Utilizing Large Language Models in Ensemble Methods to Boost Sentiment Analysis Explanations

Joshua Gompert jgomper1@jhu.edu
Johns Hopkins University

Abstract

This research explores integrating large language models, notably GPT-3.5-Turbo, into ensemble methods to amplify the efficiency and clarity of sentiment analysis classifiers. The study aims to discern the advantages of embedding GPT-3.5-Turbo within the XGBoost ensemble algorithm, examine the relationship between sentiment classification precision and the quality of explanations, and delve into including an explainability vector in the ensemble. The proposed approach encompasses data acquisition, pre-processing, foundational classifier training, and ensemble formation using XGBoost decision trees. The effectiveness of the ensemble approach with and without the large language model component is assessed, considering the influence of added explainability, and confidence ratings on the model's accuracy and comprehensibility. The results demonstrate the algorithm's increased performance in terms of precision, accuracy, recall, and explainability.

1. Introduction

In the wake of recent advancements in natural language processing (NLP) and machine learning, large language models with an uncanny ability to understand and generate human-esque text have emerged. GPT-3.5 has swiftly become a monumental breakthrough in machine learning and artificial intelligence. These models have demonstrated their prowess across many NLP tasks, including sentiment analysis—a discipline concerned with discerning the sentiment or emotional undertone in textual content (Belal et al., 2023).

Sentiment analysis underpins myriad applications, from monitoring social media pulse and dissecting customer feedback to brand reputation oversight and enhancing information domain cognizance. Historically, sentiment analysis was anchored in various machine learning techniques, including ensemble-based classifiers that amalgamate the strengths of various weaker classifiers for augmented performance (Yue et al., 2019). Nevertheless, natural language's idiosyncrasies and intricate nuances often stymie these traditional methods, culminating in constrained accuracy and explicability.

This research pivots around integrating the formidable GPT-3.5-Turbo into ensemble-based sentiment classifiers, aiming to bolster their efficacy and explicability. We postulate that folding GPT-3.5-Turbo into the ensemble will amplify the algorithm's proficiency in sentiment discernment and offer robust explanations underpinning its predictions.

Our tripartite hypothesis commences with the contention that integrating GPT-3.5-Turbo will foster enhanced transparency and explicability. Subsequently, we propose a correlation between the classifier's precision and the robustness of its explanation, coupled with its confidence level. Our final conjecture asserts that weaving in a classifier's feature vector—spanning sentiment, confidence, and explicability—will elevate the ensemble's accuracy, precision, and recall metrics.

We endeavor to satiate the burgeoning demand for precise and interpretive sentiment classifiers, harnessing the prowess of behemoths like GPT-3.5-Turbo. By scrutinizing the repercussions of GPT-3.5-Turbo's integration, delineating the relationship between sentiment classification veracity and explanatory vigor, and delving into the inclusion of explicability scores in the ensemble, we aspire to galvanize sentiment analysis scholarship, imparting insights to both practitioners and academic cognoscenti.

2. Related Works

Sentiment analysis remains a dynamic research niche within the machine learning sphere. Yue et al. (2019) offered a comprehensive review, highlighting three primary research dimensions: task, granularity, and methodology. Dual class polarity classification, employing naive Bayes and support vector machines, is a dominant strategy, as initially broached by Pang et al. (2002). The bag-of-words technique, capturing n-gram sentiment, was explored by Qu et al. (2010). Granularity-wise, Yue et al. (2019) shed light on diverse modeling scales, ranging from documents to individual words. Methodologically, the spectrum extends from supervised to unsupervised and semi-supervised paradigms.

This study emphasizes harnessing ensemble methodologies to amplify the efficacy of weaker sentiment classifiers. (Hama Aziz & Dimililer, 2021) introduced SentiXGBoost, an innovative XGBoost-based stacked ensemble method. Their approach adeptly amalgamated classifiers like Naive Bayes, K-Nearest Neighbors, Logistic Regression, and more, leveraging the XGBoost meta-classification algorithm. When evaluated across six benchmark datasets, SentiXGBoost showcased enhanced accuracy, recall, and F1-score metrics.

Large Language Models (LLMs) have emerged as top-tier tools in sentiment analysis and sentence embedding domains. The model BERT, an acronym for Bidirectional Encoder Representations from Transformers, developed by (Devlin et al., 2019), is a testament to the versatility of pre-trained language models, amenable to fine-tuning for diverse NLP challenges. A subsequent study by (Liu et al., 2019) unveiled RoBERTa, an optimization of BERT, marking a significant leap in benchmark NLP tests. To tackle the computational heft of RoBERTa, (Reimers & Gurevych, 2019) introduced Sentence-BERT (SBERT), integrating siamese and triplet network structures, dramatically cutting computational durations. In a step forward, (Wang et al., 2022) proposed FEFS3C, an SBERT variant spotlighting sentence semantic content, offering a fresh perspective on sentence similarity evaluation.

