Grocery Apps, ChatGPT, Sentiment Analysis

AP Research 27 April 2024

Research Question: “To what extent can ChatGPT assist Grocery Delivery Apps through Sentiment Analysis in comparison to other Algorithms”

Approach: To Explore
Design: Experimental

Method: Quantitative & Qualitative Combining both Comparative Analysis and Sentiment Analysis

Data Source: Primary

Method: Comparative Sentiment Analysis

This study uses a quantitative and qualitative design with Comparative Sentiment analysis to explain if the ChatGPT
the model produces the best results for Sentiment Analysis in comparison to other top-performing Sentiment Analysis Algorithms.

Data Source: Primary

Keywords:

Machine Learning, ChatGPT, Sentiment Analysis

Word Count: 4865

Discipline: Mathematics & Computer Science

Abstract

Among the 55% of the 333.3 million in the US population who use online grocery shopping, it means there is an increased demand for proper consumer service (TIDIO). Poor customer service affects both sides of the grocery spectrum. This study utilizes ChatGPT, an artificial intelligence chatbot to conduct a sentiment analysis. This algorithm could be used by marketers and a variety of businesses to speed up the review of customer complaints and concerns. By developing a response through multiple neural networks, it creates accurate and powerful responses for businesses to understand and utilize. This algorithm could effectively determine what was communicated with a 98.85% accuracy. In comparing the algorithm's metrics to other Sentiment Analysis algorithms, ChatGPT was determined to be the best algorithm for accuracy, but not timeliness. This study's findings could demonstrate ChatGPT's vast potential for training on this data and predicting future consumer reactions on specific products and actions.

Introduction

The 55% of the 333.3 million in the US population who use online grocery shopping caused an increased demand for proper consumer service (TIDIO). Considering the massive demand on proper grocery delivery, the pressure to keep up is only getting higher. Being behind on worker and customer complaints, corporations such as Doordash and Instacart have shown mistreatment to its extremes. The Los Angeles Times quotes up to "5-20% of shoppers" would be fired weekly within the same department due to the poor customer ratings system. Furthermore, Doordash has found ways to wiggle around the legal requirements of pay for up to 3.4 million gig workers and $31 billion in lost income (Reuters). The same problem is found on the other side, demonstrating potential retaliation by stealing food or poor job performance.

Having a poor response to these problems with wait times lasting up to 3 years, it's caused prolonged suffering for both audiences of the grocery shopping industry. As a measure to hasten the review process through the breakthrough of new AI, I believe that it should be taken advantage of through Sentiment Analysis. In this study, we'll be assessing ChatGPT's ability to conduct a sentiment analysis for the grocery companies who have and haven't used it on consumer reviews.

COVID-19 Effects on Grocery Shopping

Grocery shopping is no longer the same today; it has seen a paradigm shift in the way this activity is undertaken, with various drivers to this change being examined, including impacts and challenges, and more specifically, artificial intelligence (AI) as a shaping force in the current landscape.

Consumer behavior concerning grocery shopping underwent a considerable transformation due to digitization and the pandemic shock effect. According to a report published in March 2020, 42% of the US population purchased their groceries online at least once a week, marking a stark increase from 22% just 2 years ago (Soper, 2020). The two types of concerns emphasized by the National Library of Medicine — health and finance — that people had during the outbreak. On the other hand, major factors influencing consumers' uptake of online grocery delivery are notably fear concerning health safety as well as economic instability. Multiple accounts were made where workers complained about how bigger corporations like Doordash don't account for full expenses. Luckily for them, Doordash would be forced to change its pay model due to having revealed that their "tips" were being used to subsidize its payment to workers. (Freytas-Tamura, 2021)"

It is, however, a social challenge to ensure that these AI tools, such as OpenAI's ChatGPT-4, are advanced in technology but also unbiased, precise, and culturally responsive in their assessment. Companies like Instacart, Microsoft, Duolingo, and many other big enterprises include ChatGPT on their websites (Marr, 2023). Nevertheless, erroneous AI-fueled sentiment analysis may trigger an array of problems ranging from lopsided market research data to the propagation of biases in society, ending up with a failure to truly reflect people's positions on matters of importance. A misinterpreted review about the service on the Instacart platform could lead to severe consequences for the shopper. According to an Instacart Shopper in Vox, her rating "dropped to 4.96" and her earnings went "from $25 per hour to much lower, likely below New York's $15 minimum wage."(Vox)

Online Reviews

Online reviews have been a feature companies have used to get feedback for the customer's dissatisfaction or satisfaction. Customer reviews often are taken on a score from 1-5 with feedback describing why they made the score. Using customer feedback, companies have grown in massive scales because of how they would please the public. The variance and the number of reviews have both determined the popularity and the sale of products (Floyd, 2014). 70 percent of customers look at online reviews before making their final purchase decision, and 63 per cent of customers are more likely to buy the product if it has higher product ratings and positive reviews (Rauschnabel et al., 2019; MacDonald, 2018).

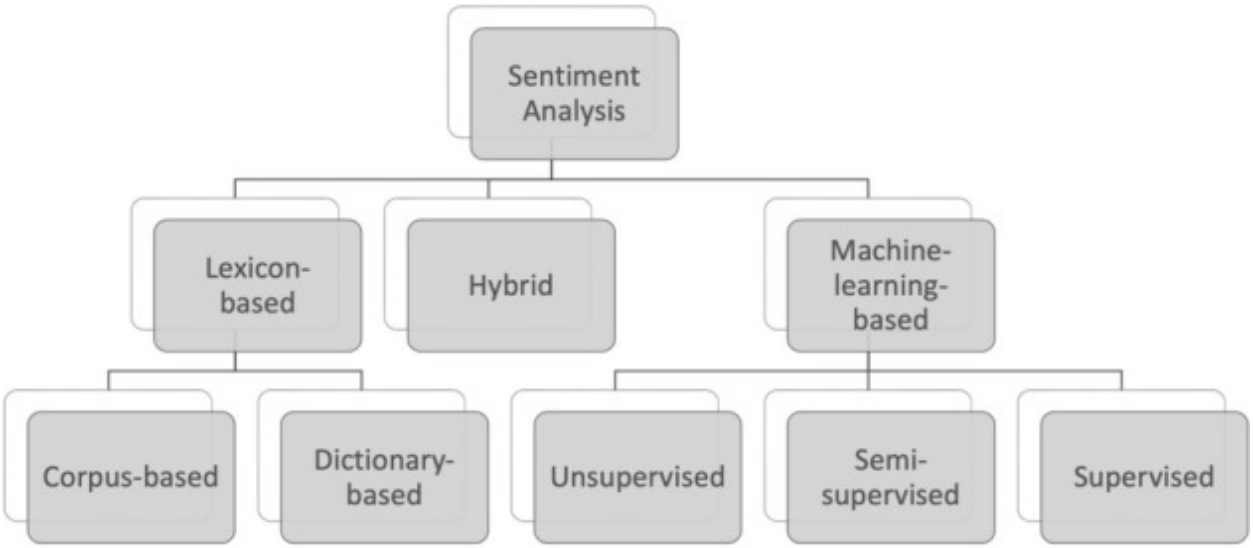
In a survey conducted by Zendesk with up to 1046 participants, 88% have been influenced to buy a product due to customer reviews (Martech). Despite this, in a study by Wenjing Duag the rating of online user reviews has no significant impact on movies' box office revenues after accounting for the endogeneity (Gesenhues, 2013). To add on, consumers have

been found to read online reviews while paying attention to review scores and other contextual information such as a reviewer's reputation and reviewer exposure (Nan Hu, 2008).

