Sentiment Analysis using VADER Classifier

Nandan Hegde
Dept. of ISE
BMS College of Engineering
Bangalore, India

Lekha G Patel
Dept. of ISE
BMS College of Engineering
Bangalore, India

Khushi Agrawal
Dept. of ISE
BMS College of Engineering
Bangalore, India

Dr. Roopa R
Dept. of ISE
BMS College of Engineering
Bangalore, India

Abstract—**Sentiment research plays an important role in understanding consumers' perceptions and feelings towards products or services. Online research has become a primary source of information for potential buyers in today's digital age. Analysing and extracting emotions from customer perspectives can provide businesses with valuable insights, enabling them to make data-driven decisions and improve their products and services accordingly. This paper focuses on a sentiment analysis framework specifically designed to analyse customers' subjective perceptions. The system aims to classify the emotions expressed in the surveys as positive, negative, or neutral. By accurately identifying sentiment, companies can gain a broader understanding of customer satisfaction levels and identify areas for improvement.**

I. INTRODUCTION

In recent years, the advent of online platforms and e-commerce has changed the way consumers interact with businesses. Consumers now have the ability to share their opinions, experiences and reviews about products and services through online surveys. These customer reviews have become valuable information for businesses and potential customers. However, the volume of research makes it impractical for companies to manually analyse it and gain insights from it. This is where sentiment analysis, a subset of natural language processing (NLP), comes into play. Sentiment analysis, also known as mind mining, is simply a process of identifying and classifying the emotions expressed in text. This includes analysing customer comments, social media posts, and other texts to determine whether the sentiments expressed are positive, negative, or neutral. Using machine learning and NLP algorithms, sentiment analytics enables companies to gain valuable insights from multiple customer perspectives, enabling them to better understand customer sentiment and make data-driven decisions The focus of this paper is to provide a comprehensive analysis of the sensitivity analysis of consumer perceptions of products. Our goal is to develop an efficient and accurate sentiment analysis system that can automatically categorize the sentiment expressed in customer surveys, help provide insights into job satisfaction levels, identify areas for improvement, and make informed decisions to improve their products and services.

Through this research, we aim to contribute to the field of sentiment analysis and provide businesses with a practical solution for effectively analysing customer reviews. By extracting sentiments from these reviews, businesses can gain valuable insights into customer preferences, improve their products and services, and ultimately enhance customer satisfaction.

The rest of this research paper is organized as follows: Section II provides a detailed review of related literature and existing approaches to sentiment analysis. Section III outlines the architecture of our model, Section IV lays insight into the methodology used by us, Section V presents the flow of the developed sentiment analysis system. Finally, Section VI concludes the paper with a summary of the findings, contributions, and avenues for future research.

Overall, this research paper aims to contribute to the advancement of sentiment analysis techniques for customer reviews on products and empower businesses with the ability to harness the power of customer feedback to drive their decision-making processes and enhance customer satisfaction.

II. DETAILED SURVEY

[1] researches Stanford Network Analysis Platform (SNAP) that functions as a general-purpose, easy-to-use operations for analysing and manipulating large networks along with high-performance system that can handle massive networks with millions of nodes and billions of edges. It contains various sections that include Related Network Analysis Systems which provides an overview of existing single-machine systems for network analysis and discuss their features, performance, and limitations. SNAP Foundations represents the foundational concepts of SNAP. According to the research SNAP serves as a robust instrument for scrutinizing graphs, delivering efficient graph adjustment, swift algorithm execution, and flexible graph portrayal. Due to these functionalities SNAP stands as a precious asset for addressing intricate graph-oriented quandaries, offering proficient, adaptable, and scalable resolutions.

Authors of [2] discuss that in the real world, social connections frequently demonstrate scale-free distributions wherein a small number of nodes possess a large number of connections, while the majority have only a few connections. Small-world phenomenon is observed, where the average distance between nodes is short and there is a high level of connectivity in the case of strong interconnections between them. Numerous graph generators have been designed to replicate these patterns, such as random graph models and preferential attachment processes. Detecting communities within a network can be approached from either a group-centric perspective or a network-centric perspective, focusing on connections across the entire network. Hence efficient graph discovery techniques are fundamental for identifying societies in expansive and complex networks. Crucial domains for further examination include scalability, community progression, utilization of detection results, and comprehending the significance of patterns in various networks.