(Belal et al., 2023) demonstrated ChatGPT's potential as a universal sentiment analyzer. Their findings underscored ChatGPT's prowess, outclassing lexicon-centric methods with a marked 20% and 25% accuracy enhancement for tweets and Amazon reviews, respectively. As Belal et al. (2023) demonstrated, ChatGPT emerges as a potent data annotation tool in sentiment analysis, sidestepping labeled data predicaments in supervised learning. This research paves the way for streamlined and precise sentiment analysis, especially for digital and social media platforms. Importantly, their work positions ChatGPT as a formidable rival to traditional lexicon-driven sentiment analysis techniques.

3. Explainable Sentiment XGBoost Algorithm (XSXGBoost)

The proposed eXplainable Sentiment XGBoost (*XSXGBoost*) algorithm extends the pioneering work of Hama and Dimililer (2021) , which employed SentiXGBoost as an ensemble technique amalgamating several weak classifiers for sentiment analysis. This innovative algorithm reimagines SentiXGBoost, integrating the GPT-3.5-Turbo Large Language Model (LLM) as a pivotal component, replacing the previously used weak classifiers. This modification infuses eXplainable AI (XAI) capabilities into sentiment analysis and fortifies its performance and generalization, positioning the LLM as a universal sentiment classifier within the ensemble (Belal et al., 2023).

Beyond classification, *XSXGBoost* harnesses GPT-3.5-Turbo to elucidate its rationale for specific text categorizations. These outputs subsequently undergo an explainability analysis within the LLM, serving as a metric to assess the quality of GPT-3.5-Turbo's explanations.

Furthermore, *XSXGBoost* combines this eXplanation Score *(XS)* metric with the GPT-3.5-Turbo classifier's Sentiment Score *(SS)* and Confidence Rating *(CR)* culminating in an eXplainability Vector (XV). This XV then functions as a supplementary feature vector, fed into the XGBoost ensemble's decision trees (DT) to elevate *XSXGBoost*'s efficacy. A comprehensive visualization of the *XSXGBoost* algorithm can be found in Figure 1.

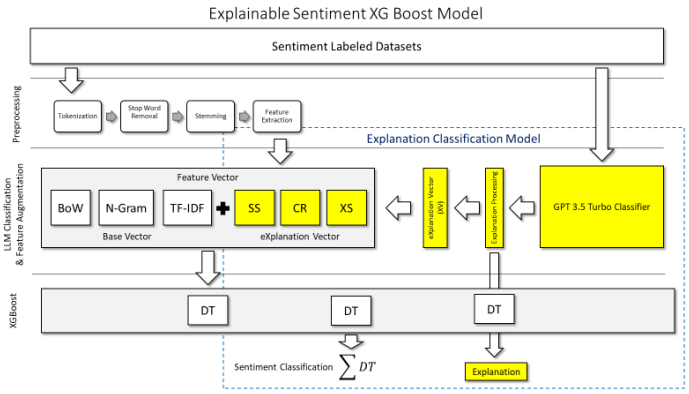

Figure 1: Explainable Sentiment XG Boost Model

3.1 Hypothesis

The core hypotheses underpinning this experiment are:

- **H1** - Incorporating GPT-3.5-Turbo within the classifier ensemble will bolster the transparency and explainability of its output.
- **H2** - GPT-3.5-Turbo's accuracy in sentiment categorization will directly correlate to the robustness of the provided explanations and the resultant confidence scores.
- **H3** - Integrating the eXplainability Vector XV as an auxiliary feature vector will significantly enhance *XSXGBoost*'s overall performance.

4. Experiment Methodology

The experiment methodology unfolds in several stages, closely aligned with the blueprint established by Hama and Dimililer (2021) for the conception of SentiBoost. These stages encompass data collection, pre-processing, feature extraction, foundational classifier training, and ensemble cultivation via XGBoost decision trees. Unique to the *XSXGBoost* model is the integration of an explanatory mechanism, leveraging the prowess of GPT-3.5-Turbo for both classification and extracting elucidations. Scores, synthesized from SS, CR, and XS, are subsequently channeled into the ensemble as supplementary features. This modification diverges from the traditional SentiXGBoost approach, where these scores supplant the role of the weak classifiers.

