many reviews, and was not pleased. However, this bench had one review, so I took a chance based on that review which was good. This bench is phenomenal. Great quality and very sturdy. Glad I got it since it was all sold out a few days later. A great addition to my flat bench. But of course, this bench serves the dual purpose of flat/incline/decline. Great bench, I love it!*!" has a 5-star rating. This means the buyer highly recommends this product to a potential buyer. Now, the review comment, on the other hand, not only discusses how good the product is or what feature of the product is best but also shares the experience that the reviewer encountered. Another feedback comment for the same product says, "*After my failed purchase from another brand name Flybird I switch to this one with a higher price hoping it would be better, but no. The chair* *came defective, as the pictures above show you. On 1 side, you can see that the rotation rod holding the backrest is visible while it is not supposed to (It needs to look like the 2nd pic that is completely closed), which makes the backrest extremely shaky. I even had to call for my dad help, and he's been working in the Machinery-Assembly industry for 30 years, and he told me better to return this piece of garbage cause with the rod-like that there is no way we can fix it*" The star rating corresponding to this review is 1-star which means this particular reviewer had a very horrible experience which was clearly explained in details by the reviewer in the comment. Hence, each star rating has a different reason behind it and each review comment consist of various experience that the purchaser encounters and different information about the product.

In some cases, the nature of star rating and comment do not match. For instance, this Amazon review of a luggage "*I want to return this. Backpack....the zipper is not working on one opening. Very disappointed since I had it on a trip and it was the first time to use it. How do I return?*" has a 5-star rating which means the reviewer should be thrilled with the product, but instead the nature of the opinion is negative. The review reflects a negative aspect of the product, which the reviewer did not like and contradicts the corresponding star rating. Another such example would be this 1-star review which holds a very positive sentiment towards the product "*It's a perfect size for traveling, and I love the different compartments. I would recommend this item for everyone. All women need a cosmetic bag, and it's pretty.*". Such reviews whose true nature of opinion contradicts the nature of star rating are considered as abnormalities. These anomalies should be identified while performing sentiment analysis of customer reviews. Generally, to clean the dataset, the identified outliers are removed. Still, each review carries unique information and opinion towards a product, so removing these outliers will result in loss of information. The proposed method, SADCM, tries to overcome these two problems a) identifying the outliers, which are defined as any customer review which carries an opposite sentiment with respect to the corresponding star rating, and b) correcting the identified outliers without removing them to prevent any information loss.

CHAPTER 4

DATASET

With the advancement in the Internet and cloud computing (Ghahramani et al., 2017), data collection has become more accessible for researchers. Even though public datasets are abundant, it is favorable to work with collected datasets because a) collected datasets can be updated regularly and b) these datasets are highly customizable according to the requirement of the algorithm. Based on a survey from Feedvisor, an article in Forbes concluded 89% of the buyers chose Amazon over other e-commerce websites to make online purchases (Kiri Masters Forbes, 2019). Each Amazon review carries multiple information like name, date, place, star rating, verified purchase, the number of buyers who find the review helpful, and the images added by the reviewer, as shown in Figure 4.1. Datasets used in this dissertation can be broadly classified into a) collected dataset and b) publicly available dataset of Amazon customer review.

Figure 4.1 An example of Amazon customer review along with its various elements.

4.1 Collected Amazon Customer Review Dataset

Amazon is one of the largest marketplaces in the continental United States. Thus, the datasets from Amazon are highly enriched in customer data and may help understand customer sentiment in a better way. Datasets used in this dissertation (SADCM-Datasets. Zip, 2021) consists of product reviews, starting from the year 2008 to 2020; we collected from Amazon.com seven popular and trending products, spanning across different domains, namely, book (Harry Potter Paperback Box Set (Books 1-7)), pharmaceutical (Turmeric Curcumin Supplement by Natures Nutrition), electronics (Echo Dot 3rd Gen by Amazon), grocery (Sparkling Ice Blue Variety Pack), healthcare (EnerPlex 3-Ply Reusable Face Mask), entertainment (Harry Potter: The Complete 8-Film Collection), and personal care (Nautica Voyage By Nautica).

The dataset scraped from Amazon consists of 35,000 Amazon customer reviews, including the product name, comment date, star rating, and the number of helpful votes. Hence, all the individual datasets consist of 5000 Amazon product reviews. Figure 4.2 shows the number of reviews against each star rating accumulated for all seven collected datasets individually. For the dataset named '*Book*', it can be noticed that more than 80% of the customer reviews have a star rating of 5 while less than 4% of the reviews hold 3, 2, or 1-star ratings, and approximately 8% of the remaining reviews correspond to a 4-star rating. This pattern is repeatedly observed for all the other datasets collected from the Amazon website. Hence, it can be concluded that the highly positive star rating (5-star) dominates the dataset by having the maximum number of corresponding reviews, and there is very few negative (1 and 2-star) and moderately positive (3 and 4-star) star ratings as compared to the former.

Figure 4.2 J-shaped distribution of the tallied reviews from all the accumulated datasets.

The skewed nature of all the datasets observed in Figure 4.2 thus results in a J-shaped distribution. There can be multiple reasons behind such bias towards extremely positive reviews. People usually agree with and write about the positive ratings and comments quickly but are generally skeptical about the negative ratings or comments. When a consumer notices an extremely positive review, it usually influences the consumer's opinion resulting in the switching of star rating. A higher rating has also been observed to easily influence a consumer to increase the valuation, while the reverse is not true (N. Hu et al., 2009). Table 4.1 represents the consumer review distribution across the different star ratings in all the collected datasets individually. The results show the same biases of customer reviews towards a 5-star rating as compared to the rest.

Table 4.1 Review Distribution across Different Star Ratings

Dataset	**5-Star Rating**	**4-Star Rating**	**3-Star Rating**	**2-Star Rating**	**1-Star Rating**
Book	4134	395	181	122	168
Electronics	3567	770	94	51	518
Entertainment	2485	1062	797	271	385
Grocery	3402	683	134	104	677
Health Care	3014	910	263	196	617
Personal Care	3287	338	641	200	534
Pharmaceutical	3190	855	184	114	657

Figure 4.3 represents a graphical distribution of the average number of helpful votes per review for all the collected datasets. It can be noticed in the dataset consisting of book

reviews that more than 60% of the helpful votes are for 1-star reviews while the rest have less than or approximately 10% of the votes individually. All the other datasets have the same tendency, which indicates that customers find the extremely negative reviews as the most helpful for making buying decisions or understanding a product. Extremely negative reviews are more critical about the product, its features, packaging, delivery, usefulness, cost, and authenticity. The buyers often expect the product that is not met because they analyze each aspect of a product carefully, which generally could have been compromised or overlooked. It becomes easier for a consumer to decide about buying a product if they understand the various aspects of a product and the extremely negative experiences of former buyers. Table 4.2 compiles the average helpful vote per customer review in each dataset. It can be observed that most customers find extremely negative reviews most informative and beneficial.

Table 4.2 Average Helpful Vote Distribution Per Review across Different Star Rating

Dataset	**5-Star Rating**	**4-Star Rating**	**3-Star Rating**	**2-Star Rating**	**1-Star Rating**
Book	6.27	1.89	9.72	10	52.3
Electronics	4.61	4.54	1.23	1.52	28.11
Entertainment	1.31	0.11	1.27	0.69	5.19
Grocery	0.77	0.35	0.23	0.58	1.45
Health Care	1.01	1.04	0.45	0.38	1.22
Personal Care	0.54	0.46	0.06	0.23	1.41
Pharmaceutical	4.19	2.34	0.43	0.72	9.64

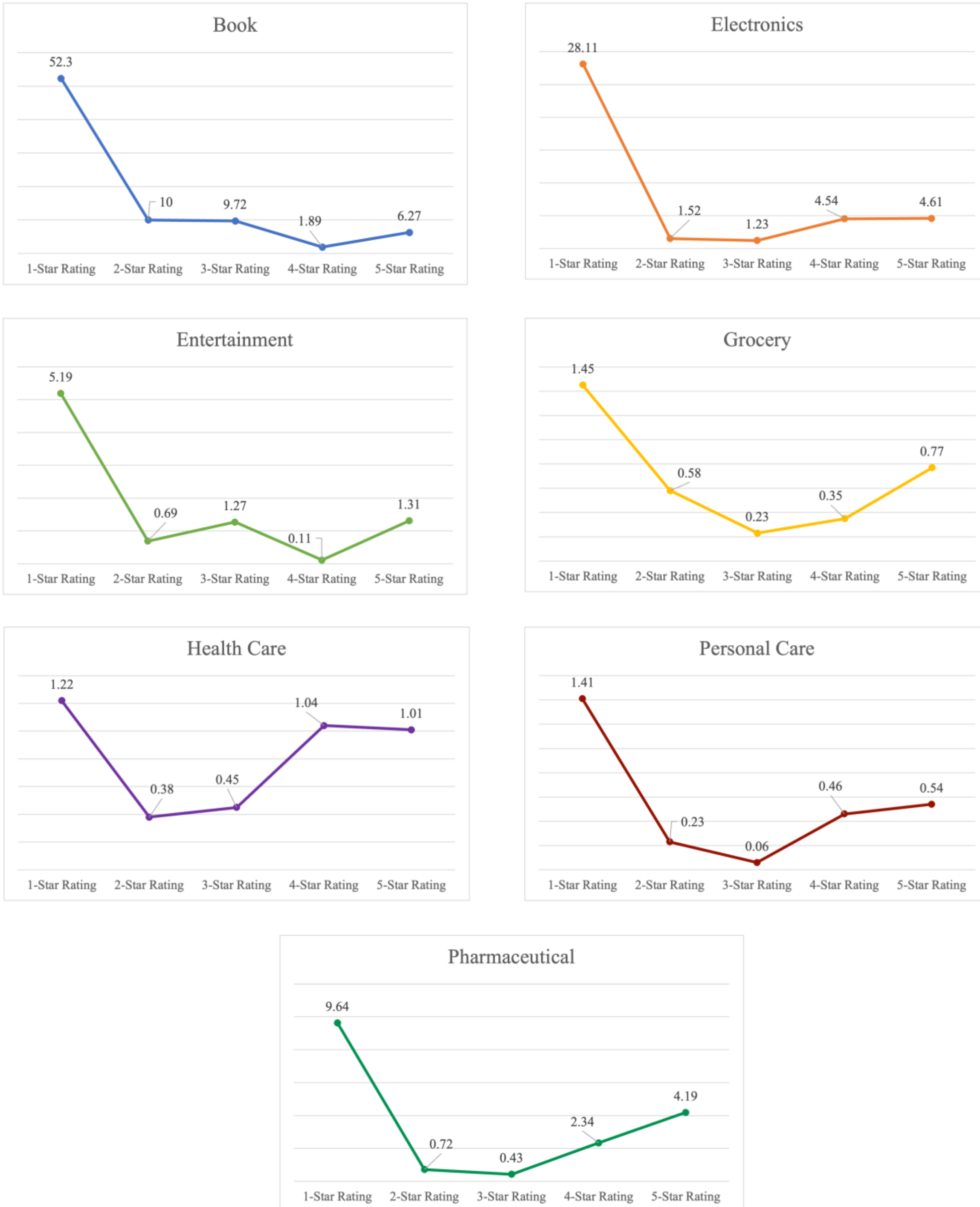


Figure 4.3 Average helpful votes per review across different star ratings.

4.2 Public Dataset of Amazon Customer Reviews

Customer Reviews (a.k.a. Product Reviews) are one of the fascinating features of Amazon. Millions of Amazon customers have contributed to the marketplace by sharing reviews, and feedback about purchased or used products. An estimated count of over 130 million reviews was recorded over the last two decades. This naturally makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others.

The data is available in TSV format from the amazon-reviews-pds S3 bucket in AWS US East Region. Each line in the data files corresponds to an individual review. The dataset contains the customer review text with accompanying metadata, consisting of three major components:

1: A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to encourage the study of the properties of the customer reviews along with their evolution, thus understanding how people rate and share their experiences with respect to products at scale.

2: A collection of reviews about products in multiple languages from different Amazon marketplaces focussed on facilitating analysis of customers' perceptions of the same products and more comprehensive consumer preferences across languages and countries.

3: A collection of reviews that have been identified as non-compliant concerning Amazon policies. This is intended to provide a reference dataset for research on detecting promotional, spam, or biased reviews.

The seven domain reviews selected are accessory, beauty, fashion, furniture, jewelry, luggage, and toy. Each domain dataset consists of approximately 100,000 Amazon customer reviews. The review distribution based on star ratings for these datasets is projected in Table 4.3.

Table 4.3 Review Distribution for Amazon Customer Review Dataset

Dataset	**5-Star Rating**	**4-Star Rating**	**3-Star Rating**	**2-Star Rating**	**1-Star Rating**
Accessory	59531	17890	8224	5438	8917
Beauty	65557	13583	7292	4982	8586
Fashion	54621	18672	10682	9608	6415
Furniture	56623	19363	9358	5480	9176
Jewelry	62892	14129	8558	5515	8906
Luggage	62138	17538	7981	5253	7090
Toy	67579	13043	7161	4330	7887

CHAPTER 5

STATISTICS-BASED ANOMALY DETECTION AND CORRECTION METHOD (SADCM)

5.1 Interquartile Range

In other words, in descriptive statistics, the Interquartile range (IQR) measures statistical depression across the dataset. The IQR may also be called the mid-spread, middle 50%, or H-spread, as it is a measure of dissemination of the middle half of a dataset. It is defined as the difference between the 75th and 25th percentiles of the data. Traditionally, a dataset can be represented using the five-number summary, which includes the lowest and highest value, median, and first and third quartile, the middle number between median and first and last number, respectively (Hussain & Aleem, 2018). These values exhibit more information about a dataset as compared to just rows and columns. Figure 5.1 is an example of the box plot distribution of a dataset.