Media platforms such as ConsumerAffairs, Reviews.io, and TrustPilot provides a space for users to share their ideas and thoughts on the quality of effort both the platforms and the shoppers have provided. This could be done through text, pictures, videos, reviews, etc. To interact with other users, the customers could promote or comment through likes and upvotes. ConsumerAffairs is filled with an advisory board filled with experts who have had up to 3+ million real-world testimonies. As for Reviews.io, they've won multiple G2 awards for having the best results, usability, and implementation in 2024. Moreover, they have up to 9000+ brands who trust this reviewing service. Finally, as for TrustPilot, they received a Transparency Report in 2022 describing TrustPilot as an online review community that receives "genuine feedback from customers".

Sentiment Analysis

Figure 1: Sentiment Analysis Chart



Sentiment analysis is a tool based on the text mining method for text analysis to find out the reader's emotion as polarity towards a particular topic (Walaa Medhat). Sentiment analysis can be broken down into three parts: Lexicon-based, Hybrid, and machine learning, which could be broken down further into Corpus-based and Dictionary-based for a Lexicon-based algorithm and Un, Semi, and normally Supervised for Machine Learning (Qi, 2023). This task has been realized when potential customers or users leave feedback in the form of reviews to help the marketers identify and rectify the existing issues. Evaluating the performance and the reviews, especially, helped the companies in enhancing their services and sales to compete in the marketplace (Watson). An emotion analysis enables us to use the opinions of the consumers for further learning. The meaning could be demonstrated by stating that unidimensional meaning implies fleeting opinions, but understanding the polarity of a message goes deeper and it allows more opinionated people (Tang et. Al, 2019).

Srivastava finds Naive Bayes to be the strongest algorithm in sentiment analysis, claiming to have an “AP value is 0.99 which shows that the training set is perfect for the model development”. However, from the perspective of Shawwal, he finds Lexicon as one of the top algorithms that gave an average accuracy of 87.33%. The Lexicon model is based on a dictionary of scores with defined sentiment scores with a scale from -1 to 1. The -1 score would be defined as negative, a score closer to 0 is neutral, and 1 would be positive. As for the Corpus Algorithm, it trains upon a large dataset and learns patterns. Online platforms are used by customers to share their opinions on goods and services. Customers also write differently; these differences in writing styles are a reflection of the diversity of the consumer base. This issue can be solved with the aid of sentiment analysis (Hu, 2012).

Figure 2: Methodology of Lexicon & Corpus Algorithm

Used in Xi's Study

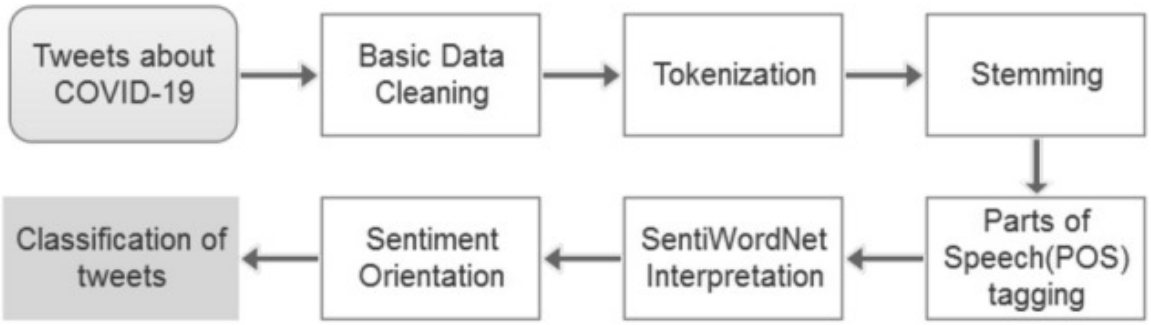
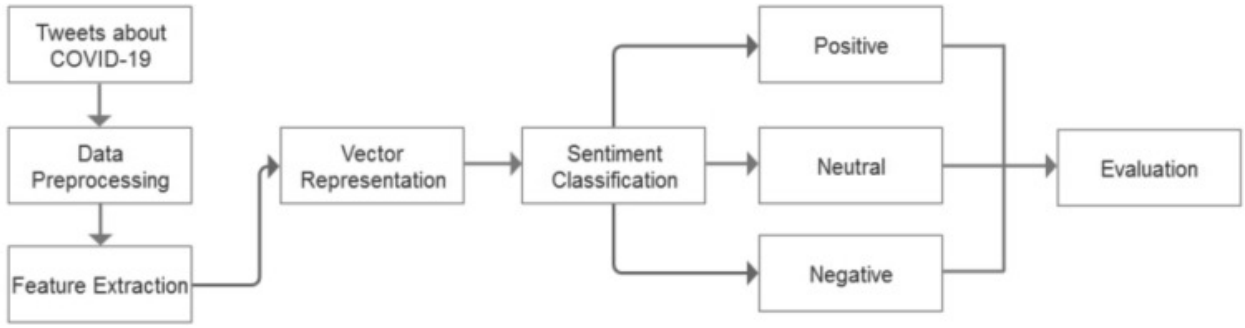


Figure 3: Methodology of Machine Learning Algorithm

Used in Xi's Study



As for the machine learning model, it's trained on a certain subset of values and creates responses based on those results. Both models undergo cleaning to maximize the accuracy potential. Sentiment analysis is an extremely effective forecasting method that may be used to construct machine learning models (Ravi and Ravi, 2015). As a measure for the accuracy of the algorithms, it's commonly separated into four categories: precision, recall, f1-score, and accuracy.

