The efficacy of SBERT models in sentiment analysis of online doctor reviews

Anonymous

1. Introduction

Freely written patient healthcare feedback contains important information which is often absent from structured surveys; there is a growing body of work which uses NLP and ML techniques to extract insights from this data. It has been established that the two key elements of analysis are polarity (sentiment analysis) and thematic categorization of the text (Wallace et al., 2014, p1098). Our hypothesis is that the SBERT text embeddings (Reimers et al., 2019). when applied to sentiment analysis in the health domain, represents a meaningful improvement in performance compared to Naïve Bayes classification. The latter has been previously endorsed as one of the most effective tools for labelling the polarity of medical free-text feedback responses (Greaves et al., 2013).

2. Literature review

The importance of textual data in evaluating quality of care is underscored by Lopez et al., who showed how a rigorous framework for classifying and categorizing online reviews reveals insights which may have been missed on traditional surveys (2012). Wallace et at. identified how sentiment scores can be used in conjunction with topic classification to generate actionable insights from large datasets (2014, p1098). Greaves et al. found that multinomial Naïve Bayes classifiers were comparable in performance relative to other NLP/ML techniques such as decision trees, bagging and vector support machines, and superior in computational efficiency thereby indicating them as a preferred tool for sentiment analysis (2013). Moreover, Georigou et al found that Naïve Bayes outperformed other ML models on healthcare survey data, further solidifying its suitability for the domain (2015.). However, Greaves et al. acknowledged key limitations in Naïve Baye's ability to label texts containing sarcasm, irony, figures of speech, and other heavily context dependent phrases (2013). Text embeddings with the BERT algorithm is a sophisticated deep learning approach to NLP which can be adapted to different tasks and domains through a fine turning process (Devlin et al, 2018.). Through use of a corpus and a neural network, BERT uses the context of surrounding text to vectorize tokens and sentences, thus could more easily see the nuances and caveats in unstructured text which Naïve Bayes is too unsophisticated to capture. Of particular interest is the SBERT algorithm as described by Reimers et al. which is posited to be more computationally efficient and accurate in sentence comparison tasks relative to other BERT based algorithms (2019). In the text classification domain, it has been suggested that SBERT used in conjunction with Random Forest classifier produces optimal performance, based on results in a multiclass classification task (Akber et al., 2023).

3. Method

We will use the dataset constructed by Lopez et al and expanded into a full dataset by Wallace et al to access the models. To ensure a fair test, both models will use the same dataset, although preprocessed in the manner which literature indicates as best practice for each model so we can compare the best-case performance of the models given the data.

3.1. Naïve Bayes model

We constructed the Naïve Bayes model using the same basic pipeline, with modifications, as proposed by Greaves et al as their model performed well on online reviews from the NHS which is of a similar nature to the online reviews in the dataset of interest (2013). We however under sampled the majority class because to improve the class imbalance, the rationale for this deviation was for consistently since the some of the SBERT models were having data balance measures implemented, and Naïve Bayes would be at an experimentally created disadvantage if not given balanced data. Additionally, the no prior polarities were calculated as described by Greaves et al. – manually labelling the sentiment of thousands of n-grams, would not be feasible given the resources available.

3.2. SBERT

The novel nature of SBERT means that there is a lack of precedents and results in the literature to guide best practices. Preprocessing has been used for SBERT to remove stop words, urls, punctuation and other artifacts of language (Akber et al., 2023). However, others have

omitted this step (Reimers et al., 2019). Given SBERT uses the context of surrounding tokens to create its vector representation of a token's semantic meaning, it would be potentially detrimental to remove tangential information in preprocessing. Another major consideration is the model used to classify the SBERT vectors, Akber et al. used a number of ML models to classify the output vectors of which they found a decision tree model called Random Forest to have produce the strongest performance, however there is considerable difference in the domain and task explored by Akber et al. and our domain, namely they were classifying personality in the framework of a psychological model, while in this case we are classifying the sentiment of texts. Furthermore, the overall task was multiclass in nature in Akber et al, while our task is binary. The experiment did not run under the assumption that Random Forest is in general the most suitable algorithm for classifying any SBERT embedded vectors including our task. Therefore. four other candidate algorithms were run: KNN, Logistic Regression, and Multilayer perceptron, they were chosen because of their popularity in previous studies (Akber et al., 2023) and capacity to work with continuous numbers which SBERT vectors are encoded in. A most frequent dummy model was used as the baseline model, it predicts the most commonly observed class in the training data. Hyperparameter selection was carried out where appropriate for each classifier, some significant improvements were seen in tuning the alpha value and hidden layer structure for MLP, the maximum tree depth and minimum impurity decrease for Random Forest and the K value for KNN.

