DETECTING FAKE REVIEWS IN E-COMMERCE USING MACHINE LEARNING: A CASE STUDY ON AMAZON

[1]**Omar Daraz** [2]**Muhammad Muneeb ur Rehman**

[1]Lahore Garrison University, Lahore Pakistan
[2]Independent Researcher
Email : munibgcu@gmail.com

Abstract
This paper explores the feasibility of using machine learning algorithms to detect fake reviews on online e-commerce platforms. The recent overreliance of consumers on online reviews to gauge the quality and desirability of e-commerce products has prompted these e-commerce brands to create fake, misleading reviews to shape customer narratives and perceptions about their products. This paper identifies current gaps in the mechanisms employed to filter out fake online reviews and analyzes how Amazon uses emerging methods like behavioral metrics, text analysis, and advanced neural networks to detect fake reviews. Given that a large enough dataset is used, these methods vastly outperform rule-based systems, showing that machine learning methods have significant potential in countering market and narrative distortion in the e-commerce world.

INTRODUCTION

Online e-commerce platforms have grown exponentially in the last couple of decades, with the internet revolution incentivizing businesses to supplement their physical businesses with a strong digital presence. Online stores offer unique benefits that enable businesses to cut massive costs. Since businesses no longer require physical outlets that are costly and difficult to maintain, they can increase their overall business revenue with digital stores. The barriers to entry for starting an online e-commerce store is also much lower than a physical store. Moreover, the scalability and various methods of online outreach mean that these stores are growing much faster than a lot of purely physical stores. Amazon has grown to be the largest e-commerce platform in the world, consistently expanding its inventories, markets, and workforce.

To encapsulate the experience and quality of their products, online e-commerce platforms rely on mechanisms such as product reviews and customer testimonials. Since customers do not have access to the physical product before buying it, unlike in physical stores, they rely on the opinions of people that have previously purchased the product and reviewed it on the platform. Reviews and testimonials have rapidly become the largest

and most important metric of a brand's trustworthiness and reliability.

Given the rise of e-commerce platforms, however, brands have opted for deceptive tactics and unfair business practices to drive consumer engagement and bypass fair growth. Several e-commerce platforms have been accused of pushing fake, pre-written, and scripted reviews on their platforms to mislead users and convince them of the quality of their products. As the growth of e-commerce platforms continues, conventional detection models are lagging behind in terms of quickly and accurately identifying and issuing warnings of fake reviews.

Instead of conventional text-based approaches, several independent detection agencies are moving towards machine learning algorithms and methods to quickly filter out fake reviews on online platforms and inform users of these reviews. These processes follow a standard approach, where they first collect data, pre-process it using tokenization, lemmatization, metadata preprocessing, and behavioral preprocessing. Then they move to feature engineering, model training, evaluation metrics, and real world deployment. This paper will be exploring how the world's largest e-commerce platform, Amazon, uses machine learning to filter and ban fake and misleading reviews.

Literature Review

Cui, M., Jin, Y. and Kwon, O. (2016) explore the details of unstructured data and how we need to invest in and bring attention towards systems of language analysis that deal with the growing mixed language usage that comes with a lot of unstructured data nowadays. Put simply, the authors argue a number of things. The internet has opened the floodgates of data and information to the public. We have more information available at our fingertips now than at any other point in history. An almost infinite amount of unstructured data brings along quite a few problems and issues. The rise of fake news, biased reporting, as well as the scientific need to classify and utilize this information to develop structures that help us use it in the best way possible are all common concerns. The authors talk about the need to analyze and study unstructured data on the basis of sentiment and intent, so as to establish a better understanding of consumer bases and understand how to detect content that is fake but is pushed in online reviews to boost the favorability of a specific brand [2].

Here are the aims/objectives that this discussion sets out to address:

1). To explore some of the existing techniques and strategies being used by mixed-code sentiment predictors and other machine learning algorithms and determine the benefits that these techniques have had for identification and classification of online content.