[3] explores three approaches in social network analysis: Graph Representation, Content Mining, and Semantic Analysis. The focus is on using graph theory to analyse social network topology, structural modelling, network structure, tie-strength, random walks, community detection, group cohesion visualization, and metrics computations.

Community detection in SNA faces the challenge of lacking ground truth information about community structures in real-world networks. As a result, many new methods have emerged in addition to those mentioned in the paper.

[4] emphasizes the significance of comprehending communication behaviour in social networks (SN) between Social Media users. It highlights the crucial role that discussions and user-generated material (UGM) play in virtual communication. The investigation revolves around constructing an accurate diagram to depict online SN discussions, reconstructing discussions, assigning suitable category tags to comments, revealing minute subjects, scrutinizing the spread and emotions of topics, and identifying recurring patterns in conversation diagrams. The objective of the study is to tackle these inquiries by utilizing extensive data sets from SN platforms. The approach suggested in this study entails scraping the internet to gather information from various social media platforms. This is then followed by text manipulation to cleanse and prepare the collected data. The subsequent steps involve designing the network and assessing their sentiment through text classification and sentiment analysis. Results are displayed in the form of graph.

[5] introduces an overview of graph mining, investigating its uses in various areas. It classifies methods, examines their principles and impacts, and also tackles prevalent challenges in graph extraction. It categorizes the different graph mining algorithms and provides an overview of their basic concepts and contributions by various authors.

Graph mining techniques can be categorized into three main types: graph clustering, graph classification, and sub-graph mining. Certain navigating challenges and their possible solutions are also discussed which include zooming, panning, node identification, link exploration which can be solved by optimizing, filtering, highlighting selected nodes or links.

The authors of the [6] proposed a method for sentiment analysis in social media using various classifiers. They experimented with Naïve Bayes, Maximum Entropy, Decision Tree, Random Forest, Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and Convolutional Neural Networks (CNN). The study found that MLP and CNN performed better than other classifiers, particularly in analyzing sentiment in Twitter data and consumer reviews. The use of sparse vector representation yielded good results for most classifiers, while dense vector representation was more effective for CNN. The paper concluded that intelligent techniques like MLP and CNN can be successfully applied to sentiment analysis in social media and other online communities. These techniques have the potential to enhance sentiment analysis accuracy and improve understanding of user opinions and emotions in social media platforms.

[7] paper proposed a system to detect and classify cyberbullying comments on Instagram. They utilized the Naïve Bayes Classifier algorithm and employed various preprocessing techniques including folding case, cleansing, tokenizing, word replacer, stop words removal, and stemming. The system aimed to accurately classify comments as either cyberbullying or non-cyberbullying. Through experimentation, the authors achieved a best accuracy of 84% using both stemming and non-stemming approaches. Stemming was found to influence the number of detected cyberbullying comments, with a higher count when stemming was applied. The system's results highlight its potential to effectively detect and mitigate cyberbullying on Instagram, contributing to the creation of a safer online environment.

Authors of [8] provide a comprehensive survey of various techniques used for sentiment analysis specifically on Twitter data. The authors address the growing importance of sentiment analysis in social media and the unique challenges associated with analysing Twitter data due to its characteristics such as limited text length and informal language. Furthermore, the authors examine various aspects related to Twitter sentiment analysis, such as feature selection, sentiment lexicons, data preprocessing, and evaluation metrics. They also discuss the impact of domain-specific sentiment analysis and the challenges of handling sarcasm and ambiguity in Twitter data.

[9] method proposed in this paper focuses on sentiment detection in Instagram, specifically targeting the emotions expressed by users through their posts. Instagram's unique features, such as hashtags and image sharing, provide an opportunity to classify users' emotions based on psychologically defined categories. The study deviates from traditional polarity-based sentiment classification and instead uses Thayer's model to classify sentiments. By extracting sentiment keywords through hashtags, sentiment categories are created, and the similarity between sentiment adjectives and keywords is measured to determine the sentiments. The experimental results demonstrate a high average accuracy rate

of 90.7% for all sentiment categories, indicating the effectiveness of the proposed method. The potential applications of this approach extend to various fields, including the analysis of social phenomena through social networking services.

Authors of [10] introduces a methodology for sentiment analysis on Twitter using a Gated Attention Recurrent Network (GARN). The proposed approach involves pre-processing the Twitter dataset, extracting sentiment-based features using a term weight-based model called Log Term Frequency-based Modified Inverse Class Frequency (LTF-MICF), and selecting optimal features using a Hybrid Mutation-based White Shark Optimizer (HMWSO). The results indicate the effectiveness of the proposed GARN approach for sentiment analysis on Twitter data. Future research is suggested to improve the feature selection process and explore larger datasets for further analysis.