4.1 Data Collection

In line with Hama and Dimililer's (2021) strategy, this experiment predominantly harnesses the benchmark datasets integral to creating the SentiXGBoost algorithm. Exclusions arose when datasets swelled to unmanageable proportions. For instance, a mere random selection of 1000 entries was extracted from the Movie Review dataset to accommodate memory restrictions.

- Sentiment Labeled Sentences (SLS), Amazon
- Sentiment Labeled Sentences (SLS), IMDB
- Stanford Twitter Sentiment Gold Standard (STS-Gold)
- Yelp Challenge Dataset
- Movie Review (Sentiment Polarity Dataset V2.0)

4.2 Pre-processing

The preliminary data treatment adheres stringently to the methodology embraced by Hama and Dimililer (2021) . To metamorphose raw data into trainable and classifiable vectors, it undergoes a transformation via tokenization, stemming, and stop word elimination—a routine exercise in NLP. Subsequent to tokenization, superfluous "stop words" that scarcely augment the contextual essence for sentiment analysis are excised. Concluding the pre-processing arc, stemming trims words to their root form by discarding affixes. However, an unaltered dataset remains preserved for subsequent classification and analysis via the GPT-3.5-Turbo classifier.

The final preparatory phase before the training onset is feature extraction. Herein, the text is vectored, readying it for classifer training. Drawing inspiration from Hama and Dimilier (2021), this study mimics numerous feature extraction modalities they employed: Bag of Words (BoW), Term Presence and Frequency, and N-gram Features. Due to looming memory constraints and the burgeoning vector dimensionality, methods like Part-of-Speech (PoS) and Opinion Lexicon feature extractions were consciously omitted.

4.3 GPT-3.5-Turbo Classifier

The GPT-3.5-Turbo Classifier uses the OpenAI GPT-3.5-Turbo pre-trained language model, accessed programmatically through the OpenAI API. The system prompt used to shape the response is:

```
Prompt: You are a sentiment analyzer. You will respond to a
user's prompt with a four-part response separated by '|'. The
first part will be a sentiment score between -1 and 1. -1 means
the user's prompt is very negative, while 1 means the user's
prompt is very positive, and a score of 0 is neutral. The second
part is your confidence in the rating, with 0 being no confidence
and 1 being extremely confident. The fourth part will be a one
to two-sentence explanation of your reasoning for rating a text
as positive, negative, or neutral. The third part will be your
grade for the explanation, with 0 being the worst and 1 being the
best. The overall response format is: Sentiment Score [-1 - 1] |
Confidence Rating [0 -1] | Explanation Score [0 - 1] | Explanation
[Free Text]"
```

This process is repeated three times and averaged to control for random variations in the GPT-3.5 algorithm output.

- $SS > 0$ are classified as positive.
- $SS \leq 0$ are classified as negative.

4.4 Explainability Analysis

The algorithm reintroduces the explanations derived from the GPT-3.5-Turbo algorithm's output to the same model. Rather than forecasting sentiment and elucidating it, this iterative step involves grading the quality of previously produced explanations. The algorithm accounts for random variations by averaging outputs from various iterations of the GPT-3.5-Turbo algorithm. The resultant XS score subsequently contributes to the construction of the XV vector embedding, which is presented as:

$$XV = [SS, CR, XS]$$

Aside from the automated grading process facilitated by GPT-3.5-Turbo, the algorithm preserves the explanations for future human interpretation. To furnish clarity, examples spanning true positives, true negatives, false positives, and false negatives accompany the primary results.

4.5 Base Classifier Training

In their work, Hama and Dimilier (2021) utilized base classifiers like Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). In contrast, the *XSXGBoost* model employs the GPT-3.5-Turbo to infuse the feature vector into the language model. The XV feature vector supplants the traditional weak classifier encoding found in the SentiXGBoost model.

4.6 Ensemble Training

The ensemble training mirrors the SentiXGBoost paradigm established by Hama Aziz and Dimililer (2021), with adjustments made to the feature vector as delineated earlier. The output from the GPT-3.5-Turbo, denoted as $XV = [SS, CR, XS]$, enriches the processed sentence vector, which then serves as input for the ensemble's XGBoost meta-classifier. In the ensemble training phase, the XGBoost algorithm progressively refines Decision Tree (DT) classifiers, emphasizing rectifying misclassifications from preceding DT iterations. The algorithm aims to minimize losses by leveraging gradient descent. The optimal values of 100 boost rounds and a decision tree depth of 3 were determined through model tuning. For datasets where the *XSXGBoost* algorithm falls short of expectations, the model evolves by reincorporating SVM, NB, LR, and RF classifier outputs, as proposed by Hama and Dimilier (2021). This augmented model is henceforth designated as *XSXGBoost+*.

