HIGH SCHOOL EDITION
Journal of Student Research

Sentiment Analysis to Identify Consumer Criticism of Artificial Intelligence

Aarav Mulinti[1] and Guillermo Goldsztein[#]

[1]Montville Township High School, USA
[#]Advisor

ABSTRACT

As artificial intelligence becomes increasingly integrated into various aspects of our lives, understanding consumer sentiment and criticism towards artificial intelligence technologies becomes pivotal for effective utilization. This study presents an approach to sentiment analysis aimed at identifying consumer criticism of artificial intelligence use in businesses. By harnessing clear analytics, administrators can enhance their understanding of consumer feedback and thereby improve artificial intelligence integration for better user experiences. Increased consumer satisfaction is important to overall business to consumer relationships and streamlined AI use will facilitate company procedures. Our methodology revolves around machine learning techniques, specifically utilizing four classifiers, Keras logistic regression, Naive Bayes, Support Vector Machine, and random forest regressor, alongside two numerical feature representations, Bag-of-Words and Term Frequency-Inverse Document Frequency. The results show that the Term Frequency-Inverse Document Frequency features combined with the random forest regressor yielded the strongest performance in identifying criticism, with F1 scores of 100 and 99 percent for no criticism and criticism, respectively.

Introduction

Artificial intelligence (AI) is being rapidly integrated in our daily lives, in education, work, or even to communicate. By harnessing the criticism behind today's technologies, AI can be integrated better in tomorrow's businesses. The most efficient way to extract consumer opinion is through sentiment analysis, which is a form of natural language processing capable of understanding how humans feel based on the way they speak, or communicate.

Especially with the rise of online consumer reviews, administrators using AI in their businesses need to prioritize understanding consumer recommendations from them. If given clear analytics, they can better understand where they can improve, rather than trying to interpret "good vs. bad" on a specific product or feature. Thus, this paper turns to machine learning, which can identify the underlying patterns of communication specifically with or without criticism in it. Rather than using generalized models to visualize sentiment analytics, where AI can be easily lost compared to product criticism, the model this paper proposes can be specifically used to understand the performance of AI in their services.

The overall study examines the best yet accessible methods of natural language processing for sentiment analysis, specifically modified to identify criticism, to propose an efficient understanding of consumers' opinions on AI.

Background

Author et al. (2019) finds that sentiment analysis is widely used in businesses. It has several tasks, including subjectivity classification, sentiment classification, opinion spam detection, implicit language detection, and aspect extraction. The core of sentiment analysis lies in the sentiment classification, where using it to find polarity on specific products or features is extremely useful to a company. Author et al. (2022) finds that sentiment analysis has advanced to find a wider range of emotions and also represent sentiment about specific topics of interest, which are usually features of a company or their products. These analyses are useful for marketing and product design, and it clearly holds important implications for the corporate world overall. A blog estimates that an increasing number of companies (80%) are going to be using sentiment analysis by 2023. (Author, 2023) Their analyzed data comes from a variety of sources, including specific consumer reviews or generalized communication (from Twitter, Instagram, and other social media) about the company and its operations.

In terms of examining weak areas of AI (noting that ChatGPT and similar tools are popular forms of AI in consumer services), several researchers (Author et al., 2023; Author et al., 2022) conducted comparison tests to assess ChatGPT-generated text and human-generated text. Additionally, Author et al. (2023) evaluated ChatGPT's bug-fixing performance using a variety of benchmarks, and Author et al. (2023) assessed ChatGPT's ability to analyze textbox questionnaires and transcribe think-aloud data due to the fact that it can offer recommendations for fixing incorrect source code. (Author et al., 2023; Author et al., 2023) Additionally, a number of other techniques were used in recent studies, including sentiment analysis and text analysis (Author et al., 2023; Author et al., 2022), to analyze the authenticity of ChatGPT responses. Author et al. (2023) also used researchers' sentiment to determine faults of ChatGPT. Another study found a way to determine barriers of AI implementation in business but with a Delphi study (extracting consensus) on a panel of experts. (Author et al., 2021)

In comparison, this paper prioritizes a model predicated on consumer interactions. Our model with consumers is important for business-to-business specificity, while the aforementioned studies do not consider them. This paper aims to add a new lens to the analysis of weak areas of technology like ChatGPT by adding to current sentiment analysis usage.