2). To understand the importance of machine learning algorithms for various business stakeholders, such as e-commerce businesses and social media.

3). To provide a real-world insight into how large e-commerce platforms such as Amazon are facing problems with fake review generation, and how they are utilizing machine learning to identify and flag fake reviews in an attempt to be more fair with the content that users are exposed to.

The first area of literature this paper addresses is the inability of conventional NLP (Neuro-Linguistic Programming) technologies to deal with dependency parsing, word-level language identification, semantic processing, machine translation, and part-of-speech tagging in fake reviews on e-commerce platforms. NLP systems struggle with properly analyzing code-mixed text, and they often fail to adequately analyze English-based phonetics and anglicisms because most of them rely on single-language resources to study multilingual cases, making them an unreliable mechanism for accurately detecting whether online reviews are false or not. The second area of this research paper

provides a general overview of the end-to-end process employed by Amazon in an attempt to counter fake reviews on the platform.

Bellegarda, J., Apple Inc (2014) writes that sentiment prediction from textual data may be defined as a categorization, in simple terms. This categorization involves the determination of words, sentences, or even larger pieces of writing into states, such as one of three states: positive, negative, or neutral. The reason why sentiment prediction is performed is to attach an emotion or a feeling with a collection of words. The word "angry" represents a feeling of emotional turbulence, distress, and dissatisfaction. An important thing to realize here is that there is a very small subject set of words that clearly depict a common emotion. Words such as happy, sad, angry, and so on. But beyond that, the subjective nature of language and the variation that comes with using even the simplest of terms means that NLP systems need to be able to efficiently predict the semantic aspect (the meaning of the words) of a variety of scenarios on the basis of context, one that learns to label pieces of text with emotion after analyzing the potential for a context-based variation in meaning. There are a host of scenarios where the usage of such predictors to gather data from non-verbal cues is important as well [4].

Ain, Q.T., Ali, M., Et al. (2017) describe how users on social media, internet forums, review sites, and blogs generate enormous heaps of data, most of which carry implicit sentiment that is incredibly useful for vendors, creators, media figures, politicians and so on. The authors talk about how unstructured data is often lost on social media platforms if effective NLP techniques to analyze this data aren't employed in a timely manner. Moreover, they talk about deep learning as a foundational model to use for fake review identification. They referred to some models and strategies that have been applied to sentiment prediction problems

and laid out a comprehensive tabular representation of the effectiveness of various models. The authors describe how the use of machine learning algorithms is understood to be a multidisciplinary field because it involves work from a lot of fields. Examples of these fields are artificial intelligence, computational linguistics, information retrieval, natural language processing, and semantics. The classification for the approaches of sentiment analysis can be done in three extraction levels: a document level, a sentence level, and an aspect level. These approaches usually define the length of each piece of language under study. The authors also pointed out that sentiment analysis depends on two kinds of techniques: machine learning-based techniques and lexicon-based techniques [5].

Analysis for fake reviews may be performed by implementing different models such as deep learning models. Some of these models are RNN (recursive neural network), DBN (deep belief networks), DNN (deep neural networks), CNN (convolutional neural networks), and RNN (recurrent neural networks). Most researchers that have used these methods have used more than one of these methods in their research, which gives us useful insight into hybrid models and new data analysis techniques [1][2][3].

Deep learning networks have been a consistent factor in a large number of previous and ongoing sentiment prediction and analysis systems. Also, a lot of researchers have used Support Vector Machines (SVM) and Multinomial Naive Bayes classifiers (MNB) as the central models in their respective experimentation towards developing and improving sentiment prediction systems [6].