Figure 4: Assessment of Algorithms

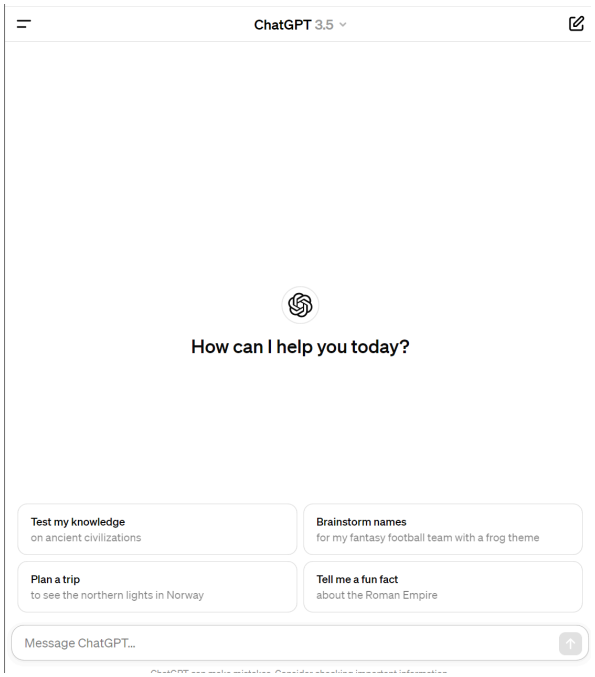
Used in Srivastava's Study

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP}
\end{aligned}
$$

The GPT-series-based model created by OpenAI under the GPT (Generative Pre-trained Transformer) architecture is a state-of-the-art illustration of generative AI (Artificial Intelligence). It is constructed of and responds to the human-like input it is fed in language (Konstantinos). Its capability in various ranges of applications, freedom to the public, and quality performance have inspired this study.

ChatGPT & Algorithms

Figure 5: ChatGPT



ChatGPT has become a breakthrough in the year of 2023. To be specific, it reached 100M monthly active users in just two months (Byte Byte Go). ChatGPT utilizes a LLM (Large Language model) - a type of neural network model that is trained on massive amounts of text data to understand and generate human languages. The model uses the training data to learn these statistical patterns and the relationships between words in the language and then utilizes this knowledge to predict the subsequent words, one at a time (Wu et al., 2023). Being fine-tuned with RLHF (Reinforcement Training from Human Feedback) by gathering feedback from people and then iteratively improving the model's performance using PPO, it allows GPT-3.5 to generate better responses tailored to specific user requests. 175 billion patterns set across 96 layers making it one of the largest deep learning models ever created. It’s organized by tokens, numerical representations of the words and trained on large chunks of internet data. Finally, it applies a GPT (Generative Pre-trained Transformer) network that works with a large dataset of texts whereby the model is taught to understand context, grammar, sentiments, slang or idiomatic

expressions. ChatGPT is capable of reasoning the message according to the context and brings out apt responses (Byte Byte Go). Therefore, besides separating words for sentiment analysis, the machine can comprehend the overall message.Despite this, it needs proper guidance though because it can create outputs that are untruthful.

With the evolution of artificial intelligence increasing the convenience of consumers in marketing, marketers must keep up with the times. ChatGpt has been found to have an accuracy of up to 92.1% for counseling comments and is still in its evolutionary stages. It continuously takes in input from users and understands natural language from humans. It's at the stage where it could answer Medical Licensing Exams with a score of up to 60%.

To begin, there were a variety of studies referring to Native Bayes as the strongest model in machine learning. It is possible to categorize customers and forecast their purchasing patterns using the Naïve Bayes classifier. It might be applied to the creation of artificial intelligence models that forecast customers' future purchase intentions. According to Wu et al. (2015), Naïve Bayes is an effective classification technique for machine learning models. A more accurate prediction model can be produced by sentiment analysis combined with the naïve Bayes classifier method (Sundararaj and Rejeesh, 2021). This narrowed down my conclusion that Native Bayes would be the strongest machine learning algorithm to compare to ChatGpt. As for the lexicon based model, a Stanford dataset showed that the best performing Lexicon algorithm was VADER showing a 72% accuracy (M. A. Al-Shab, 2020).

TF-IDF

TF-IDF Calculations

Used in Srivastava's Study

$$TF(Term\ frequency) = \frac{No.\ of\ repetition\ of\ words\ in\ sentence}{No.\ of\ words\ in\ sentence}$$

$$IDF(Inverse\ document\ frequency) = \log[(No.\ of\ sentences)/(No.\ of\ sentence\ containing\ words)]$$

$$Measure\ of\ the\ relevancy\ of\ a\ word\ to\ a\ document = TF \times IDF.$$

In its most basic form, TF-IDF calculates the relative frequency of a word in a given text in relation to its inverse proportion throughout the corpus of documents. This computation, intuitively, establishes the degree of relevance of a word in a given material. Compared to popular terms like articles and prepositions, words that are frequently used in one or a small number of texts typically have higher TF-IDF scores (Ramos, 2003). According to Tobarra et al. (2014), the number of times a word appears in the text increases the value of TF-IDF, which is then offset by the number of times the term occurs in the corpus. This makes TF-IDF the perfect feature extraction algorithm that could help determine if ChatGPT makes the right conclusion for the reasoning of the review score.

Gap

Of the Sentiment Analysis models, algorithms such as Lexicon, Natural Language Processing, Support Vector Machine, and Text mining have all been used in 10+ studies (Shaeeali, 2020). Moreover, studies would focus on the lexicon and machine learning models rather than trying to enter the AI world as much (Yuxing Qi, 2023). There haven't been any studies contributing to how the new breakthrough in ChatGPT could potentially predict consumer behavior and stop problems before they happen. Considering how ChatGPT couldn't be found among the Grocery review sentiment analysis and how it wasn't properly investigated

for proper company usage, I decided that ChatGPT was essential for study in accuracy and ability. In doing so, it could contribute to potential prediction of emotion.

Therefore, this has led me to the research question: How accurate and beneficial is ChatGPT in making a text-mining sentiment analysis on Grocery Customer Review

Method

For the sentiment analysis component of my AP research project, I have devised a methodology for systematically collecting, analyzing, and interpreting data. The first step involves acquiring a diverse dataset of textual examples that demonstrate sentiment from wide ranges of popularity. From least popular business to most popular business, I took reviews from Shipt, Instacart, and Doordash. These range from delivery reviews to company reviews to ensure that a comprehensive analysis can be done across fields. I have taken these reviews from sources including Consumer Affairs, Reviews.io, and TrustPilot from mostly verified reviewers.

Each review gets boiled down to one of four tags: positive, negative, or in some instances, neutral. I will be taking the review Description, Location, Date, and Starss from each of the reviewing websites. As a measure to protect the user's security, I will not be taking any names.

This dataset has baseline sentiment labels (positive, negative, neutral, and mixed). Few categorical sentiment labels are essential since performance generally increases with limited label sets. Therefore, we have ensured sizable reviews for context with at least 10-20 words. These labels were obtained from manual annotations provided by human raters or were extracted from datasets with pre-existing sentiment annotations.

The performance of sentiment analysis is evaluated based on metrics such as accuracy, precision, recall, and the F1 score. These calculating metrics are needed for effectiveness in the

sentiment analysis (Orozco-Arias). An algorithm's performance in a certain niche is best proven with these parameters. However, metrics like inter-rater reliability are utilized to gauge consistency in cases where human annotations are utilized.