4.7 Evaluation

The evaluation seeks to juxtapose the efficacy of the sentiment analysis ensemble method, both with and without incorporating the GPT-3.5-Turbo classifier. The study uses accuracy, precision, and recall metrics to measure performance on all five datasets. Further, the

study explores the influence of LLM's explanations on the system's accuracy and explainability through correlation analysis across the all of the results. The depth and relevance of these explanations were benchmarked against the established ground truth sentiments, thus determining the LLM's capability to offer cogent explanations.

The study also assesses the inclusion of the eXplainability Score Vector, *(XV)*, into the ensemble. This vector, which amalgamated the LLM *SS*, *CR*, and *XS*, actes as an augmented feature vector within the XSXGBoost Algorithm. Key performance indicators like precision, accuracy, and recall measured the advancements in ensemble efficiency upon integrating the *XV*. By examining the amplified efficacy courtesy of the LLM, the lucidity of its generated explanations, and the ramifications of embedding *XV* into the feature space. In two of the five datasets, the results from the *XSXGBoost+* algorithm is compared with the performance of all other classifiers in the study.

5. Results

5.1 Precision, Accuracy, and Recall

The below results show the precision, accuracy, and recall of the XSXGBoost algorithm compared to the results on the same dataset using the following other classifiers: support vector machine (SVM), multi-nominal naive Bayes (NB), logistic regression (LR), random forest (RF), XGBoost without the inclusion of *XV* results, and the binary classification from the GPT-3.5-Turbo classifier. The results for the Amazon, Yelp, and IMDB datasets are presented in Figures 2.1, 2.2, and 2.3.

The Stanford Sentiment Twitter Gold and Movies datasets results have been augmented with an additional classifier that includes the results from the SVM, NB, LR, and RF classifiers as part of the feature space. This classifier is referred to as *XSXGBoost+*. These results are displayed in Figures 2.4 and 2.5 below.

The precision, accuracy, and recall results are displayed with a confidence rating of 0.95 and obtained from performing bagging with replacement.

The *XSXGBoost* algorithm outperformed all other non-LLM-based classifiers on all datasets. The only exception where *XSXGBoost* did not outperform all other algorithms was the Movie review dataset, where the GPT-3.5-Turbo classifier was the best performing overall, and the *XSXGBoost+* algorithm outperformed the *XSXGBoost* on both accuracy and recall.