A number of machine learning models, such as Naïve Bayes and Random Forest models performed analysis on product review data [8]. Some work in this field included experiments with mood classification on blog posts. One of the researches also deals with

the review of aspect-based opinion polling from unlabelled free-form textual customer reviews without requiring customers to answer any questions

Volkova, S., Wilson, T. and Yarowsky, D. (2013) showed how different demographics of people, categorized according to age, gender, or other factors showed clearly observable variances with the sentiments they associate with language in their reviews and general opinions on social media. They chose to focus on Twitter as a medium of observation as well, because of its volume, diversity, and dynamic nature. Their results showed significant relative F-measure improvement over the gender-independent baseline 1.5% and 1% for Russian, 2% and 0.5% for Spanish, and 2.5% and 5% for English for polarity and subjectivity classification, as cited by their paper. What is most pertinent to regard as a component of this study is that they uncovered basic variances in the way male and female reviews associate sentiments with various linguistic connotations. For instance, their research cites that they discovered how the probability of women associating the term "weakness" with a positive sentiment was reasonably higher than that of men. For example, the study found that a large number of expressions of text on Twitter by female users used the word weakness in a positive sentiment ("Blueberry ice cream is my weakness!). However, the study also found that male Twitter users were more likely to use weakness in a negative connotation ("We must get rid of weakness to be real men"). The subjectivity that comes associated with language and expression is incredibly deep and scenario-based. But this study affirmed the idea that demographics such as gender have a lot to do with the sentiments that users display on social media sites like Twitter. Arguably, this finding represents the idea that fake review predictors needed to be modeled keeping in view the differences in expression that various cultures and

demographics of people might display in normal, real reviews. Only when these machines understand the variances in real reviews can they compare these differences to fake reviews and find patterns commonly and consistently present in fake reviews. The practical manifestation of this issue can work its way into issues of group identity and how a particular group (race, nationality, etc.) of people respond to review and how these responses have the tendency to portray a variety of sentiments that are brought about not by individual differences but by differences in background, differences of identity, differences of cultural evaluation [7]. These methods and techniques of analysis represent a great deal of scientific depth on their own. While aiming to develop robust and accurate fake review prediction devices, several researchers experimented with deep learning, lexical analysis, and various other data collection and analysis frameworks. These techniques simply by virtue of being employed contribute to scientific advancement one way or another. As conventional NLP techniques rapidly lose out on bargaining power because of their visible shortcomings (often a lack of accuracy, difficulty in accurately assessing the motive and emotion behind messages in mixed languages), machine learning prediction devices based on multilingual resources are rapidly achieving prominence and taking center stage as some of the best scientific tools for collecting, mapping out, categorizing, storing, analyzing, and drawing conclusions from information [8][9][10][11].

Even though earlier research has mapped our different mechanisms of fake review detection, most of this research revolves around traditional rule based systems, where the focus is on capturing patterns, such as repeated use of superlatives etc. However, as fake reviews started to grow, these traditional systems have been easily avoided by fake review systems. Secondly, even though some research exists on how within NLP, Support

Vector Machines and Random Forests can detect fake reviews by relying on metrics such as review length, n-grams, and sentiment prediction, these methods have not been adequately discussed in terms of how real world e-commerce platforms are actively using them to filter out fake reviews on their own platforms, which is the gap this study seeks to fill, by analyzing the ways in which Amazon employs machine learning methods to successfully eliminate fake reviews on the platform. Secondly, even though deep learning models are increasingly being used by individual researchers on small case studies of content, where they are using Recurrent Neural Networks and Convolutional Neural Networks, these case studies are not analyzing how e-commerce platforms themselves have actively incorporate machine learning models in their detection, which is the most useful method to understand the scale of these algorithms in the real world.

The broad gaps and challenges are:
Most models perform well on one platform but their performance fails on different types of fake reviews
Limited work has been done on real-time detection by large e-commerce brands. ● Handling more complex fake reviews remains a challenge.