Moreover, throughout the process of calculating the scores, I will be understanding all aspects of the algorithm's abilities by pointing out four categories. Drawing inspiration from Hartmann's process, I will also be separating the evaluations into four different sub-categories: Interpretability, Robustness, Efficiency, and Context. Interpretability is defined as the person's ability to understand the processes and how it got to its conclusion (Datarobot). Robustness is an algorithm's ability to keep up its accuracy despite hidden variables. Efficiency is essentially the time complexity of the algorithm. Finally context will be defined as the algorithm's ability to read the context of each review as a whole rather than the sum of individual values.

In addition, I will annotate these samples manually with their sentiments and any subtleties noted (e.g., sarcasm, mixed emotions) as a reference for evaluating ChatGPT's interpretations. We compare the sentiment analysis performance with each type of Sentiment analysis algorithm. The polarity and subjectivity will be analyzed to investigate the sentiment of each review accurately. This comparison will act as a benchmark, aiding us to understand the relative effectiveness of different sentiment analysis methodologies.

Data Cleaning

This data was converted from a sheet to a CSV file (comma separated file) to be inserted into VS code for input into multiple sentiment analysis algorithms in the Jupyter notebook. Juypter is a python based data scientist software. Additionally, Scikit-learn was installed in the jupyter considering how it is one of the most important machine learning libraries in Python (Raul Garreta). Along with Scikit-learn, other python libraries were imported such as numpy and

the panda library, which is used for working with arrays and cleaning the data. To run the data on the Lexicon-based algorithms, the reviews have been reduced to lowercase and punctuation was taken away. This would ensure that the data is run. However, for the ChatGpt algorithm, no data necessarily had to be cleaned. Rather, the data was kept the same to help the ChatGPT algorithm use context and capitalization to understand tone. Term frequency and inverse frequency was additionally used to find the relevance of the online reviews of consumers. TF-IDF takes the importance of every word and the topics that are most important to the consumers. For calculating the performance of the algorithms, we used the precision, recall, f1-score, and accuracy approach due to its high effectiveness.

ChatGpt Methodology

Regarding the methodology for ChatGPT, the LLM will be prompted with, "Conduct a sentiment analysis on these algorithms and include a brief explanation for each." The algorithm's punctuation and uppercase use will not be erased due to ChatGPT's ability to understand context. Given the raw reviews, the ChatGpt algorithm's full response will be taken in segments of 10 due to its inability to understand multiple reviews at once. This will include its word-based review of the sentiment and the reasoning for its choice.

Lexicon-Based Methodology & Machine Learning Methodology

As for the algorithms besides ChatGPT, Data input will be converted to a CSV file and fed into a Jupyter Notebook on Visual Studio Code. In Python, the choice of our data analysis software is a Jupyter Notebook. One notable example of Python software libraries is the pandas library, which provides powerful yet easy-to-use data structures and manipulation in tabular format. These are very handy while handling and cleaning up the data array, and dealing with subjectivity analysis.

After the sentiment analysis with ChatGpt and other algorithms, sklearn will be used to assess ChatGpt's accuracy in determining the factors behind the consumer's review. As for the Hybrid Model, the only difference is how the algorithm will only take the responses that both the Lexicon based model and the machine learning model can agree on.

Error Analysis

Finally, as an error analysis, I will check where ChatGpt has made its mistakes and attempt to decipher patterns in its errors and right answers. Whether it be the
In this paper, we focus on the qualitative feedback of our analysis, particularly in the cases of ambiguous or mixed-sentiment text extracts. This provides insights into how the analysis correlates with human interpretation and offers more nuanced views of sentiment detection.

Online Grocery Stores
No. of Reviews
Sources: https://www.trustpilot.com/review/shipt.com)
https://www.consumeraffairs.com/food/instacart.html
https://www.reviews.io/company-reviews/store/doordash

Table 1:

# of Negative Reviews	# of Neutral Reviews	# of Positive Reviews	Total # of Reviews
164	15	50	229

Results

Table 2:

Type of Company	Instacart	Doordash	Shipt
Total # Of Reviews	91	53	56

Table 3:

Sample 5-star Positive review: “I ordered through DoorDash today...the representative was so nice and helped to rectify the situation. I definitely will choose DoorDash over UberEats from now on.”	**Sample 1-Star Negative Review:** "We placed 2 orders from little Ceasars and decided to have them delivered... We never got our second order and the tip he complained about was on the second order... We will not use doordash ever again. And from now on I will pick up our food. We lost $100 that night..."

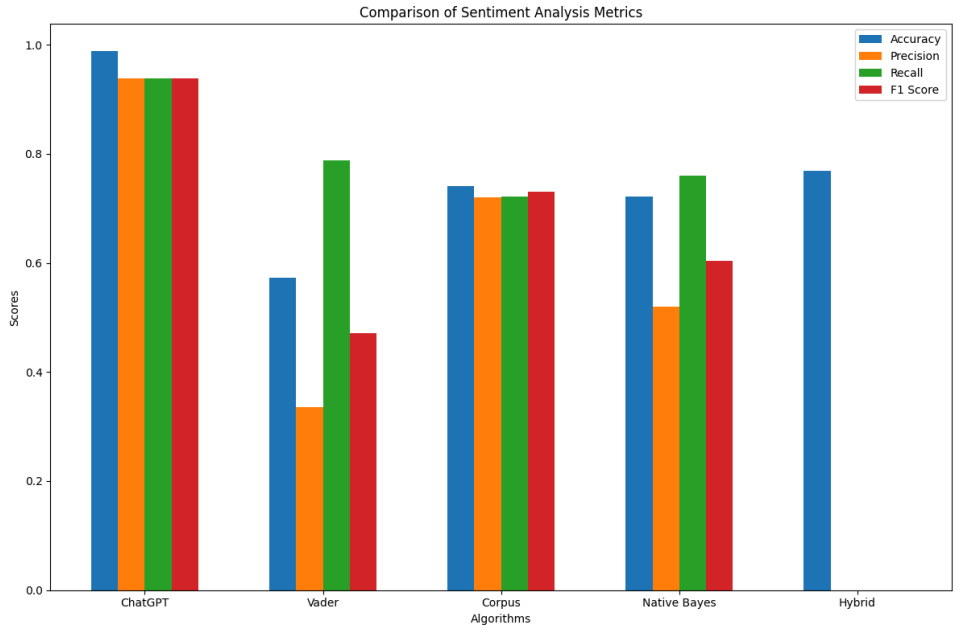


Table 4

Of the 229 reviews accumulated, the accuracy shown by TextBlob is moderate 58.60%, and the substantially low precision 35.43% accompanied by an ideal recall 86.54%. The output highlights the ability of TextBlob to make the decision of negative sentiment but also at the same time he may give out wrong messages.

A chatbot, ChatGPT, proves to be great in terms of accuracy and precision (93.75% each) and recall rates (98.85%). In addition, contrary to other models, ChatGPT demonstrates state-of-the-art precision/recall trade-off leading to its high quality sentiment distinction and least probable error rate.