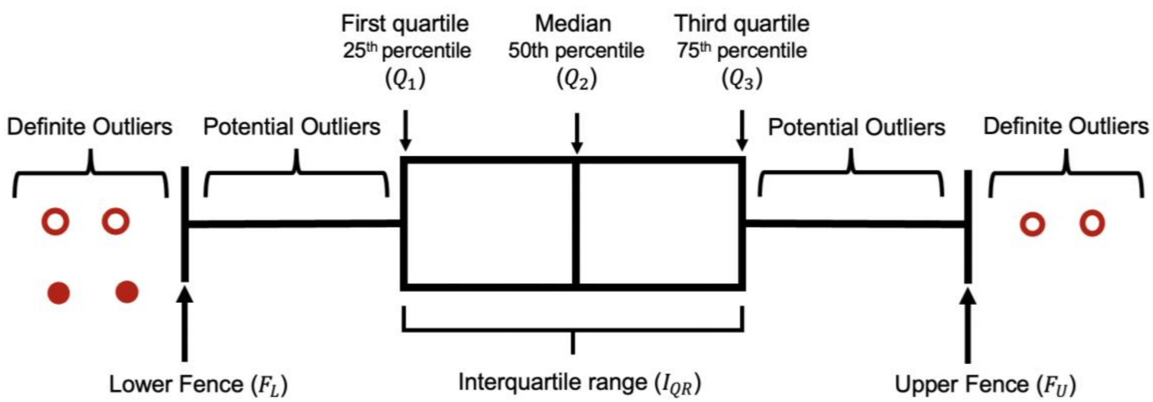


Figure 5.1 Box plot (with interquartile range) of a normal distribution for outlier detection.

Q_1 and Q_3 are the intermediate points of the first and second half of an ordered dataset, respectively, and Q_2 is the median value of a dataset. For example, in an arranged dataset $A = \{1,1,2,3,5,6,7\}$, Q_2 is 3, which is the median value or the fourth number of the dataset. Q_1 is 1 as it is the center value of the first half, 6 is Q_3 as it is the midpoint of the second half of the dataset. Businesses can use the IQR as a pointer for their income rate, benchmark other companies, or do day-to-day activities like determining the day's temperature range or measuring the range of air pollution. It can also help data analysts in studying the skewness of a dataset or in identifying outliers.

Difference between Q_1 and Q_3 is the interquartile range (I_{QR}), which reflects the spread of the dataset about the median.

$$I_{QR} = Q_3 - Q_1 \tag{5.1.1}$$

The lower and upper fences can be represented as:

$$F_L = Q_1 - 1.5I_{QR} \tag{5.1.2}$$

$$F_U = Q_3 + 1.5I_{QR} \tag{5.1.3}$$

Data in a dataset that exists beyond the bounds of F_L and F_U is an outlier. 1.5 preserves the sensitivity of the dataset. A larger scale than 1.5 would consider outliers as a datapoint, while the reverse would include data points in outliers.

In a dataset, there are two types of outliers, suspected or potential outliers and definite outliers. A potential outlier (O_P) is the data that is suspected as a possible outlier if it satisfies:

$$F_L < O_P < Q_1 \quad \text{or} \quad F_U < O_P < Q_3 \tag{5.1.4}$$

A definite outlier (O_D) is the data that is an absolute outlier if it complies with:

$$\mathrm{O_D} < \mathrm{F_L} \quad \text{or} \quad \mathrm{F_U} < \mathrm{O_D} \tag{5.1.5}$$

5.2 Definitions for SADCM

R consist of all the customer reviews in a dataset such that $R = \{r_1, r_2, r_3, \dots, r_N\}$ where, r_i denotes each review and r_i^* is the star rating of review r_i. The following definitions are presented to understand our proposed Statistics-based Anomaly Detection and correction method (SADCM).

Definition 1: r_i is positive if $r_i^* \geq 4$, where $r_i \in R$. Any review with a star rating of four or more is considered a positive star-rated review, denoted by S^+.

Definition 2: r_i is negative if $r_i^* < 4$, where $r_i \in R$. Any review with less than a four-star rating is considered a negative star-rated review, denoted by S^-.

Definition 3: $T_V(r_i) = 1\ if\ r_i \in S^+$ and $T_V(r_i) = -1\ if\ r_i \in S^-$. The target value of review r_i is 1 if it is a positive star-rated review and -1 otherwise, denoted by T_V.

Definition 4: $V_D(r_i) = d\big(T_V(r_i), C_V(r_i)\big)$, where $C_V(r_i)$ is the compound sentiment score of r_i predicted by a sentiment analysis algorithm. The value difference of review r_i is the Euclidean distance between $T_V(r_i)$ and $C_V(r_i)$ of the corresponding review, denoted by $V_D(r_i)$. Since the range of both T_V and C_V is [-1, 1], the range of V_D is [0, 2].

5.3 Proposed Algorithm

The star rating assigned to a customer's review is generally considered the comment's ideal sentiment. There are instances when a customer might have assigned a positive star rating to a review, but the nature of the feedback is negative. This 4-star Amazon customer review on a thermometer "*Purchased the thermometer to have a method to check temperatures by non-contact. The thermometer's box and content were not sealed, which bothered me because of COVID*." carries a negative sentiment but has a positive rating which is contradictory. These ratings of reviews can be corrected to their correct star rating to improve the efficiency of a sentiment analysis algorithm.

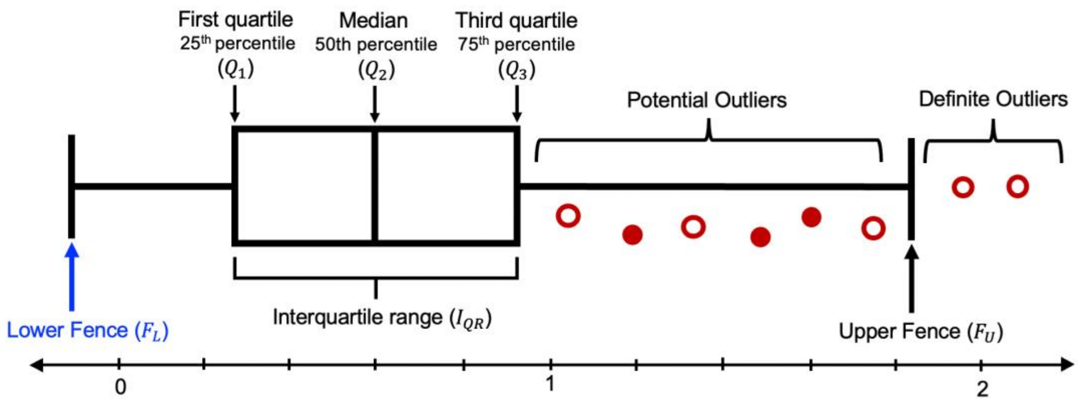
SADCM consists of two major parts, namely, a) detection of these outliers and b) correction of these identified anomalies. It has the following six steps:

Input: The input for SADCM is any dataset containing customer reviews (r_i) and their corresponding star ratings (r_i^*).

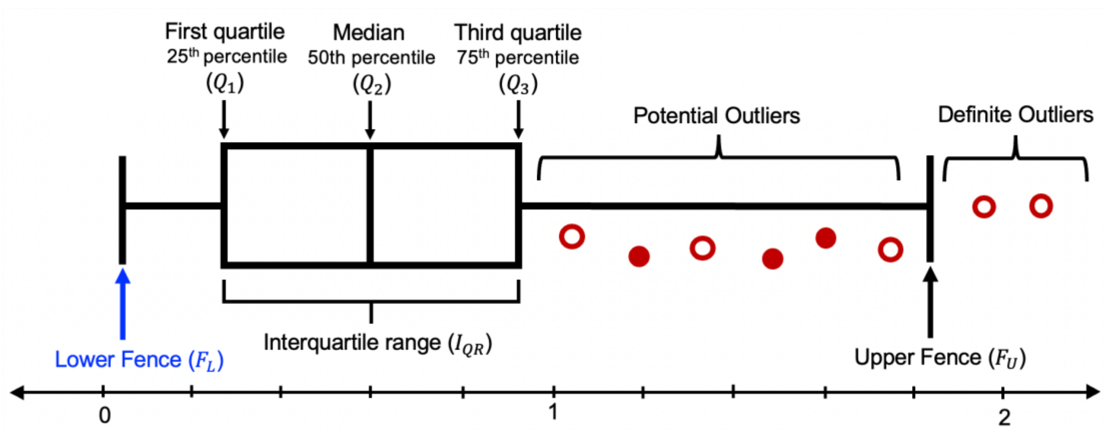
Step 1: T_V is calculated using r_i^*. If r_i belongs to S^+ then $T_V = 1$ and if r_i belongs to S^- then $T_V = -1$. Since this work focuses on binary classification of the sentiments of customer reviews, the values assigned to T_V are -1 or 1.

Step 2: V_D is calculated between T_V and C_V. The value of V_D is always positive. Since the minimum and maximum value T_V and C_V can hold is 0 and 1, the range of V_D is between 0 and 2. Figure 5.2 is an example of the box plot distribution of S^+. Since the minimum value V_D can hold is 0 Figure 5.2 (a) depicts box plot of S^+ when F_L is negative and (b) depicts box plot of S^+ when F_L is positive. Figure 5.3 is an example of the box plot distribution of S^-. Since the maximum value V_D can hold is 2 Figure 5.3 (a) depicts box plot of S^- when F_U is greater than 2 and (b) depicts box plot of S^- when F_U is less than or equal to 2

Step 3: After analyzing the dataset, it can be construed that S^+ has some reviews whose sentiment does not match the nature of star rating; hence, they are considered outliers. On the other hand, S^- has very few reviews whose opinions match the essence of their respective star rating; hence, the reviews correctly assigned to their corresponding star ratings are considered outliers. This implies that most negative comments are incorrectly rated; therefore, the outliers, in this case, would be the correctly rated comments. In other words, the incorrectly labeled reviews are all the reviews in S^-, excluding the ones which are outliers. Hence, the dataset is split into S^+ and S^-.



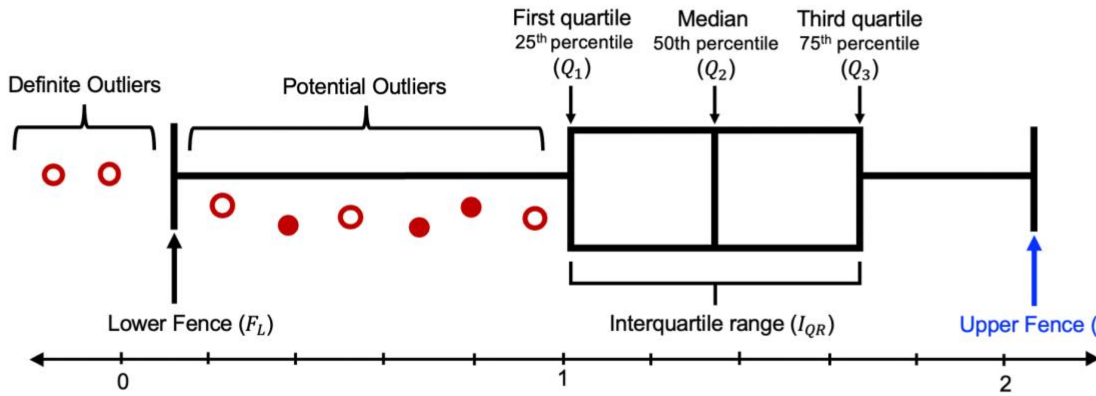
(a)



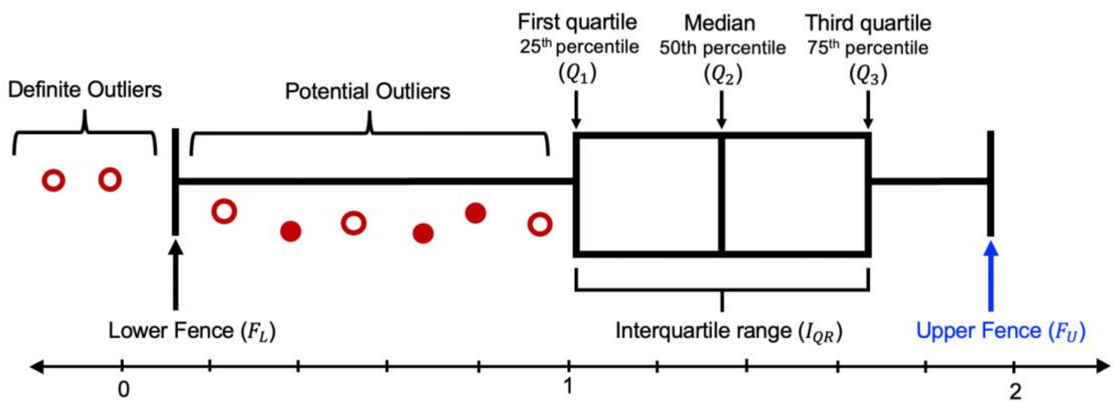
(b)

Figure 5.2 Box plot (with interquartile range) of S^+ distribution for outlier detection.

Step 4: After analyzing the dataset, it can be construed that S^+ has some reviews whose sentiment does not match the nature of star rating; hence, they are considered outliers. On the other hand, S^- has very few reviews whose opinions match the essence of their respective star rating; hence, the reviews correctly assigned to their corresponding star ratings are considered outliers. This implies that most negative comments are incorrectly rated; therefore, the outliers would be the correctly rated comments. In other words, the incorrectly labeled reviews are all the reviews in S^-, excluding the ones which are outliers. Hence, the dataset is split into S^+ and S^-.