3.3. Evaluation

Accuracy was the natural metric of evaluation since the semantic classifier's utility comes from its ability to correctly predict instances. Due to the imbalance of the dataset, macro F1 was a secondary metric to consider since correct predictions of the majority class can exaggerate the accuracy of a model, since macro F1 gives equal weight to both classes. However, almost all of the models considered had some sort of class imbalance correction applied either in preprocessing or in training therefore accuracy was cautiously used as the main metric. A standard train-test-validation split was used to separate the data, given the large sample size, it was deemed reasonable to use the train-validation split to train then evaluate the models.

4. Results

4.1. SBERT results

	precision	recall	f1-score	support
-1	0.78	0.80	0.79	1414
1	0.93	0.92	0.93	4086
accuracy			0.89	5500
macro avg	0.85	0.86	0.86	5500
weighted avg	0.89	0.89	0.89	5500

Fig.1 Random Forest results

	precision	recall	f1-score	support
-1	0.94	0.77	0.85	1804
1	0.90	0.98	0.93	3696
accuracy			0.91	5500
macro avg	0.92	0.87	0.89	5500
weighted avg	0.91	0.91	0.91	5500

Fig.2 Logistic regression results

	precision	recall	f1-score	support
-1	0.73	0.80	0.76	1346
1	0.93	0.91	0.92	4154
accuracy			0.88	5500
macro avg	0.83	0.85	0.84	5500
weighted avg	0.88	0.88	0.88	5500

Fig.3 KNN results

	precision	recall	f1-score	support
-1	0.86	0.83	0.85	1514
1	0.94	0.95	0.94	3986
accuracy			0.92	5500
macro avg	0.90	0.89	0.90	5500
weighted avg	0.92	0.92	0.92	5500

Fig.4 MLP results

4.2. Naïve Bayes (benchmark) results

	precision	recall	f1-score	support
-1	0.96	0.76	0.85	1835
1	0.89	0.98	0.94	3665
accuracy			0.91	5500
macro avg	0.92	0.87	0.89	5500
weighted avg	0.91	0.91	0.91	5500

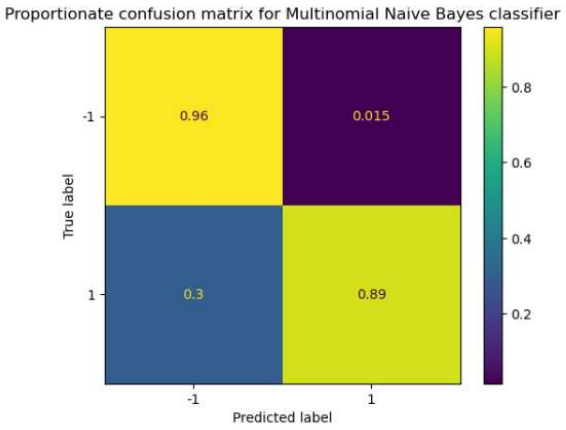
Fig.5 Multinomial Naïve Bayes results

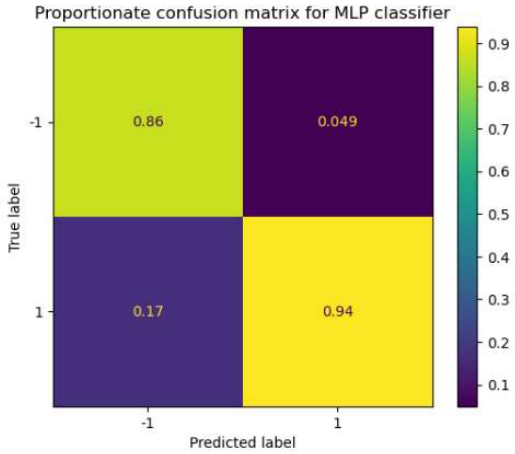
4.3. Dummy model results

	precision	recall	f1-score	support
-1	0.00	0.00	0.00	0
1	1.00	0.73	0.85	5500
accuracy			0.73	5500
macro avg	0.50	0.37	0.42	5500
weighted avg	1.00	0.73	0.85	5500