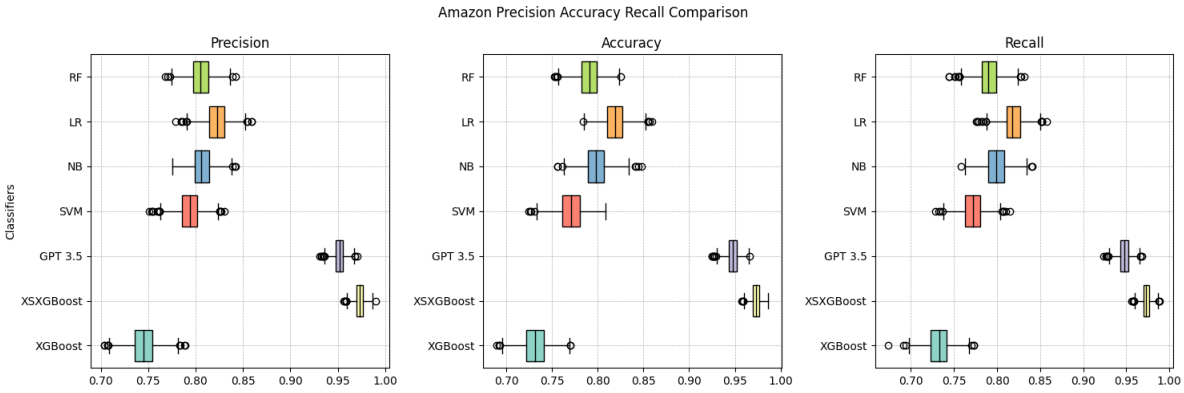

Figure 2.1: Amazon Dataset Results

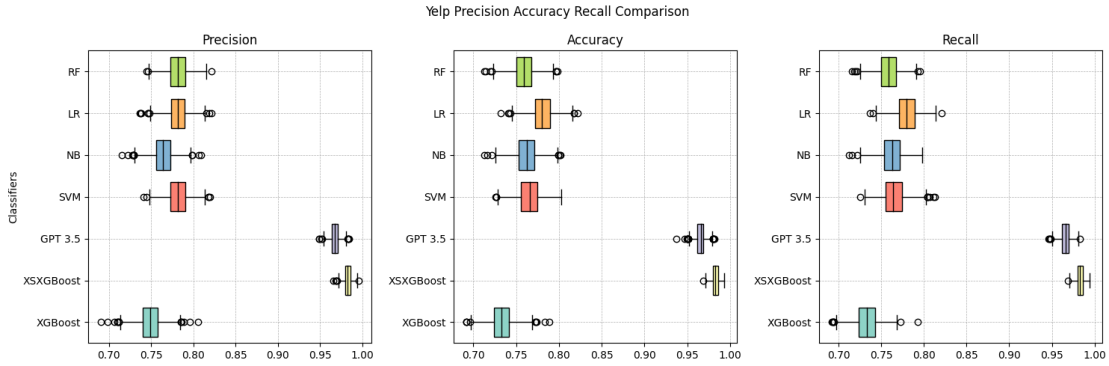

Figure 2.2: Yelp Dataset Results

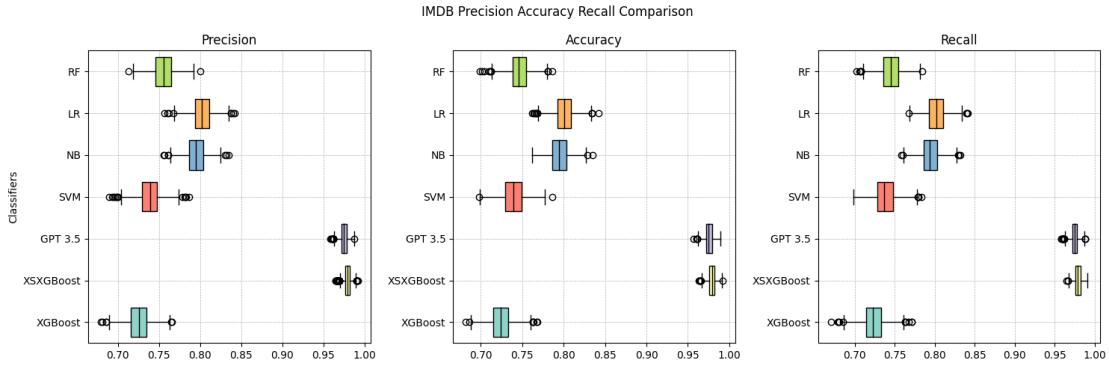

Figure 2.3: IMDB Dataset Results

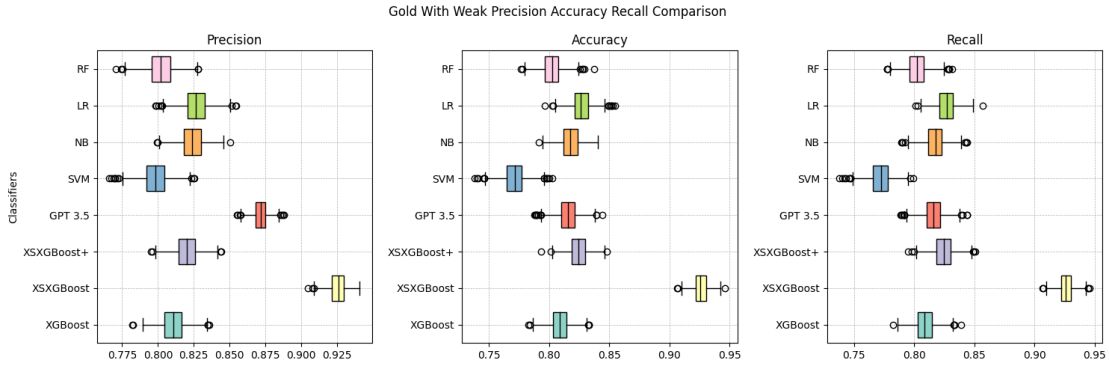

Figure 2.4: Twitter Gold Dataset Results With Weak Classifier Inclusion

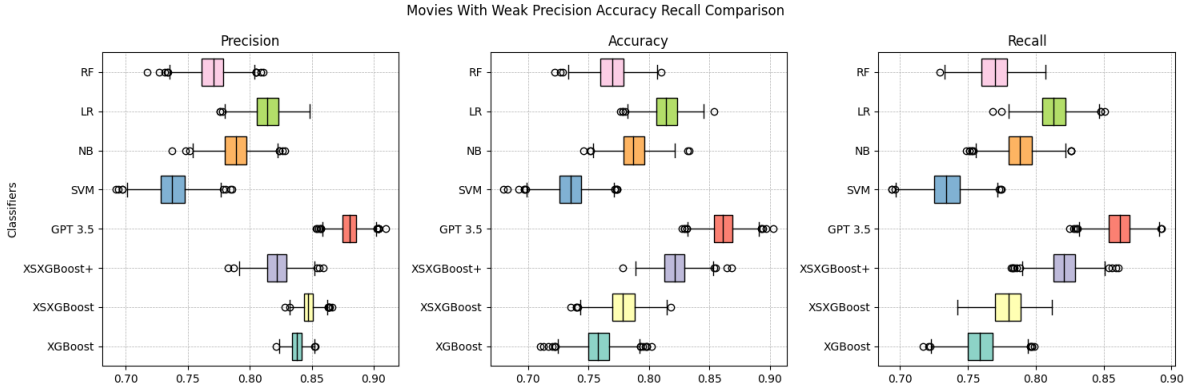

Figure 2.5: Movie Review Dataset Results With Weak Classifier Inclusion

5.2 Feature Correlation

The subsequent results delve into the interplay among the components encapsulated within the SV: SS, CR, and XS. Figure 3.1 depicts the correlation between XS and CR with the model's accuracy. Notably, the results revealed no pronounced correlation among the chosen features. However, the model had a discernible inclination to rate both features as high, irrespective of the actual accuracy.

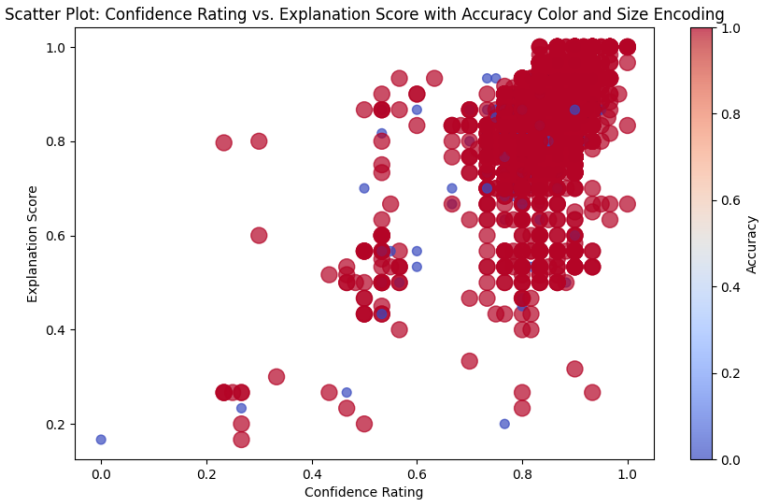