Based on corpus Naive Bayes and corpus based sentiment analysis the accuracies being achieved are 74.07 % and 72.09 % respectively. These networks process at a satisfactory level, but they fail to produce balanced accuracy in both detection cases, in particular, when it comes to recognizing negative emotions.

The hybrid classifier may be bragged with being accurate, as it can have 76.92% of accuracy, but it fails to do meaningful results as it really cannot agree on positive reviews, zeroing precision, recall, and F1 score for positive sentiment detection.

When investigating TF-IDF assessments of customer reviews and ChatGPT-produced impressions, we will discover different commonalities. Words, which include 'order', 'delivery', 'service', 'customer' and 'time', often receive high TF-IDF scores in both customer review text and ChatGPT-generated sentiments. Thus, we can observe a parallel between verbal indications singled out by ChatGPT customers and those pointed out by the language tools.

As a result, in making my conclusions for ChatGPT, I put robustness and Context for high accuracy due to its high accuracy and its ability to read the full review. As for the other algorithms, I put them at moderate and low for the two categories due to their sub-par accuracy and inability to read the context of the full review. However, on a different spectrum, ChatGPT had terrible Efficiency and interpretability due to how it would only be able to process reviews in subsets of 10 and super complex neural networks. On the other hand, the algorithms aside of ChatGPT had a high interpreablity and efficiency due to its instantaneous response and easy to understand training models.

Chart 1

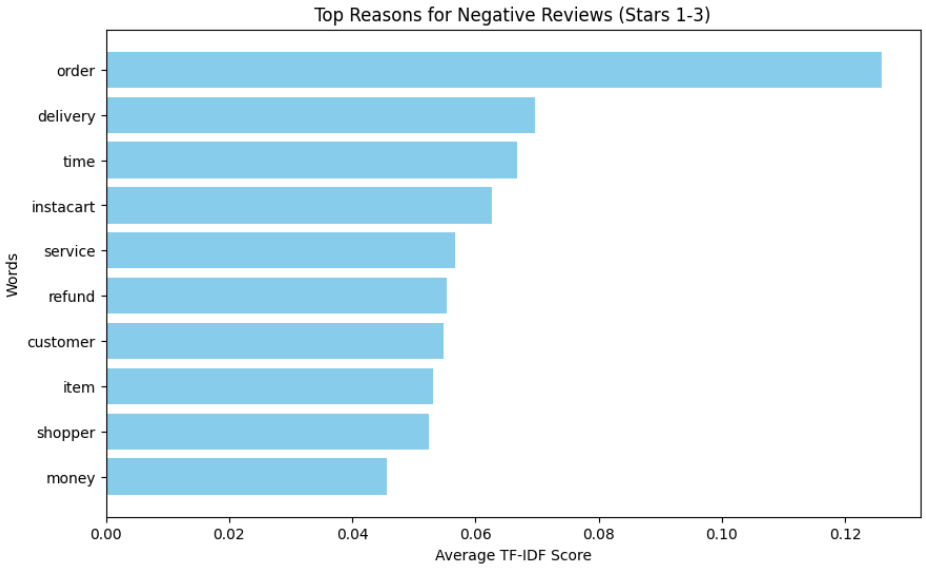


Chart 2

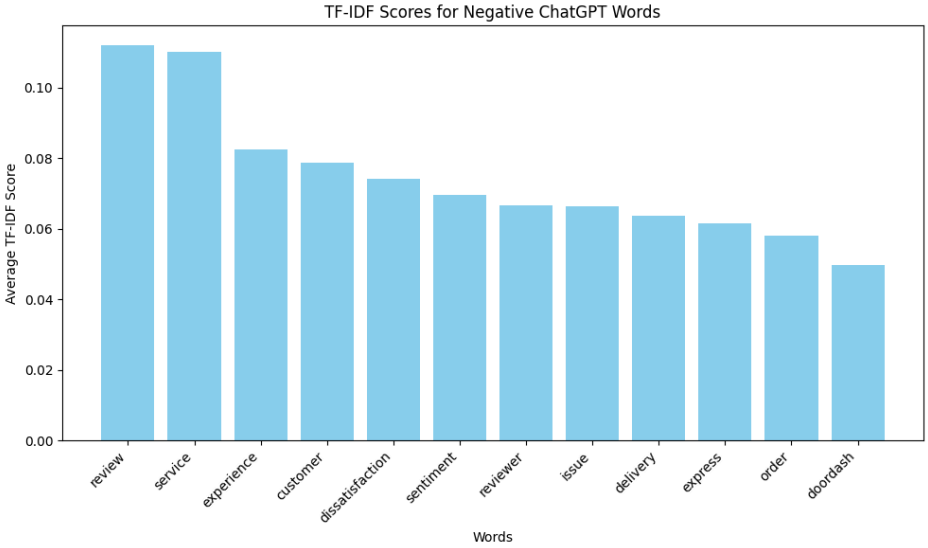


Chart 3

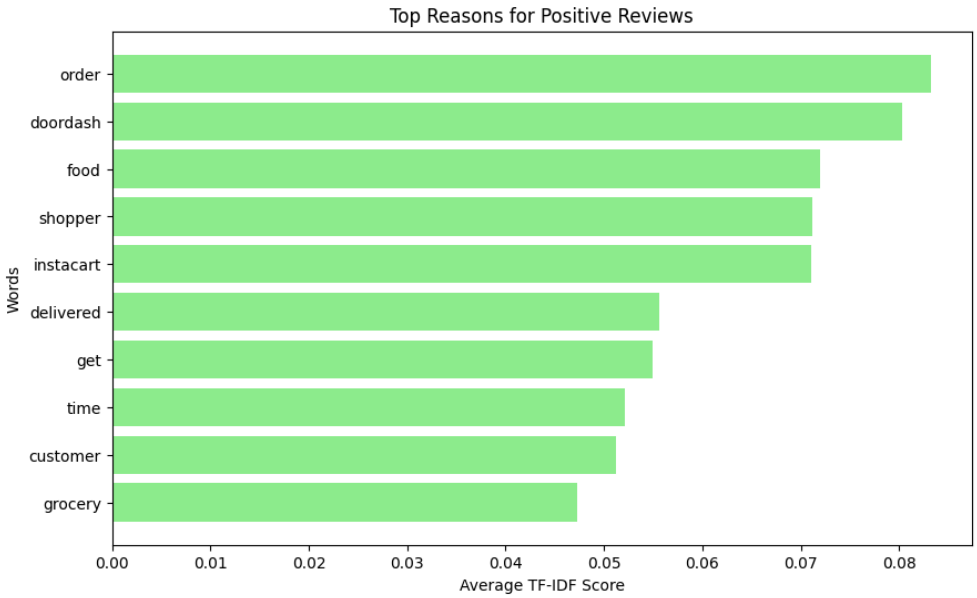


Chart 4:

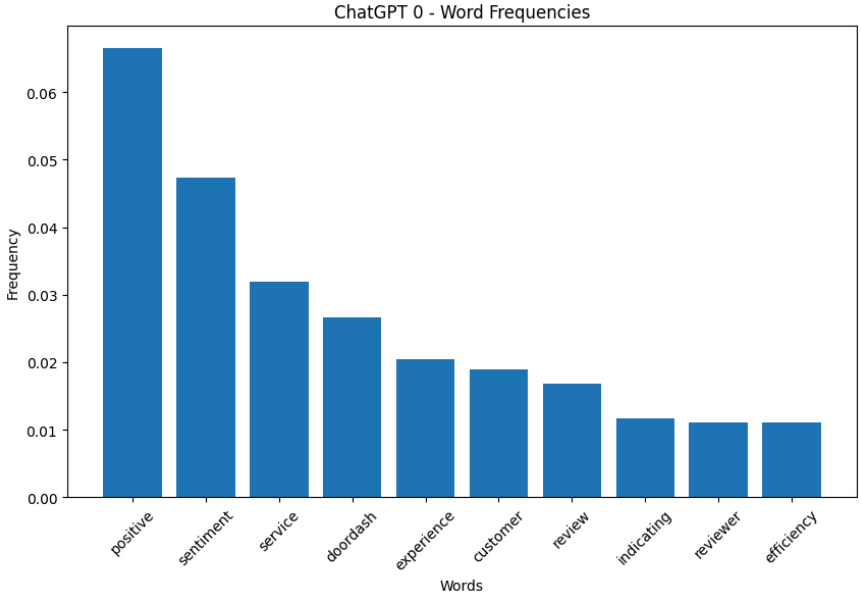


Table 3:

Model	Interpretability	Efficiency	Robustness	Context
ChatGPT	Low	Low	High	High
Hybrid	Low	Low	High	None

(VADER & Native Bayes)				
VADER	High	High	Moderate	None
Naive Bayes (NB)	Moderate	High	Moderate	None
Bag-of-Words x Native Bayes	Moderate	High	Low	None

Table 4

ChatGPT	Corpus	VADER	Native Bayes	Hybrid
A: 98.85%	A: 74.1%	A: 57.21%	A: 72.09%	A: 76.92%
P: 93.75%	P: 72%	P: 33.61%	P: 51.97%	P: n/a
R: 93.75%	R: 74%	R: 78.85%	R: 72.09%	R: n/a
F1: 93.75%	F1: 73%	F1: 47.13%	F1: 60.40%	F1: n/a

Error Analysis
Table 5

Review Content	ChatGPT Incorrect Response	Analysis
"""The driver was the only thing that was impeccable. Doordash itself besides that. The store forgot an item and doordash refused to refund even though I paid for it."""	Negative. While the reviewer appreciates the driver's effort, the overall sentiment is negative due to DoorDash's refusal to issue a refund for a missing item.	Positive for the company sense - negativity towards customers
The bigger the tip, the quicker you get your food...The drivers pick the highest priced orders first.	Neutral Sentiment: This review comments on the practice of tipping affecting delivery speed, with a neutral to slightly critical tone regarding the influence of tips on service prioritization.	Positive - unsure of what the message was directed towards

Data Analysis

Overall, the data of the reviews collected had an average score of 1.3 stars. Of these reviews, More than 20% have had on-time Delivery and more than 68% had accurate undamaged orders. Moreover, the refunds and returns process needs improvement. Most of the good 5-star reviews have come from about a year ago while the majority of the negative reviews have been prevalent in the beginning of its launch and the end. Despite the high rating of Instacart shown on the App Store, there are about the same amount of bad reviews as good reviews. Considering the high rating of Instacart, Doordash, and Shipt in the App Store and the contrasting number of

good reviews, it seems obvious that the App Store is potentially purchasing good reviews and/or faking their high rating.

Textblock has a slightly better accuracy at (67.14%) and a significant precision growth at (25.00%). It is a stronger recaller (recall rate of 93.75%) than many other tools suggesting its outstanding performance in capturing negative emotions. It follows that, TextBlob being most useful in situations where being able to capture as many negative critiques as possible wins out over the numbers of false positives considered moderate, is hinted at. This feature of the tool deserves a special mention; such a high recall rate might be regarded as very admirable, but there is a scope for improvement of the trade-off between precision and recall.

Unfortunately for the hybrid algorithm, it was able to conduct the sentiment analysis, but it was missing its analysis on any positive reviews. After further investigation, I found that the Hybrid model had negative answers where the two algorithms agreed, but not the positive reviews. Therefore, it wasn’t able to calculate the scores for Recall, F1-score, and Precision.

The conversation with ChatGPT performed with maximum accuracy, 98.85% to be precise, had both precision and recall at 93.75%. This is a perfect portrayal of the right balance of discriminating against negative feelings and correcting false alarms. ChatGPT's effectiveness suggests that it gained a better formal grip of the language that can remain more conclusive in the sense when it comes to analyzing the sentiment. It is the accuracy and especially recallability of the sentiment analysis that make such a tool a high-class device and an invaluable asset offering reliability of analysis which is vital for precise information interpretation.

TF-IDF Analysis:

TF-IDF analysis showed that words like order, delivery, and service were increased in the negative reviews and they reflected the fields such as customer service which laid the reason for customers to criticize. Numerous studies conducted in this field have demonstrated the link between the difference between negative and positive sentiments with emerging thematic disposition. Subjects in negative reviews have been found to be more concerned with rate of failure services while the happy clients focus on satisfactory service experiences. This ambivalence contrast emphasizes the role played by those service elements in the customer's overall perception of the service interaction and also introduces the use of TF-IDF in identifying focus areas needed to be improved.

Error Analysis

Sentiment analysis tools had excellent features but demonstrated limitations, particularly, when it came to complex scenarios such as sarcasm, mixed sentiment, and technical (specialized) vocabulary, the error analysis section discussed. It implies the delicate job of feeling human in automating the sentiment analysis since the complicated nature of the language the human expresses may not be interpreted with all the sophistication by some advanced algorithms.

On this era of all things digital, and when opinions are everywhere, feedback can be instant, deciphering just the meaning of words has never been more important. It not only connects the dots but it also draws a line between data and valuable insights, between customer feedback and

more quality service. Akin to that, the feelings being cut across in our social communications and conversations bring them in the picture of achieving unity, empathy, and finally, progression.

This research equally illustrates tech capacity/limitations of sentiment analysis machines and extends a metaphor for the interconnected issue of deciphering the witting and losing road of human emotions and views (both, on- and offline) by different systems.

This tool reads customer reviews to see if they are positive, negative, or neutral. This helps companies see how people's feelings change over time and what's popular.

Also, ChatGPT looks at details in reviews, like their content, feelings, product ratings, when they were written, and other info. This helps companies see what makes people act a certain way. This helps companies make better choices with their products, marketing, and how they treat customers.