Amazon's Utilization of Machine Learning Mechanisms to Filter Fake Reviews

Historically, Amazon has had to contend with various content attacks and fake reviews because of its stature as a platform. The platform has been susceptible to attacks from fake-review brokers that monetize fake reviews and sell them to individuals and agency accounts. Given the recent increase of fake review mechanisms worldwide, Amazon has adopted various mechanisms to consistently keep filtering out harmful and misleading reviews from the platform.
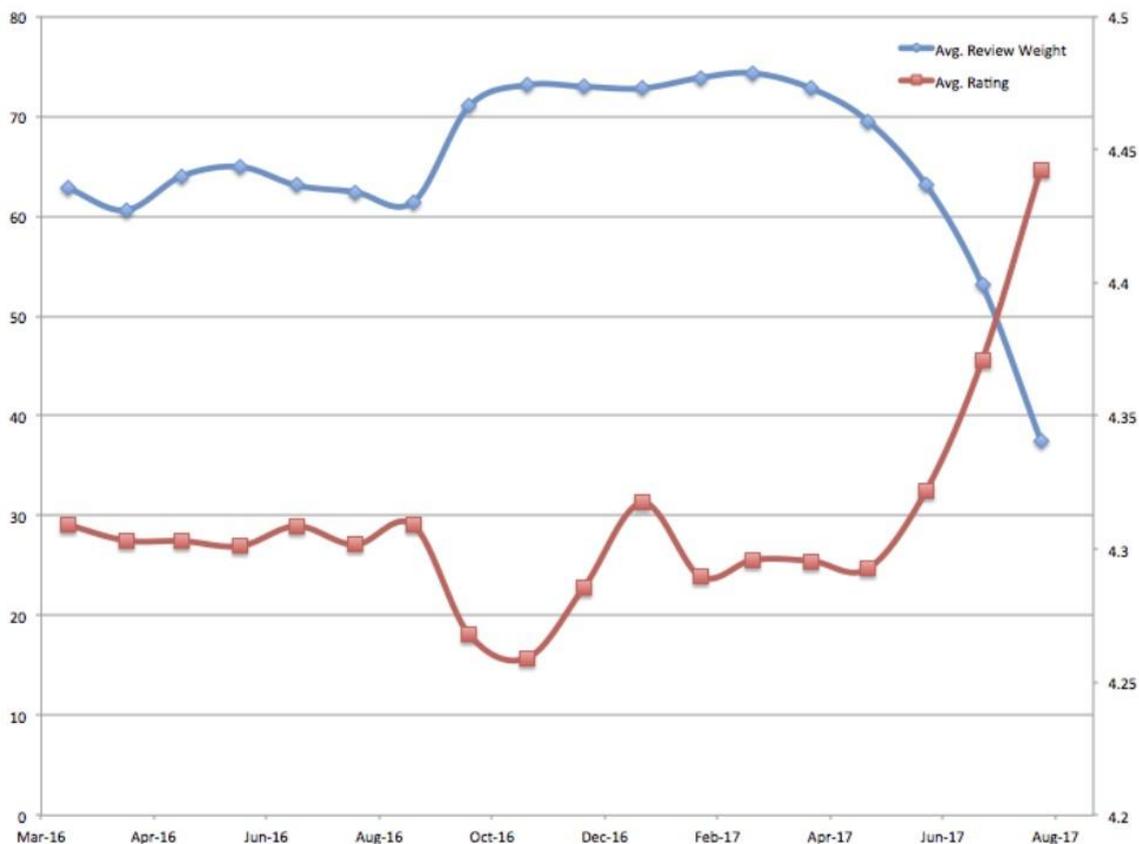
Source: Review Meta; The increase of fake reviews on Amazon, 2017 study

General Mode of Operation: Amazon has a set metric of indicators to determine if Amazon uses artificial intelligence (AI) to analyze the review for known indicators that the review is potentially fake or not. Once the review is judged to be fake, it is blocked and removed. Depending on the severity of potential harm, Amazon can pursue actions such as blocking accounts, taking away ability to review, and even legal action. Over the last two years, Amazon has observed and blocked more than 200 million fake testimonials on the online platform.