Figure 3.1: Explanation Score / Confidence Rating - Accuracy Correlation Scatter

Figure 3.2 displays the correlation matrix between the true sentiment, *XS*, *CR*, *XS*, and the results. *XS* was positively correlated to both the true sentiment and the results of the *XSXGBoost* algorithm.

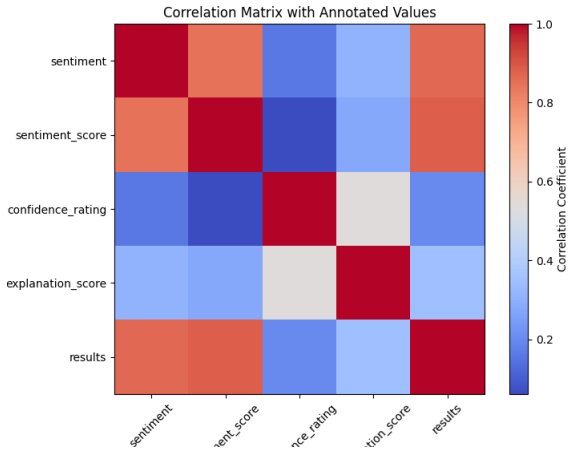

Figure 3.2: Feature Correlation Matrix

5.3 Explanations

Below is a select sample of explanations generated by GPT-3.5-Turbo during sentiment analysis. These samples are drawn from all datasets except for the Movies dataset. The

latter was omitted from this compilation due to the extended length of its input sentences. Although there was not a distinct correlation between *CR*, *XS*, and the algorithm's results, an examination of the sentences and their corresponding explanations provided a logical pathway, illuminating the rationale behind the algorithm's assignment of the respective *SS* for each input.