(a)



(b)

Figure 5.3 Box plot (with interquartile range) of S^- distribution for outlier detection.

Step 5: In S^+, if F_L is negative, then O_s can be calculated as $Q_3 + I_{QR}$ else, $F_U - I_{QR}F_L$. Since the range of V_D is [0, 2], the least value it can hold is 0. In S^- if $F_U > 2$, then O_s can be calculated as $Q_1 - I_{QR}$, else, $Q_3 - I_{QR}F_U$. We compute O_s as follows:

$$\text{For } S^{+} \quad O_s = \begin{cases} Q_3 + I_{QR}, & F_L < 0 \\ F_U - I_{QR}F_L, & F_L \geq 0 \end{cases} \tag{5.3.1}$$

$$\text{For } S^{-} \quad O_s = \begin{cases} Q_1 - I_{QR}, & F_U > 2 \\ Q_3 - I_{QR}F_U, & F_U \leq 2 \end{cases} \tag{5.3.2}$$

Step 6: In S^{+}, $V_D(r_i) \geq O_s$, if r_i is outlier. For S^{+}, customer comments, whose $V_D(r_i) \geq O_s$, are outliers. In S^{-}, if $V_D(r_i) \leq O_s$, if r_i is outlier. For S^{-}, customer comments whose $V_D(r_i) \leq O_s$, are outliers. These five steps complete the outlier detection process.

Step 7: T_V of reviews labeled as outliers in S^{+} is reversed, meaning a comment with $T_V = 1$ now gets re-labeled as -1 and vice versa. On the contrary, for S^{-}, T_V of reviews that are not labeled as outliers are reversed. This step is vital as it performs outlier correction by changing the nature of r_i^{*}.

Output: The output of the proposed algorithm is the dataset consisting of reviews with their corrected nature of star ratings which means a positive natured review is labeled as 1 and the negative natured review as -1. SADCM helps in detecting the outliers and correcting them without eliminating or modifying any review.

The above steps are realized in SADCM. After its execution, we can perform a more accurate sentiment analysis of the revised dataset, and the performance matrix of SADCM is obtained.

CHAPTER 6

SENTIMENT ANALYSIS MODELS

The sentiment analysis algorithm evaluates the nature of text in the form of sentences, paragraphs, files, or comments. The algorithm's input is a string, and the output is the corresponding sentiment that the former carries. The customer opinions can be categorized into four subcategories:

1: **Direct opinion**: The direct opinion is when a customer review directly states the opinion it carries. For example, this Amazon 1-star cast iron grill reviewer is "*HORRIBLE... don't waste your money! Got 4 of them for my son's graduation open house for a s'mores bar. Did NOT even cook one marshmallow. Wouldn't stay lit, and when it was lit it wouldn't heat up. Very disappointing!*", directly reflects a negative opinion by using words like horrible and disappointing.

2: **Comparative opinion**: When a reviewer compares the opinion of one product with another based on specific criteria, it can be classified under comparative opinions. For instance, this Amazon 2-star K-cup reviewer "*I thought being from Starbucks it would be good. however, my sister turned me onto Maud's brand and better flavor, better price for your buck.*", compares the purchased Starbucks K-cups with Maud K-cups based on flavor and price. Such comments not only give an insight into the product but also serve as micro-competitive research.

3: **Explicit opinion**: The reviews that carry detailed information about the product can be categorized as explicit reviews. This Amazon 5-star furniture review "*Perfect size for our quaint balcony space in our condo building. The vibrant chartreuse color couples perfectly with our retro robin egg-colored door. Lightweight, already assembled, and super easy to clean. The price makes this set a no-brainer (Cheaper than Target, CB2, but equal in quality). Awesome summer buy*." explicitly discusses the product's size, color, weight, maintenance, and price, thus being an explicit opinion.

4: **Implicit opinion**: Implicit opinions contrary to explicit opinions do not clearly state the opinions; instead use idioms or metaphors, which complicates the sentiment analysis process. This Amazon 3-star toy review is a perfect example of such opinions "*Smaller than described*."

Generally, sentiment analysis algorithms can be classified into a lexicon-based approach, machine learning-based approach, and deep learning-based approach. All the three approaches are discussed as follows:

1: **Lexicon-based approach**: The meaning of the word lexicon is the vocabulary of a language; hence, this approach focuses on performing sentiment analysis using a given dictionary of words. Each word is in this dictionary is labeled as positive, negative, or neutral along with its polarity, parts of speech and subjectivity classifiers, mood, and modality. One such common Python package is NLTK (Natural Language ToolKit) which consists of several such libraries.

2: **Machine learning-based approach**: Sentiment analysis using machine learning analyzes the texts for their polarity, positive or negative. The machine learning tools are trained to identify the emotion behind a text. In this process, the computer is trained to identify and read a text beyond definitions and understand things like context, sarcasm, and misapplied words.

3: **Deep learning-based approach**: Deep learning is a more intricate and detailed part of machine learning. It focuses on several types of neural networks and concepts that operate on the principle of the neurons in a brain. The datasets on which such approaches are implemented are usually enormous in size. Deep learning is implemented to create deep neural networks that solve complex binary problems, build decisions or return answers with high accuracy.

6.1 Lexicon-based Approach

The collected data is first tokenized for the lexicon-based approach, which means texts are segregated into individual words. The other two popular pre-processing steps performed include removing punctuations and stop word removal (the words that do not carry any significant meanings such as this, that, I, is, are). Figure 6.1 shows an example of such an approach.

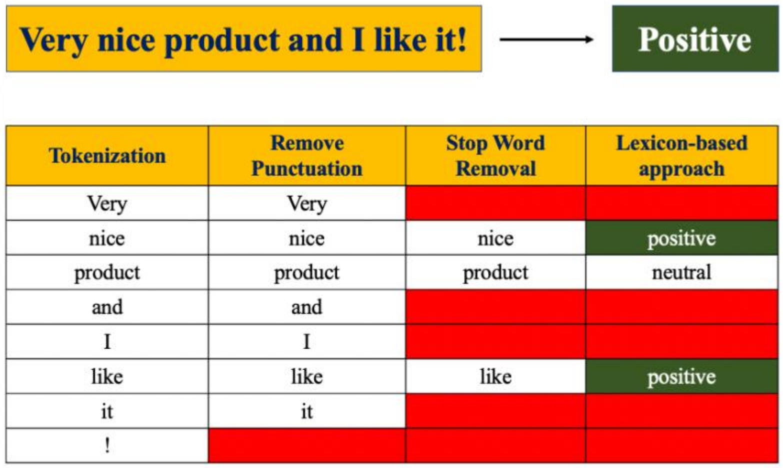


Figure 6.1 An example of steps for sentiment analysis using a lexicon-based approach.

There are several pre-trained and pre-defined lexicon approaches. Two such popular models are discussed below.

1: **VADER (Valence Aware Dictionary and Sentiment Reasoner)**: VADER is a lexicon-based sentiment analysis tool that focuses on both polarity (positive/ negative) and its intensity. It is included in the NLTK package and can be applied to the unlabelled dataset. It provides a sentiment score that maps the lexicon from the dictionary and relates it to the corresponding emotion intensities. The final sentiment score is the sum of all the corresponding scores for each word. VADER classifies words such as *'love', 'enjoy', 'happy', 'like'* as positive sentiment. On the contrary, VADER can classify *"did not love"* as a negative statement. It can also understand the emphasis of capitalization and punctuation, such as *"ENJOY!"*

```
TextBlob("Very nice product and I like it!").sentiment
Sentiment(polarity=0.9750000000000001, subjectivity=1.0)
```

Figure 6.2 Performance of VADER for the given text “Very nice product and I like it!”.

2: **TextBlob**: TextBlob is a Python library that uses NLTK to achieve its goals. It returns the polarity and subjectivity of a text. The polarity of the text lies between -1 and 1, where -1 holds a negative sentiment and 1 has a positive sentiment. Subjectivity here is defined as the level of personal opinion and general information contained in each text. Higher subjectivity indicates that the text contains more opinion and less information. TextBlob calculates the subjectivity based on the intensity of the text, where intensity determines if a word modifies the next word. For instance, in the English language, adverbs are used as modifiers “very good”.

```
sentiment_analyzer_scores("Very nice product and I like it!")
Very nice product and I like it!-------- {'neg': 0.0, 'neu': 0.459, 'pos': 0.541, 'compound
': 0.7082}
```

Figure 6.3 Performance of TextBlob for the given text “Very nice product and I like it!”.

6.2 Machine Learning-based Approach

Machine learning-based approaches for sentiment analysis usually perform better as they can tailor their decision-making to a specific type of data like tweets or reviews. However, machine learning algorithms demand much larger datasets than the sentiment lexicon-based algorithms for the model to train. One common approach is to then split each of these datasets into a training and a test dataset. Generally, 70% of the data is used to train the model, and the remaining 30% helps test the model. Several machine learning models are used for sentiment analysis, among which a few of the popular models are discussed below.

1: **Naive Bayes Classifier**: Naive Bayes classifier is a probabilistic algorithm that focuses on the naive Bayes' theorem, which assumes that the presence of a specific feature in a class is not related to the presence of any other feature. The algorithm applies the former theorem with the "naive" assumption of conditional independence between every pair of features. The following model predicts the nature of a text by calculating the probability of occurrence of each possible nature for the given text based on the trained model and then selecting the nature with maximum occurrences.

2: **Support Vector Machine**: Like Naive Bayes classifiers, a support vector machine (Kang et al., 2018; H. Zhang et al., 2020; P. Zhang et al., 2018) also requires training dataset. It uses kernel trick to transform the data, based on which the model finds an optimal boundary between possible outcomes. It classifies data into a hyperplane. Since the SVM ignores any outlier while locating the hyperplane, as shown in Figure 6.4, it can be concluded that the following model is robust to outliers.

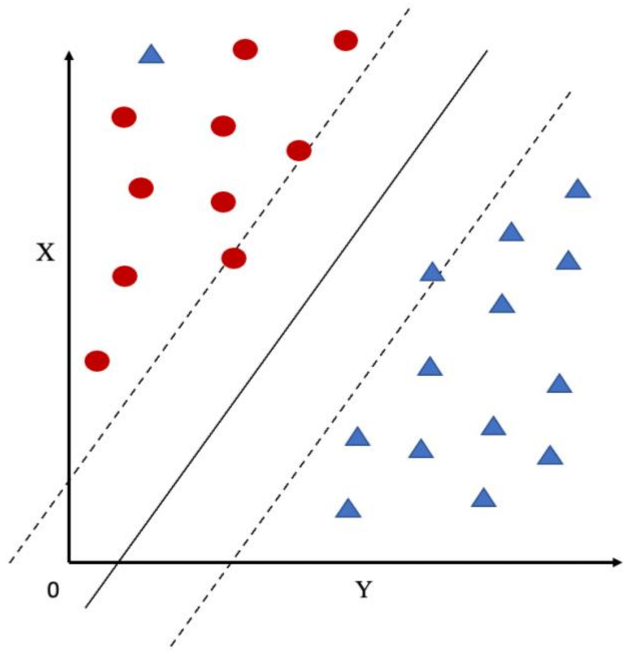


Figure 6.4 Projection of hyperplane to classify data into two classes in SVM.

3: **k-Nearest Neighbors**: The k-Nearest Neighbors (KNN) (Luo et al., 2021; L. Wang & Lu, 2019) works by finding the training examples closest to the test example. This algorithm assumes that identical things exist in direct proximity. Hence, keeping this assumption as true, KNN tries to classify the test data by finding the nearest match with the training data and then using the corresponding label to predict the test data class. This algorithm is easy to implement and simple, and it is also very versatile as it can be used for classification, regression, and to search within a dataset.

6.3 Deep Learning-based Approach

With the rapid development of neural network methods in recent years, Deep Neural Networks (DNNs) have achieved great success in many applications. ABSA's research has shifted from feature engineering methods to deep neural network methods. They can be divided into 1) Convolutional Neural Network (CNN), 2) Recurrent Neural Network (RNN), 3) Recursive Neural Network (RecNN), and 4) Memory Network (MN). In addition to the direct application of various DNNs and their variants, an attention mechanism combined with the above DNNs has become highly popular.

1: **Convolutional Neural Network**: NLP researchers have applied CNNs (Tan et al., 2021; Y. Zhang et al., 2020) to sentiment analysis, machine translation, and question answering for a long time. The input for NLP tasks is usually a sentence or a document represented by a matrix where each row represents a word, and a vector represents each word. From Figure 6.5. as an example, it can be observed that the number of words in "Good for COVID-19 but mask material is uncomfortable" is 8, and the dimension of word embedding for these eight words is chosen as 5. Hence, the input of this sentence is a matrix with dimension 8×5. It consists of a) An input layer, a matrix in which the order of word vector corresponds to the order of word in a given sentence. If a sentence has n-words and the dimension of word embeddings is k, then the matrix size is $n \times k$.; b) A convolutional layer resulting from moving a sliding window over a sentence and then applying the same convolution filter to each window in a sequence. The size of a convolution window is $h \times k$, where h is the region size of a filter matrix. After the input matrix passes the convolution layer of an $h \times k$ convolution kernel, a feature map with one column is obtained. After the convolution process is complete, a new feature ci is obtained through an activation function, e.g., tanh.;

c) A max-pooling layer method is adopted, which combines the vectors resulting from different convolution windows into a single l-dimensional vector. This method is performed by extracting the maximum value observed in the previous vector from the convolutional layer; and d) A one-dimensional vector is connected to the softmax layer for classification through the entire connection.