Machine Learning Models: Amazon utilizes cutting-edge artificial intelligence techniques to prevent vast numbers of fake reviews, manipulated ratings, and other fraudulent activities from reaching customers. By leveraging machine learning, the system analyzes a wide range of data, such as a seller's ad spending (which could increase reviews), customer-reported abuse, and suspicious behavior patterns in review histories. Sophisticated language models and natural language processing tools are employed to identify anomalies, which might suggest that reviews were incentivized through gifts or other forms of compensation. Additionally, deep graph neural networks help map and examine complex relationships between users and behaviors, identifying coordinated fraudulent activities for further investigation.

Getting into a deeper investigation of how these models work gives us the following insights:

1. Analysis of Various Data Points:
Machine learning models examine hundreds of thousands of data points across multiple reviews to identify a commonality in the type of content being posted on reviews.

Reviewer behavior: The models firstly review the frequency behind a user's review posting. Using common discernable red flags, such as a user posting a huge amount of reviews in a very small amount of time, or a user writing too many different reviews for completely unrelated products that seem unnatural, the system flags these behaviors as suspicious.

Example: A reviewer posts 30 five-star reviews within 12 hours, reviewing completely unrelated products like kitchen utensils, tech gadgets, and clothing, that have very little in common. Such behavior raises suspicion, as genuine customers typically do not review multiple products in such a short time span, especially across varied categories. Amazon's machine learning models flag these patterns and may mark the account as flagged, which will then be reviewed further.

Content review: These algorithms then analyze the type of language used in reviews using various mechanisms such as sentiment analysis, where language authenticity is checked by the amount of seemingly unnatural language used in a review. Cross-referencing multiple reviews to check for reused templates and copied content that appears in a large number of reviews.

Example: A review for a mobile phone may say: "Great product! Amazing quality, very satisfying!" If the algorithm catches this generalized feedback being repeated over and over for several products by the same reviewer or the same group of reviewers, this can alert Amazon's algorithms. These mechanisms analyze language patterns, the use of superlative texts, and more. It also checks for copy pasting, which commonly happens across fake review networks.

Purchase history: The difference of reviews between people who purchased the product and then gave reviews compared to those who never bought the product and gave reviews is then accounted for. Verified reviews from individuals who purchased the product are given a higher weightage than unverified reviews posted consistently by individuals who have not bought the respective product.

Example: A user posts a review for a laptop but did not purchase the item on Amazon.

Amazon gives more weight to reviews marked as "Verified Purchase," reducing the influence of reviews from users without purchase history. If a reviewer frequently posts feedback on items they haven't bought, this pattern is flagged as suspicious and could lead to their reviews being demoted or removed.

Networks: The interactions between sellers and reviewers are studied holistically. If an account or multiple accounts are consistently found to post overwhelmingly positive reviews for the same seller or the same product, it may be flagged as an issue. This coordination can be spotted, further analyzed, and then confirmed.

Example: A bunch of user accounts may leave a lot of collective reviews on a sole seller's listing, usually spamming within seconds or minutes of each other. Amazon's ML algorithms utilize graph theory to uncover connections between these accounts, such as timing of the review, past activity on the same set of products, and a group of shared IP addresses. Once these networks are fully mapped out, the algorithms are able to discern and identify "review farms" that work in close coordination to manipulate ratings and steer user intent. These networks are then closely monitored and the ratings are removed.

Engagement: If reviews receive multiple likes and engagement rapidly in a very short amount of time, it may suggest manipulation and coordination, so the algorithms check for this as well.

Example: A recently posted review suddenly garners an unusually high number of helpful votes in a short period, disproportionately compared to other reviews for the same product. This can indicate manipulation, where groups of users coordinate to vote up fake reviews to increase their visibility. Amazon's algorithms detect such spikes in engagement and flag them for investigation

1. Machine Learning Techniques and Tools:

Amazon employs various machine learning techniques to identify and filter out harmful content [18].