5.3.1 True Positives:

Sentence: My sister has one also and she loves it.
Explanation: The user's sister loving the product is a clear indication of positive sentiment and enjoyment, resulting in a high sentiment score.
Sentence: Server did a great job handling our large rowdy table.
Explanation: The overall sentiment of the explanations is that the user had a positive experience with the server successfully handling their large rowdy table.
Sentence: If there was ever a movie that needed word-of-mouth to promote, this is it.
Explanation: The overall sentiment of the explanations is positive, as the users believe the movie is good and worth recommending to others.
Sentence: @nicolerichie that was my favorite show/series when i was little!
Explanation: The sentiment score of 0.8 suggests a positive sentiment, as users express fondness and nostalgia for the show, indicating a positive emotional connection.

5.3.2 False Positives

Sentence: Man I so want mcdonalds breakfast..
Explanation: The user's expressions of a strong desire and enthusiasm for McDonald's breakfast indicate a positive sentiment towards it.
Sentence: A must study for anyone interested in the "worst sins" of industrial design.
Explanation: The statement is positive because it highlights the significance of studying the "worst sins" of industrial design, emphasizing the value of learning and improvement in the field.
Sentence: This short film certainly pulls no punches.
Explanation: The explanations suggest a generally positive sentiment towards the short film, as it is described as bold, uncompromising, direct, and honest. However, without further context, the specific sentiment cannot be determined.
Sentence: We've have gotten a much better service from the pizza place next door than the services we received from this restaurant.
Explanation: The overall sentiment is positive, but with some dissatisfaction mentioned. The user prefers the service from the pizza place next door, which implies a positive sentiment, but compares it to the current restaurant, suggesting some level of dissatisfaction.

5.3.3 True Negatives

Sentence: Lost all my Facebook data yesterday Need to start rebuilding today.
Explanation: The sentiment score is slightly negative due to the user's frustration and loss of Facebook data, despite mentioning the need to start rebuilding implying a negative experience.
Sentence: A cheap and cheerless heist movie with poor characterisation, lots of underbite style stoic emoting (think Chow Yun Fat in A Better Tomorrow) and some cheesy clichés thrown into an abandoned factory ready for a few poorly executed flying judo rolls a la John Woo.
Explanation: The explanations indicate a clear negative sentiment towards the movie, highlighting the use of descriptive words such as "cheap," "cheerless," and "poor," as well as criticisms of characterization and poorly executed action sequences. The mention of John Woo and Chow Yun Fat in a Better Tomorrow provides a slight hint of positivity, but overall the sentiment remains negative.
Sentence: As for the "mains," also uninspired.
Explanation: The explanations indicate that the mention of 'also uninspired' expresses a negative sentiment towards the mains, suggesting dissatisfaction, disappointment, and a lack of creativity or excitement in a neutral user prompt.

5.3.4 False Negatives:

Sentence: Much less than the jawbone I was going to replace it with.
Explanation: The majority of the explanations highlight that the sentiment score is negative because the user expresses disappointment in the product by comparing it unfavorably to their desired alternative, which supports the high confidence rating and justification for the sentiment analysis.
Sentence: All rockstars are back home: while 'some of us' freshen up, 'others' watch Magic/Lakers game, then we'll celebrate Sweden Rock in Florida
Explanation: The text overall has a positive sentiment, with mentions of rockstars being back home and celebrating Sweden Rock in Florida creating a joyful atmosphere, although there are some elements of neutrality due to possible divisions or exclusions.
Sentence: PS the only scene in the movie that was cool is when the central character finds her room blown up.
Explanation: The sentiment of the statement is slightly positive with a mention of a "cool" scene, resulting in a moderately clear and confident explanation that highlights the positive aspect of the scene. However, more specific details about why the scene was cool could improve the explanation.
Sentence: Not a weekly haunt, but definitely a place to come back to every once in a while.
Explanation: The overall sentiment of the explanations is slightly positive, indicating

that the user finds the place enjoyable and worth revisiting occasionally.

6. Discussion

6.1 Hypothesis 1 - GPT-3.5-Turbo Increases Transparency and Explanability

The results demonstrate that we can reject the null hypothesis for H1. The inclusion of GPT-3.5-Turbo provides an audit trail for why the GPT-3.5-Turbo classifier produced the results it did when deriving *SS*. As demonstrated n the examples above, even when the classifier produced erroneous results that depart from the true classification, it provided justifications in plain language as to why it did so. These results are easily interpretable by algorithm users, leading to increased transparency and trust in the algorithm as opposed to other non-LLM-based sentiment analysis models. In addition, when examining the logic behind the results, it becomes clear that it would be easy for a human to draw similar conclusions from the text, especially taken out of context, such as when performing sentiment analysis on a dataset.