Dataset

The data used in this paper is from an online data source called Kaggle. (Author, 2023) The dataset itself has 217.622 samples of tweets that are about ChatGPT (maximizing the ability to understand criticism of AI since ChatGPT and similar tools are widely-used), and the content varies as there is no prerequisite for what these tweets are about. Due to the limitation of research done to identify specific criticism in these tweets, we've created an algorithm redefining the labels. Rather than use the three classifications of positive, neutral, or negative, this paper uses an organized list of phrases (Merriam-Webster, Thesaurus.com) that clearly indicate criticism in the content of the tweet. The algorithm created a new array of labels with a binary classification system: 0 for no criticism, and 1 for criticism. This paper uses two methods for numerical representations of features, which will be further explained in the next section. The dataset is eventually split 75% for training and 25% for validation.

Materials and Methods

An important piece of building this model is creating an accurate criticism-flagger with an accessible machine learning algorithm. To determine the best, this paper uses four types of classifier - Keras logistic regression, Naive Bayes, Support Vector Machine, and random forest regressor, all of which are easily accessible, allowing our results to be widely applicable and especially useful for smaller businesses. (Author, 2019) Each of these popular and accessible classifiers can process two types of numerical features representing the text, Bag-of-Words and Term Frequency-Inverse Document Frequency (TF-IDF). In total, 8 models were tested in Google Colab Notebooks.

Features

Bag-of-Words

There are two phases to calculating this version of features. The first is building a frequency list. This means creating a dictionary indicating how many times every word appears in each type of classification. Using that dictionary, the model can calculate “frequency scores” for every sample, with one score corresponding to each classification. The scores indicate a specific pattern for the model to learn, which is what it predicts its classifications on.

TF-IDF

Term Frequency measures how often a term appears in a document. It's calculated by dividing the number of times a term appears in a document by the total number of terms in that document. The idea is to emphasize terms that occur frequently within a specific document.

$$TF(t, d) = (\text{Number of times term t appears in document d}) / (\text{Total number of terms in document d}).$$

Inverse Document Frequency (IDF) measures the importance of a term across a collection of documents. It's calculated as the logarithm of the ratio of the total number of documents to the number of documents that contain the term. The IDF value decreases as the term appears in more documents, making common terms less significant.

$$IDF(t, D) = \log((\text{Total number of documents in the collection D}) / (\text{Number of documents containing term t}))$$

The TF-IDF weight for a term in a document combines the term frequency and the inverse document frequency. It gives higher weights to terms that have high frequency within a specific document (indicating importance to that document) while also considering the rarity of the term across the entire collection.

$$TF\text{-}IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

The TF-IDF representation assigns a weight to each term in each document, forming a matrix where rows represent documents and columns represent terms. This matrix can then be used as input for machine learning algorithms.

Classifiers

Keras Logistic Regression

Based on the training samples and the binary cross-entropy error, the Keras Logistic Regression model improves its predictions of the classification over time. The prediction is a probability of how likely the sample is to be 1 (which is indicating that the sample is criticism).

Naive Bayes

The Naive Bayes classifier produces a probability of the sample being a certain class, but the naive part indicates that it assumes the features are independent of one another, which enhances the decision-making process, especially in sentiment analysis.

Support Vector Machine

By processing the text as vectors, it aims to find the optimal hyperplane that separates the classes in a high-dimensional vector space.

Random Forest Regressor

By building decision "trees" during training and combining their predictions, it makes a more accurate and stable regression prediction.