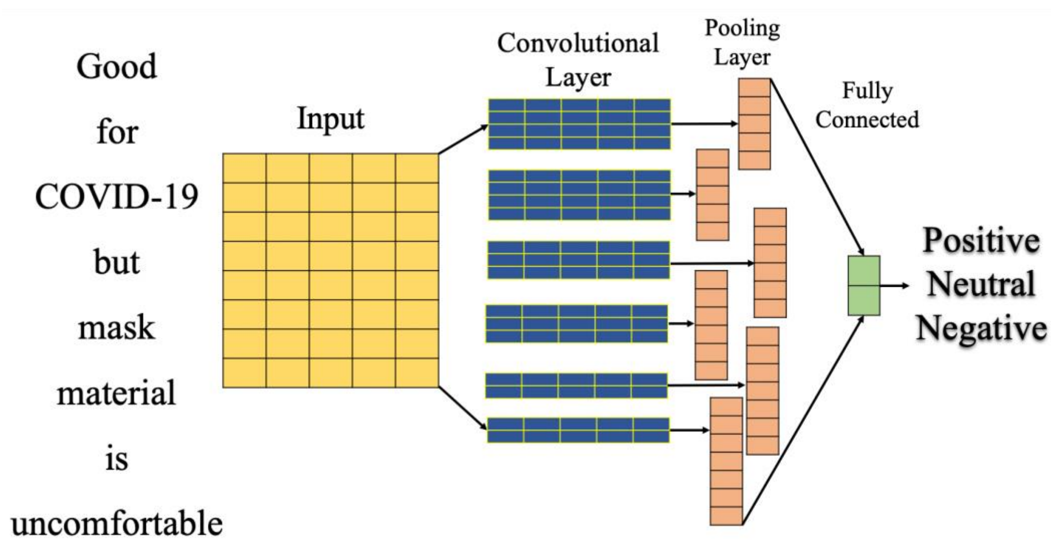


Figure 6.5 CNN model for sentence classification.

2: **Recurrent Neural Network**: RNN (S. Li et al., 2018; W. Zhang et al., 2021; Y. Zhang et al., 2020) is a very popular model that has shown great power in many NLP tasks. The main idea behind it is to use sequential information. In traditional neural networks, we assume that all inputs are independent of each other. For many tasks, this is an unrealistic assumption. If one wants to predict the next word in a sequence, it needs to know which words are in front of it. RNN performs the same operations for each element in a series, relying on its previous calculations. In theory, RNN can use arbitrarily long sequenced information, but in reality, only a few previous steps can be reviewed. Since the number of input layers in a neural network is fixed, the variable input length needs to process recurrently or recursively. The RNN realizes by dividing the variable length of input into some equal length of small pieces, which are then inputted into the network. For example, when dealing with a sentence, the sentence is treated as a sequence of words. We then input one word at a time to RNN until finishing the whole sentence. Finally, a corresponding output is produced through RNN. Figure 6.6 shows a basic RNN model. At time *t*, given *x* as input, we obtain the hidden state h_t.

3: **Recursive Neural Network**: RecNN (Nguyen et al., 2020) is similar to RNN. Its computational graph is a deep tree, but it does not have the general RNN chain structure. Unlike RNN that can handle a fixed number of input layers, RecNN does not treat a sentence as a sequence of words. RecNN encodes the information, in the shape of a tree or a graph, as a vector and maps the information into a semantic vector space. This semantic vector space satisfies some properties; for instance, vectors containing similar semantics are closer to each other in a space domain. In other words, if two sentences have the same meaning, their separately encoded vectors are close to each other. The problem of RecNN is that its structure is a tree, and its computational time is ten times more than LSTM's. Putting a tree structure over a sentence means it needs to make categorical choices. It is used to determine the words that are going to be single components. Figure 6.7 uses one example to show the different structures based on CNN, RNN, and RecNN. CNN computes vectors for every possible phrase while RNN computes vectors along with the sentence's order, and RecNN calculates compositional vectors only for grammatical phrases.

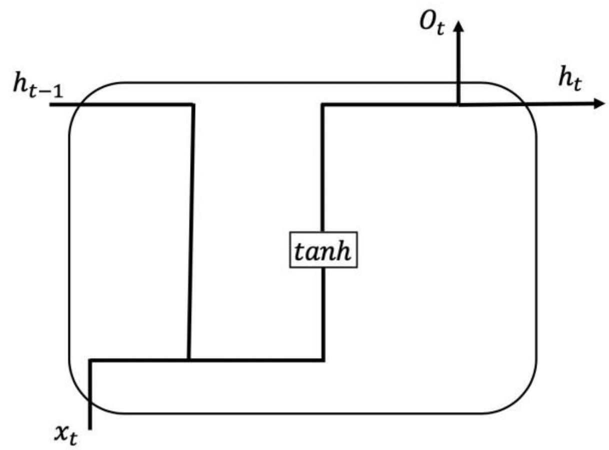
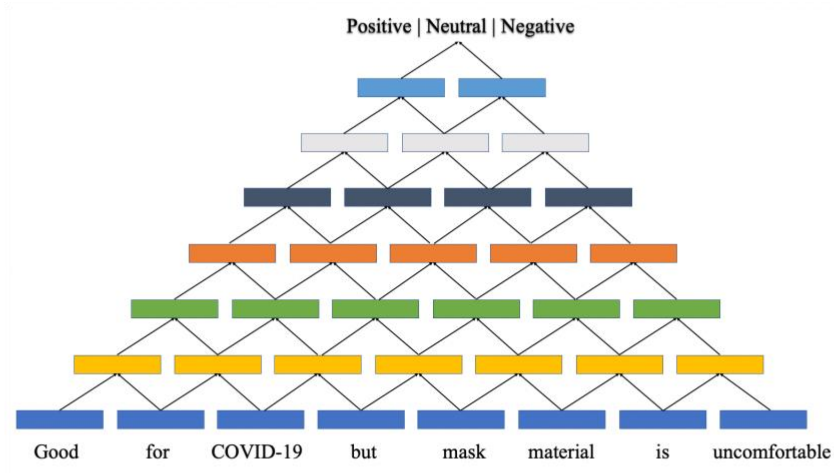
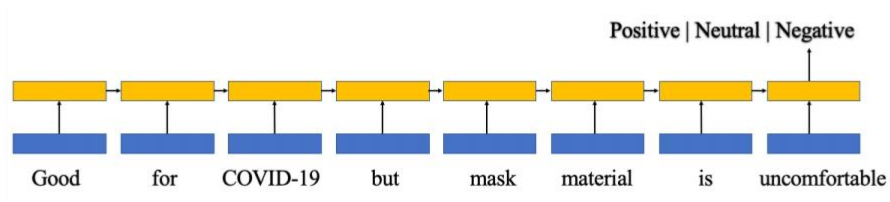


Figure 6.6 An example of a basic model for recurrent neural network RNN model.

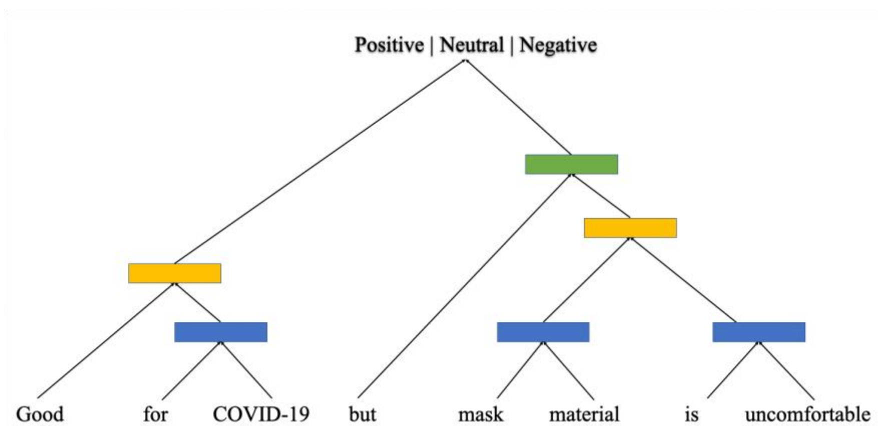
4: **Memory Network (MN)**: The basic motivation of using Memory Networks is the need for long-term memory to hold the knowledge of questions and answers or the contextual information of conversations. A traditional RNN does not perform so well in long-term memory. Popular deep learning models such as RNN, LSTM, and GRU use hidden states or an attention mechanism as their memory function, but the memory generated by them tends to be too small to record all the information expressed in a paragraph accurately. Thus, such pieces of information are lost while input is encoded as a "dense vector". Figure 6.8 shows a basic MN model, which consists of memory m and four responses $\mathbb{R}$.



(a) Convolutional Neural Network (CNN) models



(b) Recurrent Neural Network (RNN) model



(c) Recursive Neural Network (RecNN) model

Figure 6.7 A comparison example for CNN, RNN, and RecNN models.

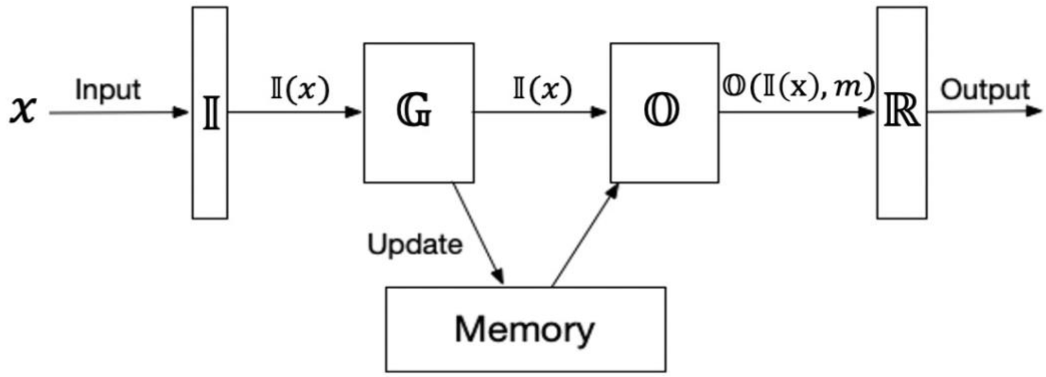


Figure 6.8 A basic example of memory network (MN) model.

CHAPTER 7

EXPERIMENTAL RESULTS

The proposed SADCM identifies and rectifies the outliers for all the seven datasets containing Amazon customer reviews of products from various domains. The algorithm used for sentiment analysis in Tables 7.1 and 7.2 is TextBlob (*Textblob*, 2020), a Python library for NLP. The experiment is performed in two stages. Initially, the algorithm is implemented to each star rating of a dataset separately to study the results. SADCM then evaluates the complete dataset at a later stage of the research.

Tables C1 – C5 in Appendix B represent the results from reviews evaluated based on the star ratings individually. For Tables C1 – C2, the least value for O_s is considered as F_U and O_s is then decremented by 0.1 till it reaches 0.8. For Table C3 – C5, the least value for O_s is considered as F_L and O_s is then incremented by 0.1 till it reaches 1.2. The results are then saved in a .csv file, evaluated manually to check the number of outliers detected correctly and incorrectly. In all the Tables, O_D represents the total number of outliers detected, O_I is the number of reviews incorrectly labeled as outliers and O_C equals the number of reviews correctly labeled as outliers. O_I and O_C are validated manually for cross-verification. SADCM is implemented for all the datasets and ratings separately.

The performance of SADCM is compared with two state-of-the-art outlier detection methods published this year: a) a class-based approach (Riahi-Madvar et al., 2021) and b) a deep-learning-based approach (Studiawan et al., 2020) and on two separate datasets a) self-scraped dataset from Amazon.com (Chatterjee, 2021), and b) publicly available Amazon customer review dataset (*Amazon Customer Reviews Dataset*, 2017). Table 7.1 represents the performance comparison of SADCM with those in (Riahi-Madvar et al., 2021) and (Studiawan et al., 2020) on all the seven collected Amazon customer reviews datasets (Chatterjee, 2021). All the three algorithms were executed on an Amazon review dataset publicly available in the amazon-reviews-pds S3 bucket in AWS US East Region (Amazon Customer Reviews Dataset, 2017), as shown in Table 7.2. These seven datasets each consists of 100,000 Amazon product reviews from various domains. For Tables 7.1 and 7.2 bold font indicates the best results obtained for each dataset. Table 7.3 compiles the metrics comparison for SADCM using p-value, T-score, and CI, where CI represents the 95% confidence interval in [x, y].

From Tables C1 – C5, it can be concluded that SADCM detects an optimal number of outliers in all the datasets and shows a perfect ratio between the correctly and incorrectly detected outliers, thus resulting in a high degree of accuracy. The accuracy decreases considerably once the value of O_s reaches one. Also, the increase or decrease in O_s for positive or negative star-rated reviews, respectively, results in a rise in incorrectly labeled outliers. It can also be concluded from Table 7.1 that the accuracy and recall percentage of SADCM for all the collected datasets outperforms those of (Riahi-Madvar et al., 2021)-(Studiawan et al., 2020). A similar conclusion can be made for Table 7.2, where the SADCM exceeds its performance for the public dataset. Hence, it is inferred that SADCM outperforms the other methods in the outlier detection and correction approach, which are outperformed by those in (Riahi-Madvar et al., 2021)-(Studiawan et al., 2020).