Natural Language Processing: Most large machine learning models for content utilize NLP, which is a breakthrough in recent years in analyzing text patterns, identifying how unnatural or repetitive they are, and more. Various aspects of the review, including the style of expression, the potential sentiment expressed, and the nature of how repetitively a review has been pasted across the platform, all are taken into account to determine if a review is genuine or if it has been generated by bots [15][16].

Behavior Analysis: Behavior tracking is also done as a part of these systems. The mechanisms of reviewers and their interactions are studied to be able to identify any potential misuse of the product review feature [14].

Graph Modeling: Intricately detailed graphs are created to understand the relationships between sellers, products, and consumers by making a "web" of their interactions. Graph theory is then employed in order to track how suspicious activity happens and how it moves from source to source, such as large groups of users leaving behind chains of fake reviews [13].

For instance, in 2021, Amazon pursued legal action against various group admins on Facebook that were offering product giveaways in exchange for fake positive reviews. When a deep graph neural network was created, it helped the company identify clusters of profiles and accounts that were associated with these schemes by taking a close look at shared behaviors, language used in reviews, and more.

Reinforcement: The learning algorithms are constantly updated through feedback looped to them. The growth of fake reviews comes with increased complexity in detecting these patterns, meaning that these models need to

keep getting better in order to spot these reviews and counter them.

AI Comparisons: AI-driven comparisons are consistently performed on different reviews to quickly identify and filter out any misleading, harmful, and bot-generated content and then drive product insights. Clustering techniques, such as K-Means, are used to identify fake product reviews in order to give consumers fair context on how a product is actually performing, how well it is perceived by people, and so on [17].

Human Moderation and Feedback

Machine learning patterns may need some degree of human oversight to be able to work effectively and counter the growing number of fake review frauds on these platforms.

Supplemented with human reviews: Reviews that have been preliminarily flagged may be handed to a team of human reviewers that verify the review and find out if there are any additional metrics they may need to check the review for, such as paid campaigns and more.

Customer reports: Individuals have the ability to report other users on the platform if they notice suspicious and unusual behavior. User reports on different reviews are also used to train AI models to be able to spot fake reviews in a more holistic way, while improving its understanding of how real people interact with content.

Statistics & Action

According to public data published by Amazon, by 2022, they processed more than 1.5 billion reviews and ratings to check for authenticity. In another preemptive effort Billions of reviews scanned: As of 2022, Amazon processed over 1.5 billion reviews and ratings posted by 125 million users globally . Moreover, 250 million fake reviews have been flagged and blocked before being posted to the platform. Since the machine learning algorithms operate in real time, it is possible to keep up with the pace at which these fake reviews are posted.

Amazon takes further action against organized fake review efforts. The company has filed lawsuits against thousands of fraudulent review groups, particularly targeting Facebook groups offering compensation in exchange for positive reviews.

Ongoing Challenges

Despite these robust efforts, 27.6% of reviews on Amazon are estimated to be fake, highlighting the continuous arms race between Amazon's detection tools and increasingly sophisticated fraudsters. As fake review tactics evolve, including AI-generated reviews and bots, Amazon continues to invest in improving its machine learning algorithms.

Fake views

Amazon products, before and after end of latest fake-review promotion on Facebook
Seven-day moving average

Source: "The Market for Fake Reviews" by Sherry He, Brett Hollenbeck and Davide Proserpio, working paper 2020

Recent Innovations

The utility of AI in multifarious domains and sectors has prompted organizations around the world to consider the large-scale utilization of AI tools and features, such as chatbots, natural language processors, and sentiment analyzers. Similarly, the increasing recognition of AI tools around the world has prompted technology companies to integrate AI within their applications, driving innovation and creating value for users. Several large product companies around the world have been experimenting with AI driven features, such as personalized content that attracts consumers, relevant suggestions based on collected user data, behaviors, and other metrics, and much more. Amazon has also utilized various innovative methods to quickly counter fake reviews and misleading content.