Results

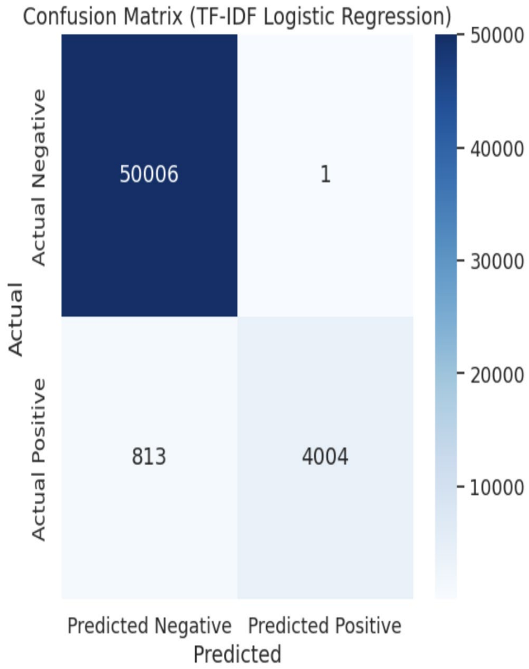
The below table indicates the results. This paper uses the F1 score, primarily because the model used an imbalanced dataset.

F1 Scores (%)	Keras Logistic Regression	Naive Bayes	SVM	Random Forest
Bag-of-Words (0)	97	93	97	97
Bag-of-Words (1)	62	44	62	65
TF-IDF (0)	99	96	100	100
TF-IDF (1)	91	10	98	99

Figure 1. All models F1 scores for both criticism (1) and non criticism samples (0)

Here's a closer look at each classifier's confusion matrix of the validation sets. They represent the true and false positives and negatives.

Logistic Regression



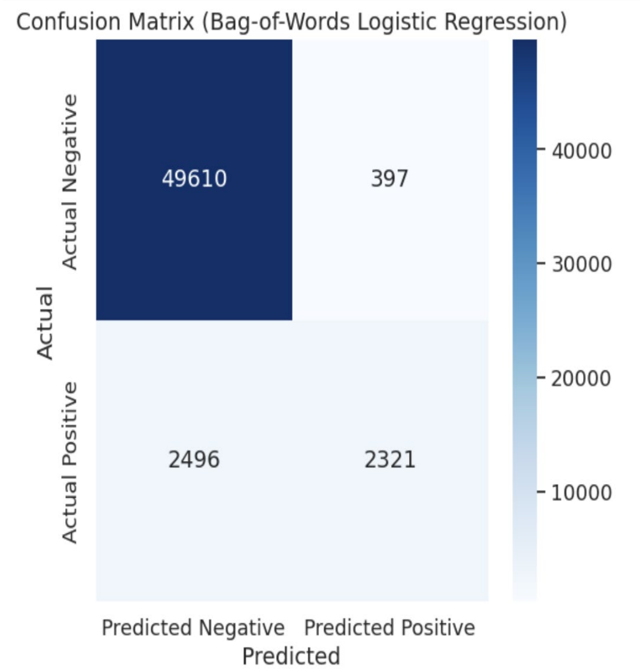
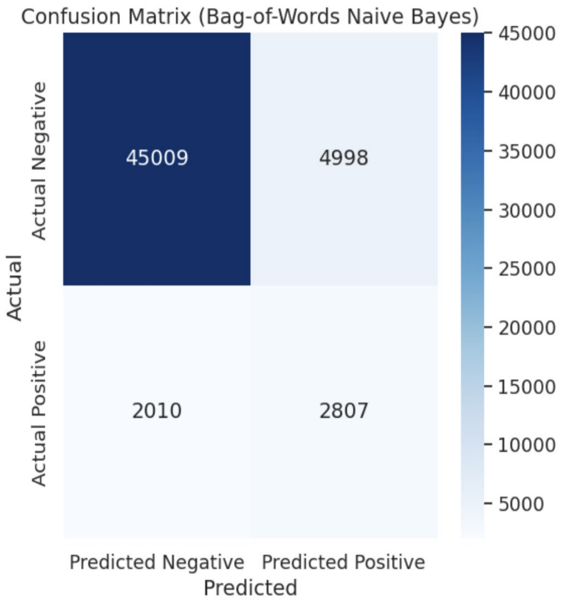


Figure 2. Bag-of-Words + Logistic Regression Results

Figure 3. TF-IDF + Logistic Regression Results

Naive Bayes


Confusion Matrix (Bag-of-Words Naive Bayes)
Actual
Actual Negative
Actual Positive
45009
4998
2010
2807
Predicted Negative
Predicted Positive
Predicted
45000
40000
35000
30000
25000
20000
15000
10000
5000