Table 7.1 Performance Comparison of SADCM on Collected Dataset

Dataset	Methods	Accuracy%	Recall%	O_D
Book	**SADCM**	**95.1**	**97.5**	**31**
	(Riahi-Madvar et al., 2021)	90.2	65.4	678
	(Studiawan et al., 2020)	91.3	65.5	955
Electronics	**SADCM**	**93.1**	**96.5**	**60**
	(Riahi-Madvar et al., 2021)	67.3	49.8	193
	(Studiawan et al., 2020)	71.3	48.5	638
Entertainment	**SADCM**	**87.6**	**93.8**	**23**
	(Riahi-Madvar et al., 2021)	67.7	51.8	158
	(Studiawan et al., 2020)	79.1	48.9	1434
Grocery	**SADCM**	**92.3**	**96.1**	**31**
	(Riahi-Madvar et al., 2021)	75.7	49.7	406
	(Studiawan et al., 2020)	85.8	48.1	1194
Health Care	**SADCM**	**93.1**	**96.5**	**43**
	(Riahi-Madvar et al., 2021)	74.8	51.1	99
	(Studiawan et al., 2020)	86.2	49.1	1025
Personal Care	**SADCM**	**93.3**	**96.6**	**31**
	(Riahi-Madvar et al., 2021)	76.3	50.9	717
	(Studiawan et al., 2020)	86.2	48.9	1177
Pharmaceutical	**SADCM**	**89.4**	**94.7**	**17**
	(Riahi-Madvar et al., 2021)	78.7	51.0	239
	(Studiawan et al., 2020)	77.3	47.2	971

Table 7.2 Performance Comparison of SADCM on Public Dataset

Dataset	Methods	Accuracy%	Recall%	O_D
Apparel	**SADCM**	**89.1**	**94.5**	**809**
	(Riahi-Madvar et al., 2021)	78.8	65.3	6404
	(Studiawan et al., 2020)	80.1	65.3	585
Beauty	**SADCM**	**90.4**	**95.1**	**936**
	(Riahi-Madvar et al., 2021)	81.2	65.4	9501
	(Studiawan et al., 2020)	83.1	65.5	643
Fashion	**SADCM**	**92.3**	**96.1**	**1061**
	(Riahi-Madvar et al., 2021)	81.6	62.2	3257
	(Studiawan et al., 2020)	81.4	62.1	604
Furniture	**SADCM**	**90.8**	**95.3**	**922**
	(Riahi-Madvar et al., 2021)	80.4	64.8	3743
	(Studiawan et al., 2020)	81.2	64.1	675
Jewelry	**SADCM**	**91.3**	**95.6**	**700**
	(Riahi-Madvar et al., 2021)	81.2	64.4	6345
	(Studiawan et al., 2020)	82.4	64.4	562
Luggage	**SADCM**	**92.1**	**96.2**	**831**
	(Riahi-Madvar et al., 2021)	82.1	63.6	4000
	(Studiawan et al., 2020)	83.3	63.8	599
Toy	**SADCM**	**90.2**	**95.1**	**662**
	(Riahi-Madvar et al., 2021)	83.2	65.7	9444
	(Studiawan et al., 2020)	84.1	65.2	634

Table 7.3 reflects that the p-value is less than 0.001, which is a piece of robust evidence against the null hypotheses. An extremely low p-value signifies that the results are not accidental, and the improvement is due to SADCM. The T-score for all the datasets is high, indicating more significant evidence against the null hypothesis. This means that there is a considerable difference between the collected star ratings from the website and the improved star ratings based on the nature of the reviews by SADCM. CI in Table 7.3 represents a 95% chance that the actual error of the model is between x +/- y. Hence, the smaller the CI, the more precise is the estimate of the model.

Table 7.3 Metrics Comparison for SADCM.

Dataset	**p-value**	**T-score**	**CI**
Book	3.21 e-6	14.35	[0.04, 0.06]
Electronics	1.43 e-6	16.67	[0.06, 0.08]
Entertainment	8.46 e-8	25.67	[0.11, 0.13]
Grocery	1.48 e-7	18.93	[0.07, 0.08]
Health Care	7.26 e-6	17.27	[0.06, 0.08]
Personal Care	1.08 e-6	17.38	[0.06, 0.07]
Pharmaceutical	3.62 e-9	23.63	[0.10, 0.12]

CHAPTER 8

CONCLUSION AND FUTURE WORK

Amazon is a global leader in the online marketplace, and potential buyers usually rely on their star ratings and product reviews for a particular item or service. In some cases, the sentiment a review carries contradicts the nature of its star rating. This dissertation addresses such problems with Amazon product reviews and ratings.

SADCM is a novel approach for identifying anomaly in a customer review dataset and rectifying it by improving their corresponding star rating. The results exhibit that the performance of the proposed algorithm surpasses other state-of-the-art approaches, and it also gives evidence for SADCM's rejection of the null hypothesis. The advantage of SADCM against most methods is that this data analysis pipeline preserves the outliers to correct them and prevents any information loss. It can be inferred that the outlier definition is different for positive and negative reviews as the minority in a dataset with positive star-rated reviews is when the nature of both reviews and the star rating contradicts. At the same time, the reverse is true for negative star-rated reviews.

The reviews for SADCM are classified into two classes a) positive and b) negative. The reviews with star ratings of 4 and 5 are considered as positive star-rated comments, while a review with star ratings of 1, 2, and 3 are considered as negative star-rated reviews. The reason for considering 3-star comments as negative can majorly be explained in the following four points: a) most of the time, three reasons a customer assigns 3-star to a product is when the consumer is either not fully satisfied with the product or the purchaser does not find the product to be unique or different; b) customer review datasets are generally highly imbalanced datasets which means the number of positive reviews dominates the dataset with high margin and so considering 3-star reviews as negative would slightly increase the number of the reviews in minority class; c) while going through the reviews manually it was observed that the 3-star reviews followed the same pattern as the 1 and 2-star rated comments which are the negative comments and lastly, d) there was a reduction in performance of SADCM when the 3-star ratings were considered under the positive label. Hence, the 3-star rated comments are classified under the negative label.

The performance of such outlier detection and correction algorithms depends on the nature of a dataset, the definition of outliers according to the algorithm, the way the outlier correction is handled. It is observed that SADCM outperforms both the state-of-the-art algorithms for all the datasets with higher accuracy and recall percentages. One of the reasons is that both the research work (Riahi-Madvar et al., 2021; Studiawan et al., 2021) define outliers differently and are designed for different datasets; where (Riahi-Madvar et al., 2021) uses datasets which are consisting of medical records while (Studiawan et al., 2021) performs outlier detection on OS log. Hence, both methods' accuracy is above 80%, but the recall percentage is comparatively lower. Recall indicates the percentage of total relevant results correctly classified by the algorithm, which is the outliers in this case. Hence, it can be said the percentage of correctly classified outliers is less, although the accuracy is higher.

From both collected and public dataset studies, it can be inferred that the customer review datasets are generally highly imbalanced irrespective of the product or its department; and they follow J-shaped distribution, meaning the number of positive rated comments dominate the dataset while the count of negatively rated comments are negligible.

To deal with such imbalanced data, usually, dataset manipulation is done by over-sampling, under-sampling, or by creating synthetic samples for the dataset (Han et al., 2021; Kang et al., 2017; Liu et al., 2019; X. Wang et al., 2019; Zheng et al., 2020; Zhu et al., 2020). This dissertation chooses to keep the authenticity of the dataset intact by not taking any action and modifying the proposed method to work on an imbalanced dataset

Another observation that can be deduced by studying the count of helpful votes in the datasets is that extremely negative reviews have the most significant number of helpful votes, which means that the potential buyers find such reviews as the most helpful ones for decision making. One of the reasons could be that a star rating usually signifies the rating for a product on a scale of 1 to 5, but it does not reflect the reason behind such rating. Sometimes, the reviewer likes the product, but the negative star rating is because of delay in shipment or rough handling of packages. The extremely negative reviews are usually highly critical about the product, its various features, the shipment, packaging, and even its use and applications. These reviews give an insight into a buyer's opinion about a product and share the experience a customer might have had and the reason behind such a negative star rating.

SADCM performs well on both collected and public datasets consisting of Amazon customer reviews, which means the method works well on any dataset comprising a star rating and Amazon review information. Hence, some of the future work of this dissertation can be summarized as follows:

1: **SADCM on other marketplace datasets**: There are other marketplaces that purchasers rely on apart from Amazon.com. These online retail stores are also a hub of customer feedbacks for various products. The proposed method can be applied to these product reviews from other marketplace datasets like eBay, Etsy, Best Buy, Target, Walmart, etc., to give a better insight into the discrepancies between star ratings and the related reviews. This will help conclude that SADCM can detect and rectify the anomalies without deleting any data to preserve the overall dataset knowledge.

2: **Implementation of SADCM on real-life scenarios**: This algorithm can be implemented in several real-life scenarios like accessing product performance, conducting market research, and flagging reviews through rating and review irregularity detection. This will help in solving such existing issues without any significant loss of information.

3: **Performance evaluation of SADCM with other NLP models**: In this paper, the sentiment analysis algorithm used is TextBlob, a Python-based NLP package. It should be interesting to study the behavior and impact of SADCM when combined with other state-of-the-art sentiment analysis algorithms like BERT, XLNet, ELECTRA, OpenAI's GPT-3, RoBERTa, or StructBERT.

4: **Application of SADCM to other areas**: The increasing use of Internet of Things and advances of Industry 4.0 (Fortino et al., 2018; Mahmud et al., 2019; White et al., 2020; N. Wu et al., 2018; Yang et al., 2021) have generated more and more big data. Yet outliers often exist in such data. The application of

SADCM to different smart industrial systems (Chen & Luo, 2018; X. Hu et al., 2018; Huang et al., 2020; F. Wang et al., 2017; Zahid et al., 2020) should be sought.

APPENDIX A

SADCM PSEUDOCODE

This Appendix projects pseudocode of the proposed algorithm Statistics-based Anomaly Detection and Correction SADCM.

Algorithm: Statistics-based Anomaly Detection and correction method (SADCM)

Input: D // dataset containing r_i and r_i^*

Output: D^*// modified dataset post outlier detection and correction

Step 1:

if $r_i^* \geq 4$ **then**

    $T_V = 1;$

else

    $T_V = -1;$

end if

Step 2:

INITIALIZE V_D to array [0];

for each r_i **do**

    $V_D[i] = d^E(T_V, C_V);$

end for

Step 3:

INITIALIZE $S^+$to array [0];

INITIALIZE $S^-$to array [0];

for each r_i^* **do**

  if $r_i^* \geq 4$ **then**

    $S^+[i] = [r_i, r_i^*, V_D[i]]$;

  else

    $S^-[i] = [r_i, r_i^*, V_D[i]]$;

  end if

end for

Step 4:

Function I_{QR} calculation (S, V_D)

  Sort (V_D);

  Let $Q_1 =$ first quartile (V_D);

  Let $Q_3 =$ third quartile (V_D); $I_{QR} = Q_3 - Q_1$;

  $F_L = Q_1 - 1.5I_{QR}$;

  $F_U = Q_3 + 1.5I_{QR}$;

  if $S \geq S^+$ **then**

    if $F_L < 0$ **then**

      $O_s = Q_3 + I_{QR}$;

    else

      $O_s = F_U - I_{QR}F_L$;

    end if

  Else

    if $F_U > 2$ **then**

      $O_s = Q_1 - I_{QR}$;

        else

            $O_s = Q_3 - I_{QR}F_U$;

        end if

    end if

    return O_s;

end Function

$O_{S+} =$ calculation (S^+, V_D);

$O_{S-} =$ calculation (S^-, V_D);

Step 5:

INITIALIZE $O^+$to array [0];

INITIALIZE $O^-$to array [0];

for each r_i in S^+ **do**

    if $V_D(r_i) \geq O_{S+}$ **then**

        $O^+[i] =$ ‘yes’;

    else

        $O^+[i] =$ ‘no’;

    end if

end for

for each r_i in S^- **do**

    if $V_D(r_i) \leq O_{S-}$ **then**

        $O^-[i] =$ ‘yes’;

    else

        $O^-[i] =$ ‘no’;

end if

end for

Step 6:

for each r_i in S^+ **do**

if $O^+[i] =$ 'yes' **then**

$T_V[i] = \text{toggle}(T_V[i]);$

end if

end for

for each r_i in S^- **do**

if $O^-[i] =$ 'no' **then**

$T_V[i] = \text{toggle}(T_V[i]);$

end if

end for

$D^* = \text{concat}\,(S^+, S^-);$

APPENDIX B

THEOREMS FOR SADCM

This Appendix states the theorems required for SADCM.

Theorem 1: $r_i \in S^+$ if $V_D(r_i) \geq O_s$, where r_i is outlier. For S^+, customer comments whose V_D is greater than or equal to the calculated outlier score (O_s) are outliers.

Proof: By contraposition, it is equivalent to prove "$r_i$is not an outlier if $V_D(r_i) < O_s$".

By method of mathematical induction, let's, consider two datasets: D_1={0.6, 0.7, 0.8, 0.9, 1.0, 1.8} and D_2={0.1, 0.11, 0.15, 0.2, 0.21, 0.8}.

CASE 1:

For D_1={0.6, 0.7, 0.8, 0.9, 1.0, 1.8}

$N = 6$

$Q_1 = ((N + 1) * \frac{1}{4})^{th}$ term

$= ((6 + 1) * \frac{1}{4})^{th}$ term

$\cong 2^{nd}$ term

$= 0.7$

Similarly,

$Q_3 = ((N + 1) * \frac{3}{4})^{th}$ term

$= ((6 + 1) * \frac{3}{4})^{th}$ term

$\cong 5^{th}$ term

$= 1.0$

I_{QR}, F_L, and F_U can be calculated as:

$I_{QR} = Q_3 - Q_1$

$= 1.0 - 0.7$

$= 0.3$

$F_L = Q_1 - 1.5 I_{QR}$

$= 0.7 - 1.5(0.3)$

$= 0.25$

$F_U = Q_3 + 1.5 I_{QR}$

$= 1.0 + 1.5(0.3)$

$= 1.45$

Since, $F_L \geq 0$, O_S= $F_U - I_{QR} F_L$

On substituting values O_S=1.375

Any value less than 1.375 is not an outlier because the values fall between the I_{QR} range.