6.2 Hypothesis 2 - Correlation of Confidence, Explanation, and Accuracy

The results demonstrate that we can not reject the null hypothesis for H2. On the contrary, there is very little correlation between *XS*, *CR*, and the accuracy of the algorithm's predictions. The preliminary conclusion is that regardless of the accuracy of the prediction, the GPT-3.5-Turbo classifier favors producing high results for both *XS* and *CR*. This tendency makes sense, seeing as the model was trained on the internet and had no reason to doubt its ability to accurately classify sentiment or explain why it did so.

6.3 Hypothesis 3 - eXplainability Vector Inclusion Improves XSXGBoost

The results demonstrate that we can reject the null hypothesis for H3. The inclusion of XV as a feature vector significantly improved the performance of the *XSXGBoost* algorithm over XGBoost. The precision, accuracy, and recall were improved in all datasets by adding XV to the feature vector. For three of the datasets, it increased the performance of the ensemble from a worst-performing algorithm to a best-performance one. On the Standford Sentiment Gold dataset, it increased the performance from a middle of a pact algorithm to the best performing. Even on the Movie dataset, its inclusion increased performance, but the results were not statistically significant. The modification of this algorithm to *XSXGBoost+* demonstrated interesting results on this dataset and provided improved recall and accuracy but decreased precision. However, extending the algorithm on the Standford Gold dataset produced negative results for all three metrics. The difference between the two corresponds to the relative performance of the other classifiers on the dataset.

7. Conclusion

The comprehensive evaluation of the sentiment analysis ensemble method incorporated with GPT-3.5-Turbo offers insightful findings on the utility and implications of such integration. The introduction of GPT-3.5-Turbo brings notable transparency and explainability to sentiment analysis, as it not only offers a decipherable audit trail for its sentiment scores but also clarifies erroneous classifications with plain language justifications. This elucidation augments the user's trust and the algorithm's transparency. However, it's pivotal to note that there is not a substantial correlation between the model's *CR*, *XS*, and overall accuracy. Intriguingly, GPT-3.5-Turbo often rates its confidence and explanations highly, regardless of the actual prediction's accuracy. The algorithm's over confidence may stem to its vast training data, which might give it an inherent trust in its capabilities. Furthermore, the introduction of the eXplainability score Vector XV into the ensemble unequivocally enhances the *XSXGBoost* algorithm's performance across multiple datasets. This integration, particularly in conjunction with other classifiers, exhibits potential benefits and pitfalls, underscoring the necessity for careful consideration based on the specific dataset. This study elucidates the tangible advantages of embedding large language models like GPT-3.5-Turbo into sentiment analysis endeavors, emphasizing the importance of explainability in fortifying algorithmic interpretability and trustworthiness.

8. Future Work

The results of this study underscore the importance of systematically comparing the impact of incorporating various large language models into ensemble methods for explainability. There is potential for future research in merging two distinct LLMs into a single ensemble to enhance feature vectors, refine explanations, and improve the grading of confidence and explanation scores. The observed high performance of the model paves the way for its application in semi-supervised learning. Exploring how an LLM fares in labeling web-scraped data, compared to human labeling, could be enlightening. Such an approach might pave the way for creating new datasets for training and testing sentiment analyzers. Moreover, the observed disconnect between confidence, explanation, and accuracy underscores the need for developing more reliable metrics to evaluate the outputs of LLMs, primarily if they cannot be solely relied upon for self-assessment.

References

Belal, M., She, J., & Wong, S. (2023). Leveraging chatgpt as text annotation tool for sentiment analysis. *arXiv*.

Bird, Steven, E. L., & Klien, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Hama Aziz, R. H., & Dimililer, N. (2021). SentiXGBoost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier. *Journal of the Chinese Institute of Engineers*, *44*(6), 562–572.

Kanakaraj, M., & Guddeti, R. M. R. (2015). Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pp. 169–170.

Kora, R., & Mohammed, A. (2023). An enhanced approach for sentiment analysis based on meta-ensemble deep learning. *Social Network Analysis and Mining*, *13*(1), 38.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach..

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Qu, L., Ifrim, G., & Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pp. 913–921.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv*.

Rosenthal, S., Farra, N., & Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518, Vancouver, Canada. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Wang, T., Shi, H., Liu, W., & Yan, X. (2022). A joint framenet and element focusing sentence-bert method of sentence similarity computation. *Expert Systems with Applications*, *200*, 117084.

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*, 617–663.