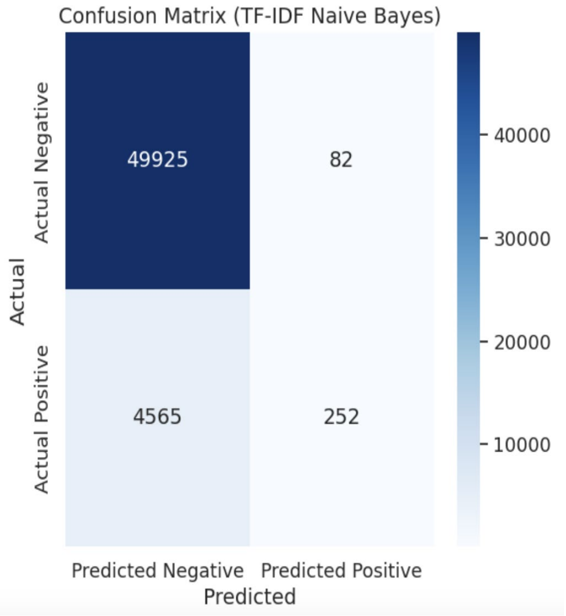
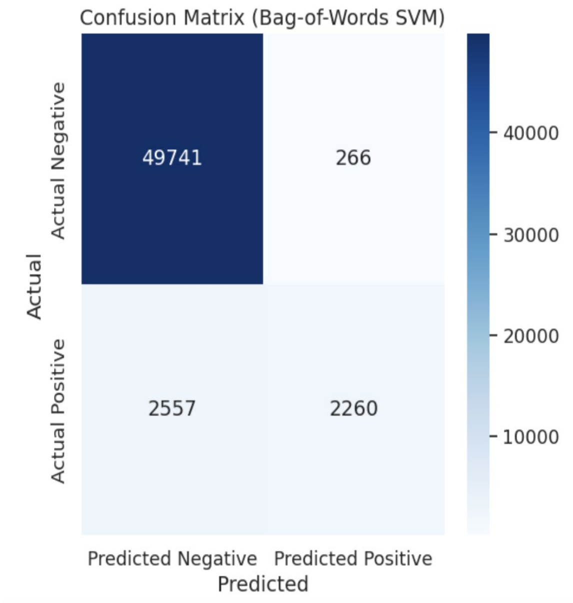

Confusion Matrix (TF-IDF Naive Bayes)
Actual
Actual Negative
Actual Positive
49925
82
4565
252
Predicted Negative
Predicted Positive
Predicted
40000
30000
20000
10000

Figure 4. Bag-of-Words + Naive Bayes Results

Figure 5. TF-IDF + Naive Bayes Results

Support Vector Machine



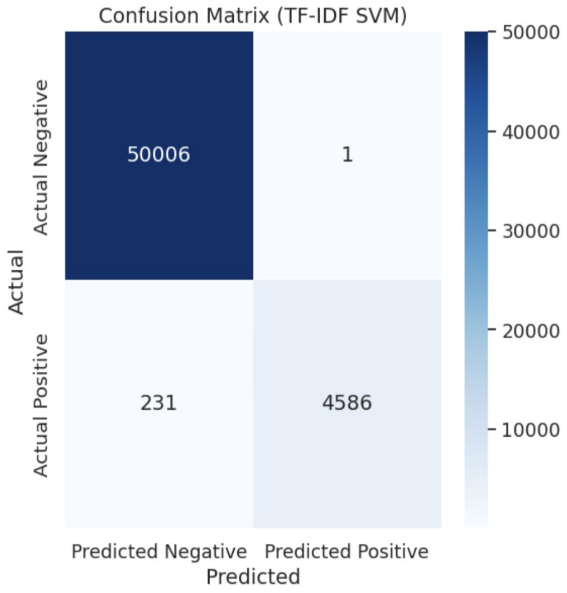
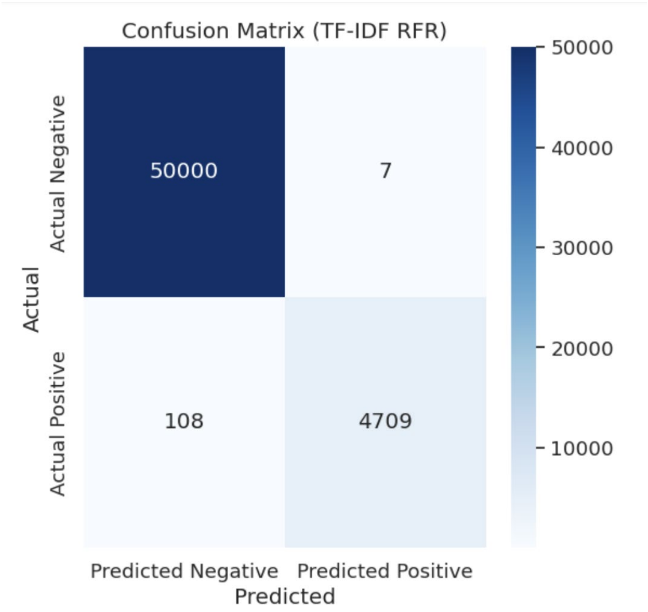