CASE 2:

For D_2={0.1, 0.11, 0.15, 0.2, 0.21, 0.8}

$N = 6$

$Q_1 = ((N + 1) * \frac{1}{4})^{th}$ term

$= ((6 + 1) * \frac{1}{4})^{th}$ term

$\cong 2^{nd}$ term

$= 0.11$

Similarly,

$Q_3 = ((N+1) * \frac{3}{4})^{th}$ term

$= ((6+1) * \frac{3}{4})^{th}$ term

$\cong 5^{th}$ term

$= 0.21$

I_{QR}, F_L, and F_U can be calculated as:

$I_{QR} = Q_3 - Q_1$

$= 0.21 - 0.11$

$= 0.10$

$F_L = Q_1 - 1.5 I_{QR}$

$= 0.11 - 1.5(0.1)$

$= -0.04$

$F_U = Q_3 + 1.5 I_{QR}$

$= 0.21 + 1.5(0.1)$

$= 0.36$

Since, $F_L < 0$, O_s=$Q_3 + I_{QR}$

On substituting values O_s=1.31

Any value less than 1.31 is not an outlier because the values fall between the I_{QR} range.

By showing that when $r_i$is not an outlier if $V_D(r_i) < O_s$, we have therefore proven that for customer comments whose V_D is greater than or equal to the calculated outlier score (O_s) are outliers, by proof of contrapositive.□

Theorem 2: $r_i \in S^-$ if $V_D(r_i) \leq O_s$, where r_i is outlier. For S^-, customer comments whose V_D is less than or equal to the calculated outlier score (O_s) are outliers.

Proof: By contraposition, it is equivalent to prove "$r_i$is not an outlier if $V_D(r_i) > O_s$"

By method of mathematical induction, let's, consider two datasets: D_1={0.2, 1.7, 1.8, 1.9, 1.95, 2.0} and D_2={0.2, 1.0, 1.11, 1.15, 1.2, 1.25}

CASE 1:

For D_1={0.2, 1.7, 1.8, 1.9, 1.95, 2.0}

$N = 6$

$Q_1 = ((N + 1) * \frac{1}{4})^{th}$ term

$= ((6 + 1) * \frac{1}{4})^{th}$ term

$\cong 2^{nd}$ term

$= 1.7$

Similarly,

$Q_3 = ((N + 1) * \frac{3}{4})^{th}$ term

$= ((6 + 1) * \frac{3}{4})^{th}$ term

$\cong 5^{th}$ term

$= 1.95$

I_{QR}, F_L, and F_U can be calculated as:

$I_{QR} = Q_3 - Q_1$

$= 1.95 - 1.7$

$= 0.25$

$F_L = Q_1 - 1.5I_{QR}$

$= 1.7 - 1.5(0.25)$

$= 1.325$

$F_U = Q_3 + 1.5I_{QR}$

$= 1.95 + 1.5(0.25)$

$= 2.325$

Since, $F_U \geq 2$, O_s=$Q_1 - I_{QR}$

On substituting values O_s=1.45

Any value greater than 1.45 is not an outlier because the values fall between the I_{QR} range.

CASE 2:

For D_2={0.2, 1.0, 1.11, 1.15, 1.2, 1.25}

$N = 6$

$Q_1 = ((N + 1) * \frac{1}{4})^{th}$ term

$= ((6 + 1) * \frac{1}{4})^{th}$ term

$\cong 2^{nd}$ term

$= 1.0$

Similarly,

$Q_3 = ((N + 1) * \frac{3}{4})^{th}$ term

$= ((6 + 1) * \frac{3}{4})^{th}$ term

$\cong 5^{th}$ term

$= 1.2$

I_{QR}, F_L, and F_U can be calculated as:

$I_{QR} = Q_3 - Q_1$

$= 1.2 - 1.0$

$= 0.2$

$F_L = Q_1 - 1.5I_{QR}$

$= 1.0 - 1.5(0.2)$

$= 0.7$

$F_U = Q_3 + 1.5I_{QR}$

$= 1.2 + 1.5(0.2)$

$= 1.5$

Since, $F_U < 2$, $Q_3 - I_{QR}F_U$

On substituting values O_s=0.9

Any value greater than 0.9 is not an outlier because the values fall between the I_{QR} range.

By showing that when $r_i$is not an outlier if $V_D(r_i) > O_s$, we have therefore proven that for customer comments whose V_D is less than or equal to the calculated outlier score (O_s) are outliers, by proof of contrapositive. □

Theorem 3: The time complexity of SADCM is $O(n)$.

Proof: Each of Steps 1, 2, 3, 5, and 6 in SADCM requires time complexity $\mathrm{O}(n)$ while Step 4 needs $\mathrm{O}(1)$. Hence, the entire algorithm has the complexity $\mathrm{O}(n)$. □

APPENDIX C

SUPPORTING EXPERIMENTAL RESULTS

This Appendix projects all the experimental tables to support the results of this work. All the experiments were conducted in Python 3.7 on a Jupyter notebook. The models were tested locally on an Apple M1 chip, 8GB of RAM, and 512GB SSD. Several Python libraries were used including NLTK 3.5, pandas 1.2.0, matplotlib 3.3.3, TextBlob 0.15.3, scikit-learn 0.19.0, NumPy 1.19.5, scipy 1.7.1, and pingouin 0.4.0. For Tables C1 – C2, the value of O_s ranges between F_U and 0.8 with a gradual decrement in steps of 0.1. For Tables C3 – C5, the value of O_s ranges between F_L and 1.2 with a gradual increment in steps of 0.1. The results, saved in a .csv file, is manually evaluated twice by two different analysts to determine the number of outliers detected correctly and incorrectly. In all the Tables, O_D represents the total number of outliers detected, O_I is the number of reviews incorrectly labeled as outliers and O_C equals the number of reviews correctly labeled as outliers. O_I and O_C are validated manually for cross-verification.

From Tables C1 – C5, it can be observed an optimal number of outliers is successfully detected in all the datasets by the proposed SADCM. This leads to a high accuracy since the number of correctly and incorrectly detected outliers reaches a perfect balance. As the value of O_s reaches 1, the sentiment analysis accuracy of the modified dataset decreases considerably and the increase and decrease of O_s for positive and negative star-rated reviews, respectively, results in a rise in incorrectly labeled outliers.

Table C.1 SADCM on 5-Star Review Comments

Dataset	o_s	o_D	o_I	o_C	Accuracy
Book	1.181	55	9	46	0.973
	1.141	**74**	**17**	**57**	**0.978**
	1.1	92	26	66	0.982
	1.0	297	87	210	0.967
	0.9	485	109	376	0.922
	0.8	1005	129	826	0.795
Electronics	1.184	35	1	34	0.929
	1.105	**75**	**5**	**70**	**0.946**
	1.1	86	8	78	0.951
	1.0	231	154	231	0.991
	0.9	573	298	275	0.853
	0.8	1178	435	743	0.61
Entertainment	1.747	15	1	14	0.886
	1.7	17	2	15	0.887
	1.6	26	4	22	0.89
	1.5	40	9	31	0.894
	1.478	**43**	**9**	**34**	**0.895**
	1.4	65	13	52	0.902
	1.3	94	15	79	0.911
	1.2	147	21	126	0.926
	1.1	257	46	211	0.96
	1.0	705	351	354	0.903
	0.9	937	454	483	0.832
	0.8	1224	590	634	0.745
Grocery	1.6	25	1	24	0.924
	1.5	32	2	30	0.926
	1.4	45	4	41	0.93
	1.355	**54**	**4**	**50**	**0.933**
	1.3	71	10	61	0.937
	1.2	102	17	85	0.947
	1.1	162	35	127	0.964
	1.0	604	245	359	0.905
	0.9	774	301	473	0.855
	0.8	1061	354	707	0.771
Health Care	1.365	26	3	23	0.938
	1.345	**28**	**3**	**25**	**0.939**
	1.3	33	7	26	0.941
	1.2	76	12	64	0.954
	1.1	119	17	102	0.969
	1.0	400	183	217	0.937
	0.9	672	291	381	0.847
	0.8	1075	408	667	0.713
Personal Care	1.687	17	1	16	0.934
	1.6	21	1	20	0.935
	1.5	45	3	42	0.942
	1.425	**50**	**3**	**47**	**0.945**
	1.4	60	9	51	0.947
	1.3	78	9	69	0.953
	1.2	100	16	84	0.96
	1.1	161	43	118	0.979
	1.0	671	233	438	0.861
	0.9	801	295	506	0.82
	0.8	1039	375	664	0.745
Pharmaceutical	1.75	13	1	12	0.896
	1.7	14	1	13	0.897
	1.6	28	4	14	0.901
	1.5	**38**	**7**	**31**	**0.903**
	1.4	48	15	33	0.906
	1.3	77	28	49	0.914
	1.2	130	57	73	0.929
	1.1	242	114	128	0.961
	1.0	1203	146	1057	0.769
	0.9	1459	177	1282	0.697
	0.8	1744	207	1616	0.595

Table C.2 SADCM on 4-Star Review Comments

Dataset	O_s	O_D	O_I	O_C	Accuracy
Book	1.174	5	1	4	0.981
	1.138	**6**	**1**	**5**	**0.986**
	1.1	6	1	5	0.986
	1.0	14	6	8	0.977
	0.9	26	8	18	0.922
	0.8	49	12	37	0.817
Electronics	1.194	9	2	7	0.928
	1.114	**21**	**2**	**19**	**0.939**
	1.1	23	4	19	0.941
	1.0	94	27	67	0.991
	0.9	260	122	138	0.835
	0.8	522	205	317	0.588
Entertainment	1.587	1	0	1	0.872
	1.5	4	0	5	0.881
	1.4	5	0	5	0.884
	1.365	**5**	**0**	**5**	**0.884**
	1.3	13	0	13	0.908
	1.2	19	0	19	0.926
	1.1	30	1	29	0.958
	1.0	73	18	55	0.914
	0.9	95	33	62	0.849
	0.8	127	52	75	0.754
Grocery	1.568	4	0	4	0.918
	1.5	7	0	7	0.922
	1.4	12	1	11	0.929
	1.326	**14**	**1**	**13**	**0.932**
	1.3	15	1	14	0.934
	1.2	22	6	16	0.944
	1.1	39	10	29	0.969
	1.0	115	62	53	0.919
	0.9	148	69	79	0.871
	0.8	198	82	116	0.797
Health Care	1.312	9	1	8	0.927
	1.269	**13**	**1**	**12**	**0.931**
	1.2	21	2	19	0.94
	1.1	44	9	35	0.965
	1.0	133	54	79	0.936
	0.9	214	93	121	0.847
	0.8	361	125	236	0.685
Personal Care	1.69	7	0	7	0.943
	1.6	7	0	7	0.943
	1.5	13	2	11	0.95
	1.429	**14**	**2**	**12**	**0.952**
	1.4	15	2	13	0.953
	1.3	18	3	15	0.956
	1.2	25	4	21	0.964
	1.1	41	6	35	0.983
	1.0	184	123	61	0.849
	0.9	213	147	66	0.815
	0.8	264	187	77	0.755
Pharmaceutical	1.75	1	0	1	0.894
	1.7	2	0	2	0.896
	1.6	4	1	3	0.898
	1.5	**5**	**1**	**4**	**0.9**
	1.4	8	3	5	0.903
	1.3	13	4	9	0.91
	1.2	26	9	17	0.927
	1.1	44	17	27	0.951
	1.0	243	139	104	0.79
	0.9	290	153	137	0.729
	0.8	383	161	222	0.609

Table C.3 SADCM on 3-Star Review Comments

Dataset	O_S	O_D	O_I	O_C	Accuracy
Book	0.834	2	1	1	0.951
	0.9	3	1	2	0.967
	1.0	6	1	5	0.983
	1.006	**6**	**1**	**5**	**0.983**
	1.1	9	4	5	0.935
	1.2	16	10	6	0.822
Electronics	0.817	8	1	7	0.933
	0.9	24	6	18	0.953
	0.993	**59**	**8**	**51**	**0.997**
	1.0	65	11	54	0.994
	1.1	179	79	100	0.851
	1.2	376	210	166	0.604
Entertainment	0.279	4	0	4	0.87
	0.3	4	0	4	0.87
	0.4	5	0	5	0.872
	0.5	8	0	8	0.876
	0.530	**9**	**0**	**9**	**0.878**
	0.6	14	1	13	0.886
	0.7	16	1	15	0.889
	0.8	29	3	26	0.909
	0.9	46	7	39	0.936
	1.0	146	65	81	0.907
	1.1	210	90	120	0.808
	1.2	279	143	136	0.7
Grocery	0.572	1	1	0	0.902
	0.6	1	1	0	0.902
	0.7	2	1	1	0.91
	0.771	**4**	**1**	**3**	**0.925**
	0.8	5	2	3	0.932
	0.9	10	4	6	0.97
	1.0	21	12	9	0.947
	1.1	29	18	11	0.888
	1.2	40	26	14	0.805
Health Care	0.615	4	1	3	0.889
	0.7	5	1	4	0.893
	0.8	9	**1**	**8**	0.908
	0.812	**9**	**1**	**8**	**0.908**
	0.9	18	7	11	0.942
	1.0	51	24	27	0.931
	1.1	68	32	36	0.866
	1.2	113	66	47	0.695
Personal Care	0.241	2	0	2	0.902
	0.3	2	0	2	0.902
	0.4	2	0	2	0.902
	0.5	3	0	3	0.907
	0.53	**3**	**0**	**3**	**0.907**
	0.6	7	1	6	0.929
	0.7	9	2	7	0.94
	0.8	12	2	10	0.956
	0.9	16	5	11	0.978
	1.0	40	18	22	0.891
	1.1	46	22	24	0.858
	1.2	59	33	26	0.788
Pharmaceutical	0.482	1	0	1	0.893
	0.5	1	0	1	0.893
	0.6	1	0	1	0.893
	0.7	1	0	1	0.893
	0.701	**1**	**0**	**1**	**0.893**
	0.8	2	1	1	0.904
	0.9	8	3	5	0.989
	1.0	37	25	12	0.723
	1.1	44	31	13	0.648
	1.2	62	46	16	0.457