[11] introduces a Machine Learning approach to Sentiment Analysis, it begins with data collection, then the pre processing of data is done. After feature extraction we can now perform classification task using classification algorithms like Naïve's algorithm. The paper even discusses various regression tasks and mentions support vector machines. It compares two distinct approaches, a rule-based approach using NLP algorithms and an automatic approach using ML techniques.

Authors of [12] explore a Deep Learning approach to the same. They make use of DNNs for Aspect based sentiment analysis. CNN and RNN models are also used by them. They make use of pretrained models too. However, it mentioned that the model does not perform very well as the ASBA and DL techniques are still in the early stages of development.

[13] explore a challenging dataset and builds effective models based on it. The methodology used by them involves a CapsNet model and a CapsNet-BERT model, the latter being obtained from the pretrained BERT model. Sentence level classifiers achieve good results on a few datasets, however CapsNet outperforms non_BERT baselines on 4 out of 6 datasets.

[14] focuses on cross domain text sentiment analysis using a combination of BERT model, CNN and an adversarial model. First, the input data, including sentence and aspect word representations from both domains, undergoes feature extraction using BERT and CNN with shared weights. BERT extracts the semantic information of the sentences, and CNN further extracts key local features. Dimensionality reduction is applied to the features with high semantic information. The output features of CNN serve as inputs to adversarial classifiers and sentiment classifiers. The results showed that the aspect-level cross-domain sentiment analysis with BERT preprocessing achieved better accuracy and F1 scores compared to the method without BERT preprocessing. The inclusion of the gated activation unit also improved the results compared to sentence-level sentiment analysis without gating.

[15] proposes a framework that aims to identify and extract important topics (aspects) from the text. The proposed technique uses part-of-speech tagging, syntax tree parsing, and frequent itemset mining to extract frequent nouns and noun phrases. Linguistic rules are then applied to filter and eliminate redundant aspects. Sentiment prediction is done and a summary is generated. Results show that sentiment classification performed well, however, the aspect extraction task had poor results.

III. ARCHITECTURE

The server is deployed on Redner which is an cloud stream for web apps and acts as the base of the architecture. Web application is built using streamlit which is a python framework. The software is also coded using python. First page is scraped, then the reviews are cleansed, tokenized, stemmed and tokenized to get the sentiment value.

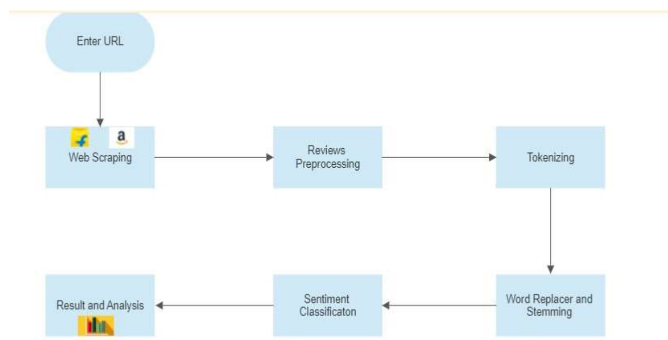

Fig. 1. Proposed Architecture

IV. METHODOLOGY

A. Scraping E-Commerce Page

Web Scraping is done with python's module called beautiful Soup. It is a Python package for parsing HTML and XML documents. It creates a parse tree through which we can extract the data. Difficult part is getting through websites server for scrape. Since we are scaping reviews of an e-commerce website, we need strong headers to bypass the blocking bots of the website. Once we are in, scarping can be done easily. Next part is to scrape multiple pages. This can be done using a complex loop function.

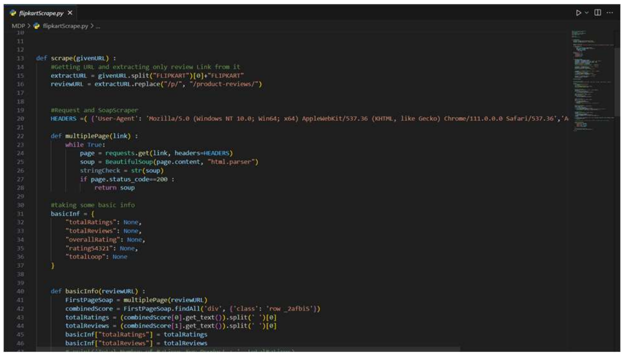

Fig. 2. Scrape Function

B. Reviews pre-processing

Pre-processing is converting the long data into short text to perform other processes such as classification, detecting emotions, and providing sentiment analysis. Duplicate rows are dropped, unfilled names are filled and data is sorted according to date. Emojis are taken out and stored for further processing. This pretty much sums up the job of pre-processing.