Implications & Limitations

Through the analytics, a picture of the problem is drawn with the sentiment analysis tools doing their bit to alter the digital communication and customer feedback scenario within a complex landscape. Ethical concerns seem to be raised, giving an indication of many different ways this technology could be used like violation of users' privacy, data security risks and negative review representations.

In spite of the skepticism and tons of studies, it can shed light on an alternative theory for ChatGPT in which it represents a potent tool in the context of understanding human emotions with fast response times and balanced precision and recall rates. Though, this says something

about the unlimited nature of the sentiment analysis as the tiny samples may provide no enough data to analyze the results and ability of humans to overcome machine errors.

The mentioned liabilities point out the problem of the automatic tools in which they cannot take into account the subtleness of the language like humor and mixed feelings. In addition there have been testing errors. Unfortunately, the reviews I've come across were skewed negatively, which affected the recall score and precision of some of the models due to increased sensitivity to negative reviews. Moreover, the reviews were manually assessed by only myself. Having done the study on my own to assess the polarity of the reviews, there is potential bias. As for the Hybrid Reviews, if I had more reviews to work with, I could've given the other score of the hybrid model. Finally, due to being unable to find any accuracy measures for the neutral reviews, I've chosen not to include it into my study. Sentiment analysis is the critical part of the surveys deciphering to the customer feedback and precautionary provides the optimal decision-making process, still we need to consider the ethics of it and the need for further refinement of technologies as the digital world evolves.

Future Directions

The scholars who are in charge of resolving sentiment analysis ethical matters are likely to keep the check on responsibility and just use these tools. This directly leads to a need to create complex regulations for big data protection, artificial intelligence security, and algorithmic transparency. Multisectoral partnering between ethics, law and technology will be a key component in creation of those frameworks. At the same time, the algorithms for sentiment analysis accuracy have been in a great demand to ensure their perfection. Nevertheless, some

shortcomings are ongoing in the undoubtedly successful procedure of understanding contextualized sentiment and hidden insults that many people often use in their utterances.

Researchers should cut across machine learning to ensure that grammar and the language never differs widely across multiple linguistic and cultural environments. This can be done by usage of deep learning models, natural language understanding algorithms and other language processing applications. Besides digital content diversifying, multimodal sentiment analysis that focuses on the exploration of the avenue is also seen as a promising new looking direction. The integration of these forms of data, such as text , images, audio, and video, enables researchers across a wide range of disciplines to access more nuanced levels of human feelings and understandings, which in turn enriches research in the area of emotion and perception. Such an interdisciplinary approach, by nature, sets the scene for a more comprehensive understanding of the complex dynamics of digital communication systems and the constant interplay between language, emotions, and context in determining the behavioral patterns and the outcome of the conversations.

Gatherered CSV File

Review,Detail, Category, Sentiment

Coding

Vader Sentiment Analysis

```
import csv
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import precision_recall_curve

import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')

def sentiment_scores(sentence):
    sid_obj = SentimentIntensityAnalyzer()
    sentiment_dict = sid_obj.polarity_scores(sentence)
    # Return the compound score which represents the overall sentiment
    return sentiment_dict['compound']

def analyze_csv_sentiments(csv_file_path):
    true_labels = []
    sentiment_scores_list = []

    with open(csv_file_path, 'r', encoding='utf-8') as file:
        csv_reader = csv.reader(file)
        next(csv_reader)  # Skip the header row
        for row in csv_reader:
            # Assuming each row has a column with text to analyze
            # and another column with true sentiment labels
            # Modify the indices if the text and labels are in different
columns
            sentence = row[0]
            true_label = int(row[1])  # Assuming sentiment labels are 0
(negative) or 1 (positive)
```

```
            true_labels.append(true_label)

            sentiment_score = sentiment_scores(sentence)
            sentiment_scores_list.append(sentiment_score)

    # Calculate precision and recall for different thresholds
    precisions, recalls, thresholds = precision_recall_curve(true_labels,
sentiment_scores_list)

    # Plot precision-recall curve
    plt.figure(figsize=(8, 6))
    plt.plot(recalls, precisions, marker='.')
    plt.xlabel('Recall')
    plt.ylabel('Precision')
    plt.title('Precision-Recall Curve')
    plt.grid(True)
    plt.show()

if __name__ == "__main__":
    # Path to your CSV file
    csv_file_path = 'C:\\Users\\ethan\\Downloads\\Review, Sentiment -
Sheet1 (2).csv'
    analyze_csv_sentiments(csv_file_path)
```

Native Bayes Sentiment Analysis

```
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_predict
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, confusion_matrix,
precision_score, recall_score, f1_score, precision_recall_curve
from sklearn.pipeline import make_pipeline
import matplotlib.pyplot as plt

# Step 1: Load your data
df = pd.read_csv('C:\\Users\\ethan\\Downloads\\reviews, sentiment -
Review, Sentiment - Sheet1 (2).csv')
```

```
# Assuming your CSV has columns 'review' for the review text and
'sentiment' for the sentiment

# Step 2: Preprocess the data (Basic preprocessing included in
Vectorization)
X_train, X_test, y_train, y_test = train_test_split(df['reviews'],
df['sentiment'], test_size=0.2, random_state=42)

# Step 3: Vectorize the text data
vectorizer = TfidfVectorizer(stop_words='english', max_features=10000)

# Step 4: Train the Naive Bayes Model
model = make_pipeline(vectorizer, MultinomialNB())
model.fit(X_train, y_train)

# Step 5: Evaluate the Model
predictions = model.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted')
recall = recall_score(y_test, predictions, average='weighted')
f1 = f1_score(y_test, predictions, average='weighted')

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
print("Confusion Matrix:\n", confusion_matrix(y_test, predictions))

# Step 6: Plot Precision-Recall Curve
y_scores = cross_val_predict(model, X_test, y_test, cv=3,
method="predict_proba")[:, 1]
precisions, recalls, thresholds = precision_recall_curve(y_test, y_scores)

plt.figure(figsize=(8, 6))
plt.plot(recalls, precisions, "b-", linewidth=2)
plt.xlabel("Recall", fontsize=16)
plt.ylabel("Precision", fontsize=16)
plt.title('Precision-Recall Curve', fontsize=18)
plt.grid(True)
plt.show()
```