Table C.4 SADCM on 2-Star Review Comments

Dataset	O_s	O_D	O_I	O_C	Accuracy
Book	0.793	1	0	1	0.978
	0.9	2	0	2	1
	0.9706	**2**	**0**	**2**	**1**
	1.0	4	1	3	0.956
	1.1	7	3	4	0.891
	1.2	17	10	7	0.673
Electronics	0.827	5	1	4	0.933
	0.9	11	3	8	0.955
	1.0	21	5	16	0.992
	1.001	**25**	**6**	**19**	**0.993**
	1.1	61	22	39	0.859
	1.2	124	65	59	0.627
Entertainment	0.346	1	0	1	0.82
	0.4	1	0	1	0.82
	0.5	1	0	1	0.82
	0.566	2	0	2	0.825
	0.6	2	0	2	0.825
	0.7	2	0	2	0.825
	0.8	8	1	7	0.855
	0.9	23	3	20	0.93
	1.0	50	13	37	0.935
	1.1	68	18	50	0.845
	1.2	95	31	64	0.71
Grocery	0.342	1	0	1	0.894
	0.4	2	1	1	0.903
	0.5	2	1	1	0.903
	0.585	**2**	**1**	**1**	**0.903**
	0.6	3	2	1	0.913
	0.7	4	2	2	0.923
	0.8	7	2	5	0.951
	0.9	8	2	6	0.961
	1.0	22	7	15	0.903
	1.1	29	12	17	0.836
	1.2	38	17	21	0.75
Health Care	0.599	1	0	1	0.887
	0.7	2	0	2	0.892
	0.798	**6**	**0**	**6**	**0.913**
	0.8	7	1	6	0.918
	0.9	13	4	9	0.948
	1.0	41	20	21	0.908
	1.1	58	29	29	0.821
	1.2	91	55	36	0.653
Personal Care	0.512	1	0	1	0.947
	0.6	1	0	1	0.947
	0.7	1	0	1	0.947
	0.734	**1**	**0**	**1**	**0.947**
	0.8	1	0	1	0.947
	0.9	4	0	4	0.973
	1.0	21	9	12	0.877
	1.1	24	10	14	0.85
	1.2	32	16	16	0.78
Pharmaceutical	0.25	0	0	0	0.862
	0.3	0	0	0	0.862
	0.4	**0**	**0**	**0**	**0.862**
	0.5	0	0	0	0.862
	0.6	0	0	0	0.862
	0.7	0	0	0	0.862
	0.8	0	0	0	0.862
	0.9	6	5	1	0.980
	1.0	18	15	3	0.784
	1.1	22	19	3	0.705
	1.2	23	20	3	0.588

Table C.5 SADCM on 1-Star Review Comments

Dataset	O_S	O_D	O_I	O_C	Accuracy
Book	0.706	10	2	8	0.931
	0.8	24	4	20	0.956
	0.9	32	5	27	0.97
	0.914	**33**	**5**	**28**	**0.972**
	1.0	72	25	47	0.959
	1.1	110	55	55	0.892
	1.2	190	123	67	0.752
Electronics	0.809	2	1	1	0.929
	0.9	13	1	12	0.958
	0.987	**25**	**5**	**20**	**0.989**
	1.0	31	6	25	0.994
	1.1	89	30	59	0.844
	1.2	202	106	96	0.55
Entertainment	0.309	2	0	2	0.838
	0.4	3	1	2	0.84
	0.5	5	1	4	0.846
	0.545	**6**	**1**	**5**	**0.851**
	0.6	10	1	9	0.853
	0.7	16	1	15	0.865
	0.8	25	2	23	0.882
	0.9	48	7	41	0.925
	1.0	127	47	80	0.926
	1.1	185	64	121	0.818
	1.2	262	108	154	0.674
Grocery	0.357	2	0	2	0.933
	0.4	2	0	2	0.933
	0.5	6	3	3	0.939
	0.6	7	3	4	0.94
	0.606	**8**	**3**	**5**	**0.942**
	0.7	9	5	4	0.943
	0.8	13	6	7	0.949
	0.9	32	11	21	0.977
	1.0	121	77	44	0.89
	1.1	168	110	58	0.821
	1.2	229	153	76	0.731
Health Care	0.608	5	2	3	0.933
	0.7	8	3	5	0.938
	0.795	**14**	**4**	**10**	**0.948**
	0.8	14	4	10	0.948
	0.9	28	8	20	0.97
	1.0	92	46	46	0.925
	1.1	147	77	70	0.836
	1.2	244	148	96	0.679
Personal Care	0.312	5	0	5	0.922
	0.4	7	0	7	0.925
	0.5	11	0	11	0.931
	0.575	**12**	**0**	**12**	**0.933**
	0.6	13	1	12	0.933
	0.7	17	3	14	0.94
	0.8	24	3	21	0.951
	0.9	37	8	29	0.971
	1.0	137	86	51	0.876
	1.1	167	105	62	0.831
	1.2	231	156	75	0.733
Pharmaceutical	0.35	2	0	2	0.874
	0.4	2	0	2	0.874
	0.5	3	1	2	0.876
	0.566	**3**	**1**	**2**	**0.876**
	0.6	5	3	2	0.88
	0.7	11	6	5	0.891
	0.8	22	16	6	0.913
	0.9	42	31	11	0.951
	1.0	166	135	31	0.808
	1.1	211	172	39	0.722
	1.2	280	236	44	0.588

APPENDIX D

TEXTBLOB

TextBlob is a python library for Natural Language Processing (NLP). It actively uses Natural Language ToolKit (NLTK) to accomplish its tasks. NLTK is a library that provides access to many lexical resources, allowing users to fulfill tasks such as categorization, classification, and many more. TextBlob is simple words is a library that supports complex analysis and operations on textual data.

For the lexicon-based sentiment analysis approach, the nature of a text is defined by its semantic orientation and the intensity a word carries in the sentence. This requires a pre-defined dictionary that classifies a word as negative and positive. Generally, a text message will be represented by a bag of words. After assigning individual scores to all the words, the final sentiment is calculated by some pooling operation like taking an average of all the sentiments.

TextBlob returns the polarity and subjectivity of a sentence. Polarity value lies between [-1,1], -1 defines a negative sentiment while 1 defines a positive sentiment, and the polarity of negation words is usually reversed. TextBlob consists of semantic labels that help with fine-grained sentiment analysis, even for emoticons, exclamation marks, emojis, etc. The range of subjectivity value lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher the subjectivity, the more personal opinion content in a text rather than factual information. TextBlob has one additional parameter, which is intensity. TextBlob calculates subjectivity by looking at the ‘intensity’. Intensity determines if a word modifies the next word. For instance, in English, adverbs are used as modifiers such as ‘very good’. The polarity and subjectivity calculation for the following sentence, “I do not like this example at all, it is too boring”, is -1for polarity as the sentence holds a negative nature and subjectivity is 1 as the sentence carries more personal opinion than factual information.

For the sentence “This was a helpful example, but I would prefer another one”. It returns 0.0 for both subjectivity and polarity , which is not the most satisfactory answer we’d expect. It is likely that if the library returns exactly 0.0, your sentence didn’t contain any words that had a polarity in the NLTK training set or because TextBlob uses a weighted average sentiment score over all the words in each sample. This readily diffuses out the effect of sentences with widely varying polarities between words in our case. ‘helpful’ and ‘but’.

Installing and executing TextBlob is very simple. Figure D.1 shows all the steps to install TextBlob and also performs sentiment analysis using TextBlob on this 5-star Amazon review of Gucci sunglasses “If I could do the head roll and finger-popping online, it would indicate that I LOVE my new glasses!! They look good on my face and just exude class and (money) even though I’m senior poor! But in all sincerity, I have a fetish for sunglasses and am very pleased with my purchase”. The results show that the polarity and subjectivity score of all the three sentences in the review individually. Since the polarity score for all three sentences is positive, it can be concluded that the nature of sentiment the following customer opinion carries is positive. Also, it can be noticed that the subjectivity value reflected for the three sentences is above 0.5 and towards 1, which means the sentences contain information but majorly consist of personal opinions about the product.

```
#STEP 1: Install TextBlob to the machine
!pip install -U textblob

#STEP 2: Install corpora or dictionary
python -m textblob.download_corpora

#The setup is complete

#Import TextBlob to the code
from textblob import TextBlob

#For example we use an Amazon 5-star customer review of Gucci sunglasses
review = "If I could do the head roll and finger popping on line, it would indicate that

blob = TextBlob(review)
for sentence in blob.sentences:
    print(sentence.sentiment)
```

```
Sentiment(polarity=0.3565340909090909, subjectivity=0.5272727272727272)
Sentiment(polarity=0.09999999999999998, subjectivity=0.6000000000000001)
Sentiment(polarity=0.65, subjectivity=1.0)
```

Figure D.1 Basic steps to setup TextBlob and execute sentiment analysis on an Amazon customer review.

REFERENCES

A. Benlahbib and E. H. Nfaoui, “Aggregating customer review attributes for online reputation generation,” *IEEE Access*, vol. 8, pp. 96550–96564, 2020.

A. Hussain and M. Aleem, “GOCJ: Google cloud jobs dataset for distributed and cloud computing infrastructures,” *Data*, vol. 3, no. 4, pp. 38, 2018.

A. Sapegin and C. Meinel, “K-metamodes: Frequency-and ensemble-based distributed K-modes clustering for security analytics,” *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020.

A. White, A. Karimoddini, and M. Karimadini, “Resilient fault diagnosis under imperfect observations - a need for industry 4.0 era,” *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 5, pp. 1279–1288, 2020.

A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, “Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches,” *IEEE 23rd International Conference on Computational Science and Engineering (CSE)*, 2020.

Amazon customer reviews dataset. [Online]. Available: https://s3.amazonaws.com/amazon-reviews-pds/readme.html. [Accessed: 20-Nov-2021].

“Anomaly detection,” Wikipedia, 17-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Anomaly_detection. [Accessed: 20-Nov-2021].

“API reference¶,” API Reference - TextBlob 0.16.0 documentation. [Online]. Available: https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob. [Accessed: 20-Nov-2021].

B. C. Neagu, G. Grigoras, and F. Scarlatache, “Outliers discovery from smart meters data using a statistical based data mining approach,” *10th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, 2017.

B. Pang and L. Lee, “A sentimental education,” *42nd Annual Meeting on Association for Computational Linguistics*, 2004.

B. Zhang, D. Xu, H. Zhang, and M. Li, “STCs lexicon: Spectral-clustering-based topic-specific Chinese sentiment lexicon construction for social networks,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1180–1189, 2019.

B. Zhang, X. Li, X. Xu, K.-C. Leung, Z. Chen, and Y. Ye, “Knowledge guided capsule attention network for aspect-based sentiment analysis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2538–2551, 2020.

C. A. Iglesias and A. Moreno, “Sentiment analysis for social media,” *Applied Sciences*, vol. 9, no. 23, pp. 5037, 2019.

C. Shofiya and S. Abidi, “Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, pp. 5993, 2021.

C. V. García-Mendoza, O. J. Gambino, M. G. Villarreal-Cervantes, and H. Calvo, “Evolutionary optimization of ensemble learning to determine sentiment polarity in an unbalanced multiclass corpus,” *Entropy*, vol. 22, no. 9, pp. 1020, 2020.

D. Kumar Gupta, K. Srikanth Reddy, Shweta, and A. Ekbal, “PSO-ASENT: Feature selection using particle swarm optimization for aspect based sentiment analysis,” *Natural Language Processing and Information Systems*, pp. 220–233, 2015.

Diana Kaemingk // October 30, “20 online review stats to know in 2019,” Qualtrics, 17-Aug-2021. [Online]. Available: https://www.qualtrics.com/blog/online-review-stats/. [Accessed: 20-Nov-2021].

F. Harrag, A. Alsalman, and A. Alqahtani, “Prediction of reviews rating: A survey of methods, techniques and hybrid architectures,” *Journal of Digital Information Management*, vol. 17, no. 3, pp. 164, 2019.

F. T. Saputra, S. H. Wijaya, Y. Nurhadryani, and Defina, “Lexicon addition effect on lexicon-based of Indonesian sentiment analysis on Twitter,” *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2020.

F. Wang, T. Xu, T. Tang, M. C. Zhou, and H. Wang, “Bilevel feature extraction-based text mining for fault diagnosis of Railway Systems,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 49–58, 2017.

G. de la Torre-Abaitua, L. F. Lago-Fernández, and D. Arroyo, “A compression-based method for detecting anomalies in textual data,” *Entropy*, vol. 23, no. 5, pp. 618, 2021.

G. Fortino, W. Russo, C. Savaglio, W. Shen, and M. Zhou, “Agent-oriented cooperative smart objects: From IOT system design to implementation,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 11, pp. 1939–1956, 2018.

G. Guibon, P. Bellot, and M. Ochs, “From emojis to sentiment analysis,” in *WACAI, Lab-STICC; ENIB; LITIS*, 2016.

G. Li, Q. S. Zheng, L. Zhang, S. Z. Guo, and L. Y. Niu, “Sentiment infomation based model for Chinese text sentiment analysis,” *IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, 2020.