C. Tokenizing

Tokenizing is an essential step in sentiment analysis that involves breaking down a text document into smaller units called tokens. In this process, the text is divided into words, sentences, or even sub word units, depending on the specific requirements of the analysis. Tokenization serves as a foundation for further analysis by providing a structured representation of the text data. Various tokenization techniques can be employed, such as whitespace-based tokenization, which separates words based on spaces, or more advanced methods like word tokenization using natural language processing (NLP) libraries, which consider linguistic rules and context. By applying word replacer and stemming techniques as part of the preprocessing phase, sentiment analysis algorithms can normalize text, reducing noise and variability, and enhancing the overall effectiveness of sentiment analysis on e-commerce page data.

D. Word Replacer and Stemming

Word replacer and stemming are two techniques commonly used in text analysis and sentiment analysis to preprocess and standardize textual data. Word replacer involves replacing specific words or phrases in a text document with their corresponding synonyms or alternative representations. This technique aims to reduce lexical variations and enhance the consistency of the text, which can improve the accuracy of sentiment analysis. Stemming, on the other hand, is a linguistic process that reduces words to their base or root form, known as the stem. By removing prefixes and suffixes, stemming reduces different word forms to a common base, allowing for more effective analysis of sentiment. For example, words like "running," "runs," and "ran" would all be reduced to the stem "run.

E. Classification using VADER

The VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon is a widely used tool for sentiment analysis in natural language processing. It is specifically designed for analyzing social media texts and other short, informal messages. The VADER lexicon consists of a pre-compiled list of words and phrases, each of which is assigned a sentiment score ranging from -1 to 1, indicating the negativity, positivity or neutrality of the term. By leveraging the VADER lexicon, sentiment analysis algorithms can quickly and effectively determine the sentiment polarity of e-commerce page data, aiding in understanding customer opinions and satisfaction levels. The lexicon incorporates not only individual word scores but also considers context, grammar rules, and punctuation to accurately determine the sentiment of a given text. VADER accounts for the intensity and polarity of sentiment, allowing it to handle both strongly positive or negative sentiments and nuanced sentiments.

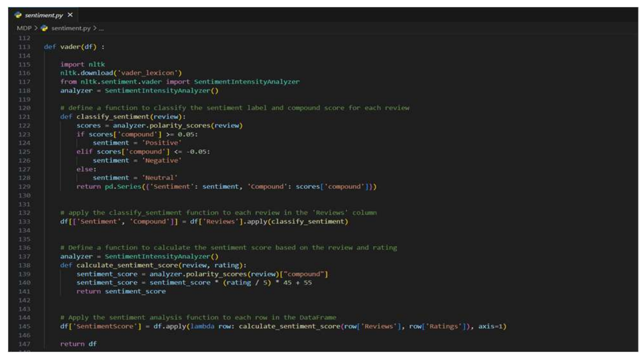

Fig. 3. VADER Classifier

F. Building The web app

Web app is built using streamlit, which is an open-source Python framework for constructing dynamic web apps. In the home page we can enter the URL of our desired product. The scarpe function runs through it and gives the data with multiple graph analysis. The first graph shows how is the sentiment trend over the months, with various labels for user understanding. It display's number of positive, negative and neutral sentiments. Graph lines are depicted with various colors for faster visualization. Bottom of the page contains bar and pie charts depicting total number of reviews.