Figure 6. SVM Bag-of-Words Confusion Matrix

Figure 7. SVM TF-IDF Confusion Matrix

Random Forest Regressor



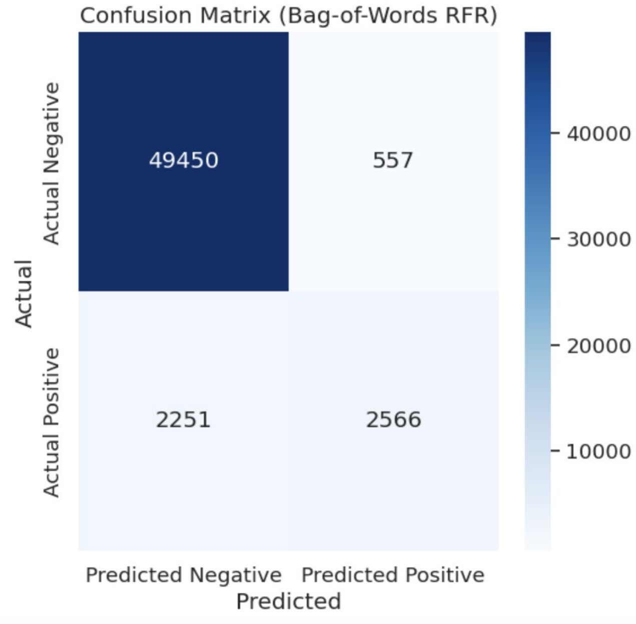


Figure 8. RFR Bag-of-Words Confusion Matrix

Figure 9. RFR TF-IDF Confusion Matrix

Discussion

Evidently, the TF-IDF features with the Random Forest Regressor classifier produced the strongest results, with 100% for samples without criticism and 99% for samples with criticism, and was most likely to pinpoint criticism in the text corpus. The weakest model was the Naive Bayes classifier with TF-IDF features, which had a 96% for samples with criticism and merely 10% for those without. Hyperparameters didn't change the F1 scores at either end of the spectrum, although the learning rate adjusted in the SVM model increased the accuracy. Now, this highly accessible model capable of detecting criticism can serve as a filter for marketing teams around the world, specifically for those using AI in their businesses and especially those using ChatGPT.

While the results are promising, there is room for improvement. The data itself needs to be expanded, covering more counts of criticism. Currently, the classification model is trained on a hard line of phrases/words indicating criticism, but incorporating examples that have underlying sarcasm or criticism will create a model capable of learning more complex patterns for its decision-making, reducing the risk for error on unseen data. The dataset is also specifically trained on tweets about ChatGPT, so combining more datasets on AI, from more platforms as well, to create a more holistic model would help generalize the classification.