For instance, Amazon introduced AI-generated product summary reviews that take trusted, approved reviews and turns them into snippets of useful content. The content for these snippets is always taken from verified reviews, which makes sure that credible feedback from people that have actually used the product can influence the AI-generated summaries.

Conclusion

Several individual entities and companies are innovating with machine learning technology in the e-commerce domain more than ever before. In these times, it is pertinent to study the application and specific usage of these ML algorithms in different e-commerce niches, faced with various types of fake reviews, and their specific utilization by large e-commerce platforms like Amazon. While the literature still has gaps in terms of broader and larger testing of machine learning features on huge amounts of e-commerce data, academia can lean on the corporate world and drive its research in this domain by learning from and researching the effectiveness of mechanisms that companies like Amazon are employing on their e-commerce pages in real time.

References

1) Jabbar, J., Urooj, I., JunSheng, W. and Azeem, N., 2019, May. Real-time sentiment analysis on E-commerce applications. In 2019 IEEE 16th international conference on networking, sensing and control (ICNSC) (pp. 391-396). IEEE.
2) Cui, M., Jin, Y. and Kwon, O., 2016. A method of analyzing sentiment polarity of multilingual social media: A case of Korean-Chinese languages. Journal of

Intelligence and Information Systems, 22(3), pp.91-111.

3) Turney, P.D., 2002, July. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

4) Bellegarda, J., Apple Inc, 2014. Sentiment prediction from textual data. U.S. Patent 8,682,649.

5) Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., and Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6), p.424.

6) Xing Fang, Justin Zhan. Sentiment analysis using product review data. Journal of Big Data 2015.

7) Volkova, S., Wilson, T., and Yarowsky, D., 2013, October. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1815-1827).

8) Ledalla, S., and Mahalakshmi, T.S., 2018. Multilingual Sentiment Analysis of Hinglish Tweets. Indian Journal of Public Health Research & Development, 9(12), pp.1627-1631

9) Anjaria, M., and Guddeti, R.M.R., 2014. A novel sentiment analysis of social networks using supervised learning. Social Network Analysis and Mining, 4(1), p.181.

10) Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

11) A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, IEEE Transactions on Knowledge and Data Engineering, VOL. 23, NO. 3, MARCH 2011.

12) Mukku, S.S., Choudhary, N. and Mamidi, R., 2016. Enhanced Sentiment

13) Classification of Telugu Text using ML Techniques. SAAIP@ IJCAI, 2016, pp.29-34.

14) He, S., Hollenbeck, B., Overgoor, G., Proserpio, D. and Tosyali, A., 2022. Detecting fake-review buyers using network structure: Direct evidence from Amazon. Proceedings of the National Academy of Sciences, 119(47), p.e2211932119.

15) Liu, W., He, J., Han, S., Cai, F., Yang, Z. and Zhu, N., 2019. A method for the detection of fake reviews based on temporal features of reviews and comments. IEEE Engineering Management Review, 47(4), pp.67-79.

16) Tabany, M. and Gueffal, M., 2024. Sentiment analysis and fake amazon reviews classification using SVM supervised machine learning model. Journal of Advances in Information Technology, 15(1), pp.49-58.

17) Elmogy, A.M., Tariq, U., Ammar, M. and Ibrahim, A., 2021. Fake reviews detection using supervised machine learning. International Journal of Advanced Computer Science and Applications, 12(1).

18) Elmurngi, E.I. and Gherbi, A., 2018. Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques. J. Comput. Sci., 14(5), pp.714-726.

19) Salminen, J., Kandpal, C., Kamel, A.M., Jung, S.G. and Jansen, B.J., 2022. Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, p.102771.