Fig. 4. Sentiment Analysis graph

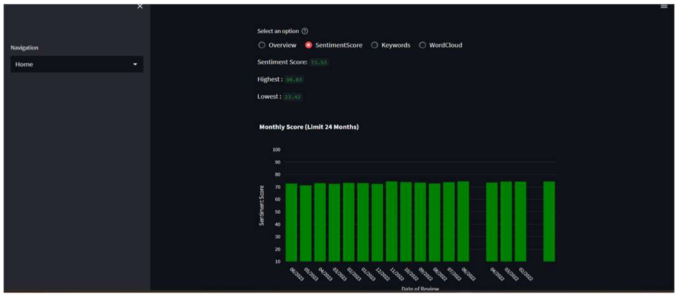

Fig. 5. Average and Monthly Sentiment Scores

Next Section of graphs depicts Average Sentiment Score. For the above graph (Fig 5) the overall sentiment score is 73.53 with the highest being 98.83 and lowest being 23.42. Anything below 60 is an average product and score under 40 is poor. A bar chart depicting monthly scores is given from last 24 months. We can analyse that product has been stable with handling reviews with stagnant performance. For 2 months product has been out of stock. For the metric given below (Fig 6) is an word-cloud which analyses all positive and negative words and gives a cloud sector of words. Higher the size means the word has been used more times. The same has been provided for negative also. One more unique feature is ability to highlight emoji count and give important keywords from the user reviews data.

Fig. 6. Insights - Deviations from Ideal Values

Further improvements are to be made in providing complex vader classifiers for more accurate scores and decrease the time required for scraping. Compared to other sentiment sites we have worked on emojis and concentrated on keywords.

G. Deploying the web app on Render

We have to first list out all the pip files used. The project has used 20 pip files fore building the products. This has to be listed in requirements files. Once this is done, deploying becomes an easy process. There are other services such as Heroku or amazon AWS which are paid services. Paid services provide much faster deployment and service doesn't go down. For free service Render and Netlify are good options. Render offers wide range of services for python files. Due to these reasons Render is best option for python modules. It is important to keep checking back the hosting service every month, since it might go down or the service might become paid.

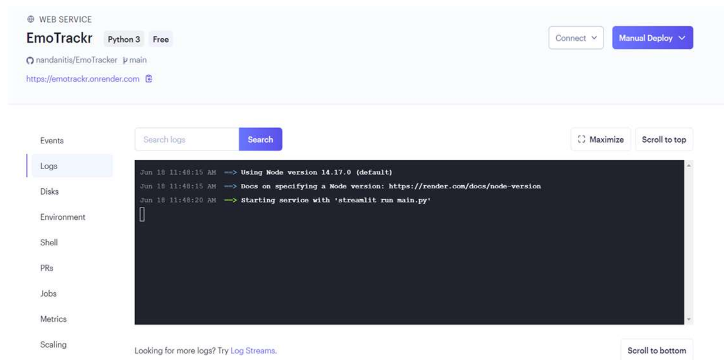

Fig. 9. Render Cloud Service

V. FLOW OF PROPOSED SERVICE

The user opens the website and enters the URL. Next code is converted to review link and web pages are scraped. Data is converted to dataframe and preprocessing steps takes place. Words are tokenized and stored, this data is stemmed to single word. Next Vader Classifier is applied to obtain the sentiment and sentiment scores. Using this graph's are built and displayed on web page. User can go to different pages to analyse different graphs. There is a contact us page and about us to know about service and what is the website doing

VI. RESULTS

The proposed sentiment analysis methodology, which combines web scraping, pre-processing, tokenizing, word replacement, stemming, and classification using the VADER lexicon, was implemented and applied to a dataset obtained from an e-commerce website. The results of the analysis provide valuable insights into the sentiment trends and customer opinions related to the analysed product. The sentiment analysis system successfully scraped the e-commerce website and collected a substantial number of reviews. The preprocessing steps, including dropping duplicate rows and filling unfilled names, ensured the data's cleanliness and organization. The tokenization process effectively divided the text into words, enabling further analysis.

By employing word replacement techniques, lexical variations were reduced, and the consistency of the text was enhanced. Stemming further normalized the textual data by reducing different word forms to their base or root form. These preprocessing techniques contributed to reducing noise and variability in the data, enhancing the accuracy of the sentiment analysis. The classification process utilizing the VADER lexicon proved to be highly effective in determining the sentiment polarity of the e-commerce page data. The sentiment scores, ranging from -1 to 1, accurately reflected the positivity, negativity, or neutrality of the expressed opinions. The incorporation of context, grammar rules, and punctuation in the VADER lexicon enabled nuanced sentiment analysis, capturing both strongly positive or negative sentiments and subtle variations.