The classifier methodology could also improve (at the expense of small-scale accessibility), with access to industrial computing powers or even higher-scale GPUs. Possible alternatives include RNNs, or Recurrent Neural Networks, which are well-suited for sequential data. In its neural network, it has connections that loop back to previous nodes, allowing it to incorporate contextual data in its learning process. Another method is transformers, which are highly complex to allow for a parallel learning process where the model can capture multiple relationships between the text of one sample at the same time. Resource requirements are definitely a bigger factor, especially when you consider the use of stronger transformers like BERT and GPT. Using Aspect Based Sentiment Analysis can also streamline the criticism-flagging by sorting which aspects specifically are being criticized by the consumer.

Conclusion

This study addressed the idea of understanding user sentiment and criticism with an application to the integration of AI in businesses, which provides valuable insights for businesses utilizing AI technologies. Using various classifiers and feature representations showed the strength of TF-IDF features with the random forest regressor in identifying criticism relating to AI. There still remain opportunities for enhancement. Future steps include expanding the dataset to cover a wider range of criticisms and exploring more advanced machine learning techniques. Ultimately, the proposed method equips businesses with an accessible tool to navigate the consumer opinions and criticisms regarding AI, promoting its integration in its services.

Acknowledgments

I would like to thank Dr. Guillermo Goldsztein for guiding me through this process.

References

Bérubé, M., Giannelia, T., & Vial, G. (2021, January 5). Barriers to the implementation of AI in organizations: Findings from a delphi study. ScholarSpace.
https://scholarspace.manoa.hawaii.edu/handle/10125/71425
Criticize synonyms: 84 synonyms & antonyms for criticize. Thesaurus.com. (n.d.).
https://www.thesaurus.com/browse/criticize

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022, January 1). Comparing scientific abstracts generated by CHATGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.12.23.521610v1

Jayaswal, V. (2020a). TFIDF Visualization. Medium. Retrieved from https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d.

Kumar, K. (2021). NLP Bag of Words and TF-IDF Visualization. Medium. Retrieved from https://koushik1102.medium.com/nlp-bag-of-words-and-tf-idf-explained-fd1f49dce7c4.

Kutela, B., Msechu, K., Das, S., Kidando, E. (2023, January). Chatgpt's scientific writings: A case study on traffic safety. https://www.researchgate.net/publication/367335184_ChatGPT's_Scientific_Writings_A_Case_Study_on_Traffic_Safety

Lakhanpal, S., Gupta, A., & Agrawal, R. (2023, August 16). Leveraging explainable AI to analyze researchers' aspect-based sentiment about chatgpt. arXiv.org. https://arxiv.org/abs/2308.11001
Merriam-Webster. (n.d.). 84 synonyms & antonyms of criticize. Merriam-Webster. https://www.merriam-webster.com/thesaurus/criticize

Sobania, D., Briesch, M., Hanna, C., Petke, J. (2023). An Analysis of the Automatic Bug Fixing Performance of ChatGPT. https://www.computer.org/csdl/proceedings-article/apr/2023/021400a023/1P7EqXY3ccw

Soni, N., Sharma, E. K., Singh, N., & Kapoor, A. (2019, May 3). Impact of artificial intelligence on businesses: From Research, Innovation, market deployment to future shifts in business models. arXiv.org. https://arxiv.org/abs/1905.02092

Thota, A. V. (2019, January 6). Applied machine learning: Naive bayes, linear SVM, logistic regression, and Random Forest. LinkedIn. https://www.linkedin.com/pulse/applied-machine-learning-naive-bayes-linear-svm-logistic-thota/

Tobane, W., de Winter, J. (2023, August) Using CHATGPT for Human-Computer Interaction Research: A Primer. https://www.researchgate.net/publication/367284084_Using_ChatGPT_for_Human-Computer_Interaction_Research_A_Primer

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022, February 7). A survey on sentiment analysis methods, applications, and Challenges - Artificial Intelligence Review. SpringerLink. https://link.springer.com/article/10.1007/s10462-022-10144-1

What is sentiment analysis and which businesses need sentiment analysis: ITech. iTech India. (2023, February 23). https://itechindia.co/us/blog/what-is-sentiment-analysis-and-which-businesses-need-sentiment-analysis/#:~:text=The%20latest%20sentiment%20analysis%20data,to%2080%25%20(Bain%26Company).