Hybrid Model Sentiment Analysis

```
import nltk
nltk.download('vader_lexicon')
import pandas as pd
from sklearn.model_selection import train_test_split

# Load your dataset
df = pd.read_csv('C:\\Users\\ethan\\Downloads\\Review, Sentiment - Sheet1
(2).csv')

# Assuming your dataset has the columns "text" for the review and
"sentiment" for the labels
X_train, X_test, y_train, y_test = train_test_split(df['review'],
df['sentiment'], test_size=0.3, random_state=42)

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline

model = make_pipeline(TfidfVectorizer(), MultinomialNB())

# Fit the model
model.fit(X_train, y_train)

# Predict with the model
ml_predictions = model.predict(X_test)

from nltk.sentiment import SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()

# Apply VADER to each review and decide on positive or negative based on
compound score
vader_predictions = [1 if sia.polarity_scores(text)['compound'] > 0 else 0
for text in X_test]
```

```
import numpy as np
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score

# Get consensus predictions
consensus_predictions = np.where(ml_predictions == vader_predictions,
ml_predictions, -1)

# Where consensus predictions are -1 (indicating disagreement), use ML
predictions
hybrid_predictions = np.where(consensus_predictions == -1, ml_predictions,
consensus_predictions)

# If there are no positive predictions, choose ML predictions for positive
reviews
if not np.any(hybrid_predictions == 1):
    hybrid_predictions = ml_predictions

# Evaluate the hybrid model
print("Evaluation Metrics:")
print("Hybrid Model Accuracy:", accuracy_score(y_test,
hybrid_predictions))
print("Hybrid Model Precision:", precision_score(y_test,
hybrid_predictions))
print("Hybrid Model Recall:", recall_score(y_test, hybrid_predictions))
print("Hybrid Model F1 Score:", f1_score(y_test, hybrid_predictions,
average='binary'))
```

References

Al-Shabi, M. "Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining." IJCSNS 20.1 (2020): 1.

Annika E. (2023, October 19). The benefits (and issues) of AI for sentiment analysis. LinkedIn. https://www.linkedin.com/pulse/benefits-issues-ai-sentiment-analysis-annika-englund/

ChatGPT and Open-AI models: A preliminary review. (2023, May 26). MDPI. https://www.mdpi.com/1999-5903/15/6/192

"Biden's Labor Department Takes Aim at Uber and DoorDash with New Rule Aimed to Cut Down on Number of 'independent Contractors'." Fortune, 9 Jan. 2024, fortune.com/2024/01/09/biden-labor-department-rule-independent-contractors-gig-economy-uber-lyft-doordash/.

Diyasa, I. Gede Susrama Mas, et al. "Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob." IOP Conference Series: Materials Science and Engineering. Vol. 1125. No. 1. IOP Publishing, 2021

Duan, Wenjing, Bin Gu, and Andrew B. Whinston. "Do online reviews matter?—An empirical investigation of panel data." Decision support systems 45.4 (2008): 1007-1016.

Floyd, Kristopher, et al. "How online product reviews affect retail sales: A meta-analysis." Journal of retailing 90.2 (2014): 217-232.

Gesenhues, Amy. (n.d.). Survey: 90% Of Customers Say Buying Decisions Are Influenced By Online Reviews. Martech. https://martech.org/survey-customers-more-frustrated-by-how-long-it-takes-to-resolve-a-customer-service-issue-than-the-resolution/

Hartmann, Jochen, et al. "More than a feeling: Accuracy and application of sentiment analysis." International Journal of Research in Marketing 40.1 (2023): 75-87.

Hu, N., Bose, I., Koh, N.S. and Liu, L. (2012) 'Manipulation of online reviews: an analysis of ratings, readability, and sentiments', Decision Support Systems, Vol. 52, No. 3, pp.674–684 [online] https://doi.org/10.1016/j.dss.2011.11.002.

Hu, Nan, Ling Liu, and Jie Jennifer Zhang. "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects." Information Technology and management 9 (2008): 201-214.

Impact of COVID-19 on online grocery shopping discussion and behavior reflected from Google trends and geotagged tweets. (n.d.). PubMed Central (PMC).

Iosifidis, Vasileios, and Eirini Ntoutsi. "Large Scale Sentiment Learning with Limited Labels." *Proceedings of KDD '17*, Halifax, NS, Canada, 13-17 Aug. 2017, Leibniz University Hannover & L3S Research Center. DOI: https://doi.org/10.1145/3097983.3098159.

MacDonald, M., 2018. Why online store owners should embrace online reviews, Available at: www.shopify.com/blog/15359677-why-online-store-owners-should-embrace-online-reviews.

Marr, Bernard. (2023, May 30). 10 Amazing Real-World Examples Of How Companies Are Using ChatGPT In 2023. Forbes. https://www.forbes.com/sites/bernardmarr/2023/05/30/10-amazing-real-world-examples-of-how-companies-are-using-chatgpt-in-2023/?sh=2c7204451441

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113.

Nytimes.com. (2020, November 30). The New York Times - Breaking News, US News, World News and Videos.

Online shopping statistics: Ecommerce trends for 2023. (2023, August 4). Tidio. https://www.tidio.com/blog/online-shopping-statistics/

Orozco-Arias, Simon, et al. "Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements." Processes, vol. 8, no. 6, 27 May 2020, http://www.mdpi.com/2227-9717/8/6/638/s1.

Qi, Yuxing, and Zahratu Shabrina. "Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach." Social Network Analysis and Mining 13.1 (2023): 31.

Ravi, K. and Ravi, V. (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications', Knowledge-Based Systems, November, Vol. 89, pp.14–46 [online] https://doi.org/10.1016/j.knosys.2015.06.015.

Shaeeali, N. Sakinah, Azlinah Mohamed, and Sofianita Mutalib. "Customer reviews analytics on food delivery services in social media: A review." IAES Int. J. Artif. Intell 9.4 (2020): 691.

Sopher, E. (2021, March 19). "Instacart's harsh ratings system hurts grocery delivery people like me". Vox. https://www.vox.com/first-person/22338325/instacart-grocery-delivery-ratings-system

Srivastava, Gautam. "A study on reviews of online grocery stores during COVID-19 pandemic using sentiment analysis." International Journal of Logistics Economics and Globalisation 9.3 (2022): 205-222.

TAYLOR SOPER. (n.d.). COVID-19 crisis sparks 'inflection point' for online grocery — and huge revenue for Amazon. GeekWire. https://www.geekwire.com/2020/analyst-covid-19-crisis-sparks-inflection-point-online-grocery-huge-revenue-amazon/

Watson, Forrest, and Yinglu Wu. "The impact of online reviews on the information flows and outcomes of marketing systems." Journal of Macromarketing 42.1 (2022): 146-164.

Wiessner Daniel. (2024, January 9). Biden administration issues rule that could curb 'gig' work, contracting. https://www.reuters.com/world/us/biden-administration-issues-rule-that-could-curb-gig-work-contracting-2024-01-09/#:~:text=Reclassifying%20independent%20contractors%20as%20companies,lost%20income%2C%20the%20group%20said

Wu, Nemin, and Lan Mu. "Impact of COVID-19 on online grocery shopping discussion and behavior reflected from Google Trends and geotagged tweets." Computational Urban Science 3.1 (2023): 7.

Wu, Tianyu, et al. "A brief overview of ChatGPT: The history, status quo and potential future development." IEEE/CAA Journal of Automatica Sinica 10.5 (2023): 1122-1136.