The web app built using Stream lit provided an intuitive interface for users to input the desired product URL and visualize the sentiment analysis results. The generated graphs illustrated the sentiment trends over time, the distribution of positive, negative, and neutral sentiments, average and monthly sentiment scores, and insightful word clouds highlighting important keywords. The analysis of the sentiment trends revealed valuable information about customer satisfaction levels and the overall sentiment associated with the analysed product. The average sentiment score, calculated as 73.53, indicated an above-average product satisfaction. The monthly sentiment scores showed a stable performance, with occasional out-of-stock periods. The word clouds provided a comprehensive overview of the positive and negative words frequently mentioned in the reviews.

Overall, the results of the sentiment analysis using the proposed methodology demonstrated its effectiveness in understanding customer opinions and satisfaction levels related to the analysed e-commerce product. The system's unique features, such as the focus on emojis and keywords, contributed to capturing more nuanced sentiments. However, there is room for improvement, particularly in enhancing the accuracy of the VADER classifier and reducing the scraping time.

VII. CONCLUSION

In conclusion, this research paper focuses on the development of a sentiment analysis framework designed to analyse customers' subjective perceptions in online surveys. By accurately classifying the emotions expressed in customer feedback as positive, negative, or neutral, businesses can gain valuable insights into customer satisfaction levels and identify areas for improvement. The proposed sentiment analysis system offers an efficient and accurate approach to automatically categorize sentiments and provide actionable insights for decision-making. By leveraging sentiment analysis techniques, businesses can make data-driven decisions, enhance their products and services, and ultimately improve customer satisfaction. This research contributes to the field of sentiment analysis and provides a practical solution for businesses to effectively analyse customer reviews and extract meaningful sentiments. Future research can further explore advanced techniques and approaches to enhance the accuracy and applicability of sentiment analysis in the context of customer perceptions.

REFERENCES

[1] JURE LESKOVEC and ROK SOSIC, Stanford Universityˇ, Stanford University "SNAP: A General-Purpose Network Analysis and Graph-Mining Library"

[2] Lei Tang and Huan Liu Computer Science & Engineering Arizona State University L.Tang@asu.edu, Huan.Liu@asu.edu "GRAPH MINING APPLICATIONS TO SOCIAL NETWORK ANALYSIS"

[3] Andry Alamsyah1, Budi Rahardjo2, Kuspriyanto2 "Social Network Analysis Taxonomy Based on Graph Representation"

[4] Marco Brambilla 1,∗ , Alireza Javadian Sabet 2 , Kalyani Kharmale 3 and Amin Endah Sulistiawati "Graph-Based Conversation Analysis in Social Media"

[5] Swati A. Bhavsar & Varsha H. Patil & Aboli H. Patil "Graph partitioning and visualization in graph mining: a survey"

[6] Kassinda Francisco, Dr. Chandra J. "Sentiment Analysis on Social Media Data Using Intelligent Techniques." ,International Journal of Engineering Research and Technology. ISSN 0974-3154 (2019).

[7] Muhammad Zidny Naf'an, Novanda Alim Setya Nugraha, " Sentiment Analysis of Cyberbullying on Instagram User Comments" (2019).

[8] Vishal A. Kharde, S.S. Sonawane. " Sentiment Analysis of Twitter Data: A Survey of Techniques." International Journal of Computer Applications (0975 – 8887) (April 2016)

[9] Shweta Gangrade, Nirvishesh Shrivastava, Jayesh Gangrade, "Instagram Sentiment Analysis: Opinion Mining." (2019)

[10] Nikhat Parveen, Prasun Chakrabarti, Bui Hung and Amjan, " Twitter sentiment analysis using hybrid gated attention recurrent network." Journal of Big Data(2023)

[11] P Ancy Grana*1, Jawaharlala Nehru Engineering College, MGM University, India, SENTIMENT ANALYSIS OF TEXT USING MACHINE LEARNING MODELS.

[12] Hai Ha Do, PWC Prasad∗ , Angelika Maag, Abeer Alsadoon, DEEP LEARNING FOR ASPECT-BASED SENTIMENT ANALYSIS: A COMPARATIVE REVIEW

[13] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, Min Yang, CHALLENGE DATASET AND EFFECTIVE MODELS FOR ASPECT-BASED SENTIMENT ANALYSIS.

[14] Ning Liu and Jianhua Zhao, A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text.

[15] Edison Marrese-Taylor a, Juan D. Velásquez a, Felipe Bravo-Marquez b, Yutaka Matsuo c, TOURISM PRODUCTS USING AN ASPECT-BASED OPINION MINING APPROACH.