G. Saranya, G. Geetha, C. K, M. K, and S. Karpagaselvi, “Sentiment analysis of healthcare tweets using SVM classifier,” *International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, 2020.

H. Du, Q. Ye, Z. Sun, C. Liu, and W. Xu, “Fast-ODT: A lightweight outlier detection scheme for categorical data sets,” *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 13–24, 2021.

H. Fan, W. Du, A. Dahou, A. A. Ewees, D. Yousri, M. A. Elaziz, A. H. Elsheikh, L. Abualigah, and M. A. Al-qaness, “Social media toxicity classification using deep learning: Real-world application UK Brexit,” *Electronics*, vol. 10, no. 11, pp. 1332, 2021.

H. Lakkaraju, R. Socher, and C. Manning, “Aspect specific sentiment analysis using hierarchical deep learning,” *NIPS Workshop on deep learning and representation learning*, pp. 1–9, Dec. 2014.

H. Liu, I. Chatterjee, M. C. Zhou, X. S. Lu, and A. Abusorrah, “Aspect-based sentiment analysis: A survey of deep learning methods,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 6, pp. 1358–1375, 2020.

H. Liu, M. C. Zhou, and Q. Liu, “An embedded feature selection method for Imbalanced Data Classification,” *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 703–715, 2019.

H. Studiawan, F. Sohel, and C. Payne, “Anomaly detection in operating system logs with deep learning-based sentiment analysis,” *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2020.

H. Wang, M. J. Bah, and M. Hammad, “Progress in outlier detection techniques: A survey,” *IEEE Access*, vol. 7, pp. 107964–108000, 2019.

H. Zahid, T. Mahmood, A. Morshed, and T. Sellis, “Big Data Analytics in Telecommunications: Literature Review and Architecture Recommendations,” *IEEE/CAA Journal of Automatica Sinica*, pp. 1–22, 2019.

H. Zhang, Y. Li, Z. Lv, A. K. Sangaiah, and T. Huang, “A real-time and ubiquitous network attack detection based on deep belief network and support vector machine,” *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 3, pp. 790–799, 2020.

H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, “Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection,” *Neurocomputing*, vol. 407, pp. 50–62, 2020.

I. Ahmed, A. Dagnino, and Y. Ding, “Unsupervised anomaly detection based on minimum spanning tree approximated distance measures and its application to hydropower turbines,” *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 654–667, 2019.

I. Chatterjee, (2021, December 7). *Amazon Customer Review Dataset*. Harvard Dataverse. Retrieved December 7, 2021, from https://doi.org/10.7910/DVN/UUB774.

J. Chen and X. Luo, “Randomized latent factor model for high-dimensional and sparse matrices from industrial applications,” *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 2018.

J. Kim, M. Park, H. Kim, S. Cho, and P. Kang, “Insider threat detection based on user behavior modeling and anomaly detection algorithms,” *Applied Sciences*, vol. 9, no. 19, pp. 4018, 2019.

J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” *8th International Conference on Machine Learning 2001 (ICML 2001)*, pp. 282–289, 2001.

J. Park, “Framework for sentiment-driven evaluation of customer satisfaction with Cosmetics Brands,” *IEEE Access*, vol. 8, pp. 98526–98538, 2020.

J. Singh and P. Tripathi, “Sentiment analysis of Twitter data by making use of SVM, random forest and decision tree algorithm,” *10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 2021.

K. Abdalgader and A. A. Shibli, “Experimental results on customer reviews using lexicon-based word polarity identification method,” *IEEE Access*, vol. 8, pp. 179955–179969, 2020.

K. Chakraborty, S. Bhattacharyya, and R. Bag, “A survey of sentiment analysis from social media data,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

K. R. Jerripothula, A. Rai, K. Garg, and Y. S. Rautela, “Feature-level rating system using customer reviews and review votes,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp. 1210–1219, 2020.

L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent Twitter sentiment classification.,” *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

L. Màrquez and H. Rodríguez, “Part-of-speech tagging using decision trees,” *Machine Learning: ECML*, pp. 25–36, 1998.

L. Wang and J. Lu, “A memetic algorithm with competition for the capacitated Green Vehicle Routing problem,” *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 2, pp. 516–526, 2019.

L. Wang, J. Niu, and S. Yu, “Sentidiff: Combining textual information and sentiment diffusion patterns for Twitter sentiment analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2026–2039, 2020.

L. Zheng, G. Liu, C. Yan, C. Jiang, M. Zhou, and M. Li, “Improved TrAdaBoost and its application to transaction fraud detection,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp. 1304–1316, 2020.

M. Afzaal, M. Usman, and A. Fong, “Tourism mobile app with aspect-based sentiment classification framework for tourist reviews,” *IEEE Transactions on Consumer Electronics*, vol. 65, no. 2, pp. 233–242, 2019.

M. Almaghrabi and G. Chetty, “Improving sentiment analysis in Arabic and English languages by using multi-layer perceptron model (MLP),” *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020.

M. Corain, P. Garza, and A. Asudeh, “DBSCOUT: A density-based method for scalable outlier detection in very large datasets,” *IEEE 37th International Conference on Data Engineering (ICDE)*, 2021.

M. Cui, J. Wang, A. R. Florita, and Y. Zhang, "Generalized graph laplacian based anomaly detection for spatiotemporal Micropmu Data," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3960–3963, 2019.

M. E. Basiri, M. Abdar, A. Kabiri, S. Nemati, X. Zhou, F. Allahbakhshi, and N. Y. Yen, "Improving sentiment polarity detection through Target identification," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 113–128, 2020.

M. F. Schmitt and E. J. Spinosa, "Outlier detection on semantic space for sentiment analysis with Convolutional Neural Networks," *International Joint Conference on Neural Networks (IJCNN)*, 2018.

M. Hu and B. Liu, "Mining and summarizing customer reviews," *ACM SIGKDD international conference on Knowledge discovery and data mining - KDD*, 2004.

M. Li, Y. Ma, and P. Cao, "Revealing customer satisfaction with hotels through multi-site online reviews: A method based on the evidence theory," *IEEE Access*, vol. 8, pp. 225226–225239, 2020.

M. Nguyen, G. H. Ngo, and N. F. Chen, "Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 461–473, 2020.

M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," *9th International Workshop on Semantic Evaluation (SemEval)*, 2015.

M. Riahi-Madvar, B. Nasersharif, and A. A. Azirani, "Subspace outlier detection in high dimensional data using ensemble of PCA-based subspaces," *26th International Computer Conference, Computer Society of Iran (CSICC)*, 2021.

M. Shanmugam, A. Agawane, A. Tiwari, and R. V. Deolekar, "Twitter sentiment analysis using novelty detection," *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020.

M. T. Mahmud, M. O. Rahman, M. M. Hassan, A. Almogren, and M. Zhou, "An Efficient Cooperative Medium Access Control Protocol for wireless IOT networks in Smart World System," *Journal of Network and Computer Applications*, vol. 133, pp. 26–38, 2019.

N. Ding, H. X. Ma, C. G. Zhao, Y. H. Ma, and H. W. Ge, "Driver's emotional state-based data anomaly detection for vehicular ad hoc networks," *IEEE International Conference on Smart Internet of Things (SmartIoT)*, 2019.

N. Hu, J. Zhang, and P. A. Pavlou, "Overcoming the J-shaped distribution of product reviews," *Communications of the ACM*, vol. 52, no. 10, pp. 144–147, 2009.

N. Wu, Z. Li, K. Barkaoui, X. Li, T. Murata, and M. C. Zhou, "IOT-based smart and Complex Systems: A guest editorial report," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 69–73, 2018.

O. Oyebode, F. Alqahtani, and R. Orji, “Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews,” *IEEE Access*, vol. 8, pp. 111141–111158, 2020.

P. Jadon, D. Bhatia, and D. K. Mishra, “A bigdata approach for sentiment analysis of Twitter data using naive Bayes and SVM algorithm,” *Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN)*, 2019.

P. Verma, M. Sinha, and S. Panda, “Fuzzy c-means clustering-based novel threshold criteria for outlier detection in electronic nose,” *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1975–1981, 2021.

P. Y. Zhang, S. Shu, and M. C. Zhou, “An online fault detection model and strategies based on SVM-grid in clouds,” *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 2, pp. 445–456, 2018.

Q. Hou, M. Han, and Z. Cai, “Survey on data analysis in social media: A practical application aspect,” *Big Data Mining and Analytics*, vol. 3, no. 4, pp. 259–279, 2020.

Q. Kang, L. Shi, M. C. Zhou, X. S. Wang, Q. D. Wu, and Z. Wei, “A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4152–4165, 2018.

Q. Kang, X. S. Chen, S. S. Li, and M. C. Zhou, “A noise-filtered under-sampling scheme for imbalanced classification,” *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4263–4274, 2017.

R. Poli, J. Kennedy, and T. Blackwell, “Particle swarm optimization,” *Swarm Intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang, “A survey on opinion mining: From stance to product aspect,” *IEEE Access*, vol. 7, pp. 41101–41124, 2019.

“The role of customer product reviews,” eMarketer, 02-Nov-2010. [Online]. Available: https://www.emarketer.com/Article/Role-of-Customer-Product-Reviews/1008019. [Accessed: 20-Nov-2021].

S. Ali, G. Wang, and S. Riaz, “Aspect based sentiment analysis of ridesharing platform reviews for Kansei Engineering,” *IEEE Access*, vol. 8, pp. 173186–173196, 2020.

S. Han, K. Zhu, M. C. Zhou, and X. Cai, “Information-utilization-method-assisted multimodal multiobjective optimization and application to credit card fraud detection,” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 856–869, 2021.

S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth, and Shubham, “Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation,” *IEEE Access*, vol. 8, pp. 26172–26189, 2020.

S. Jamshidi Nejad, F. Ahmadi-Abkenari, and P. Bayat, “A combination of frequent pattern mining and graph traversal approaches for aspect elicitation in customer reviews,” *IEEE Access*, vol. 8, pp. 151908–151925, 2020.

S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad, “NRC-Canada-2014: Detecting aspects and sentiment in customer reviews,” *8th International Workshop on Semantic Evaluation (SemEval)*, 2014.

S. Li, M. C. Zhou, and X. Luo, “Modified primal-dual neural networks for motion control of redundant manipulators with dynamic rejection of harmonic noises,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4791–4801, 2018.

S. Vosoughi, H. Zhou, and deb roy, “Enhanced Twitter sentiment classification using contextual information,” *6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2015.

S. Yang, Y. Wen, L. He, M. C. Zhou, and A. Abusorrah, “Sparse individual low-rank component representation for face recognition in IOT-based system,” *IEEE Internet of Things Journal*, pp. 1–1, 2021.

S. Zhang, D. Zhang, H. Zhong, and G. Wang, “A multiclassification model of sentiment for e-commerce reviews,” *IEEE Access*, vol. 8, pp. 189513–189526, 2020.

“Sentiment analysis,” Wikipedia, 20-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Sentiment_analysis. [Accessed: 20-Nov-2021].

T. Hu, B. She, L. Duan, H. Yue, and J. Clunis, “A systematic spatial and temporal sentiment analysis on geo-tweets,” *IEEE Access*, vol. 8, pp. 8658–8667, 2020.

U. Yaqub, M. Ali Malik, and S. Zaman, “Sentiment analysis of Russian IRA troll messages on Twitter during US presidential elections of 2016,” *7th International Conference on Behavioural and Social Computing (BESC)*, 2020.

V. Yadav, P. Verma, and V. Katiyar, “E-commerce product reviews using aspect based Hindi sentiment analysis,” *International Conference on Computer Communication and Informatics (ICCCI)*, 2021.

W. Li and B. Xu, “Aspect-based fashion recommendation with attention mechanism,” *IEEE Access*, vol. 8, pp. 141814–141823, 2020.

W. Zhang, J. Wang, and F. Lan, “Dynamic hand gesture recognition based on short-term sampling neural networks,” *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 1, pp. 110–120, 2021.

X. Hu, J. Cheng, M. Zhou, B. Hu, X. Jiang, Y. Guo, K. Bai, and F. Wang, “Emotion-aware cognitive system in multi-channel cognitive radio ad hoc networks,” *IEEE Communications Magazine*, vol. 56, no. 4, pp. 180–187, 2018.

X. Luo, X. Wen, M. C. Zhou, A. Abusorrah, and L. Huang, “Decision-tree-initialized dendritic neuron model for fast and accurate data classification,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2021.

X. Wang, H. Jiang, and B. Yang, “A K-nearest neighbour medoid-based outlier detection algorithm,” *International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2021.

X. Wang, Q. Kang, J. An, and M. Zhou, “Drifted twitter spam classification using Multiscale Detection Test on K-L Divergence,” *IEEE Access*, vol. 7, pp. 108384–108394, 2019.

Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis,” *4th ACM international conference on Web search and data mining - WSDM*, 2011.

Y. Wang, Q. Chen, M. Ahmed, Z. Li, W. Pan, and H. Liu, “Joint inference for aspect-level sentiment analysis by deep neural networks and linguistic hints,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

Y. Zhang, B. Xu, and T. Zhao, “Convolutional multi-head self-attention on memory for aspect sentiment classification,” *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 4, pp. 1038–1044, 2020.

Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: A statistical framework,” *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

Z. Huang, X. Xu, H. Zhu, and M. C. Zhou, “An efficient group recommendation model with Multiattention-based Neural Networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4461–4474, 2020.

Z. Tan, J. Chen, Q. Kang, M. C. Zhou, A. Abusorrah, and K. Sedraoui, “Dynamic embedding projection-gated convolutional neural networks for text classification,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2021.

Z. Wu, S. Huang, R. Zhang, and L. Li, “Video review analysis via transformer-based sentiment change detection,” *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020.