Fig.5 Most Frequent dummy model

4.4. Confusion matrices





4.5. Results description

For the SBERT data, the multi-layer perception classifier had the best performance with an accuracy and macro f1 of 92%, while KNN had the worst performance with a accuracy of 84% and an F1 of 88%. Contrary to previous results the Random Forest classifier was not the best model for classifying SBERT text embeddings with only 89% which indicates that both techniques have useful predictive capabilities in this dataset.

5. Discussion

5.1. Performance analysis

It seems that SBERT in conjunction with a multilayer perception slightly outperforms Naïve Bayes which has previously been described as one of the most suitable models for sentiment analysis of online health care reviews (Greaves et al., 2013.). This is apparent through it's higher Accuracy, Marco F1, Macro Precision and Macro recall scores compared to Naïve Bayes. On the other hand, the robustness of Naïve Bayes is underlined in how none of the SBERT classifiers other than MLP managed to outperform it in terms of accuracy or macro F1, despite using bag of words which is considered a less sophisticated text vectorization approach than BERT. One possible explanation for this surprising result is that vector length for SBERT is only 384 compared to the 768 length of the original BERT model (Delvin et al., 2018), each value of the vector represents some property of a word or sentence, it follows that a longer vector could encode semantics at a higher resolution, and that this might be necessary in this case for exceptional performance. It is apparent that both models overpredicted positive sentiments as both models had considerably higher recalls for the positive class. This was expected as the under sampling reduces the class imbalance rather than eliminating it. It seems that the slight relative advantage of SBERT-MLP can be attributed to lower susceptibility to misclassifying positive reviews as negative reviews as 30% of BOW-MNB negative predictions were incorrect compared to only 17% for SBERT-MLP as seen in the confusion matrices.

5.2. Limitations

One major confounding factor was the omission of the "prior polarity" preprocessing step for Naïve Bayes which involved the manual classification of the 1000 most common words in the data set and the 1000 most common 2-grams (Graves et al., 2013). If this process had been implemented, Naïve Bayes may have had a better performance. This is particularly of concern given the narrow margin between the performance of the novel SBERT model and Naïve Bayes, which should considerably detract from any strong conclusion drawn in favor of SBERT. Conversely the Multi-Layer-Perceptron may not have performed as well as it could have due to resource constraints preventing comprehensive parameter tuning, only a few combinations of hidden layer structures, alpha values and iteration maximums were searched over, there are also a number of other parameters which could have been tuned such as step size in gradient descent. Another key issue was how the Greaves approach and our approach did not share any classification models in common, this makes it harder to

attribute increases in performance to the effectiveness of the SBERT text encodings or to the properties of the classification models used. Lastly the selection space for SBERT classifiers was only a small sample of the available models, for instance a future avenue of research could be the use of Convolution-Neural-Networks as a classifier for SBERT vectors, as this hybrid architecture as already been proven as particularly effective in sentiment analysis, albeit for BERT rather than SBERT though it is expected they would have relatively similar behaviors given their close relation (Huang et al., 2021.). Due to the various limitations having opposing implications for the comparison of SBERT and Naïve Bayes in this case, we will not draw any conclusion based on the limitations, but in light of these complications, the final conclusion should be reasonably conservative.

6. Conclusion

While the data suggests that SBERT when combined with MLP performs better than Naïve Bayes which has been posited as one of the best sentiment classifiers in the online health care review domain. However, the margin between the key performance metrics of the Naïve Bayes model and the best SBERT model was close despite the purported sophistication of SBERT. Therefore, there is not a definitive conclusion for our hypothesis that SBERT embeddings are a meaningful improvement over Naïve Bayes in online health review sentiment analysis. However, it is reasonable to say that SBERT is likely at least approximately similar in performance to Naïve Bayes based on the results found, since it's performance in the SBERT-MLP model was slightly better in evaluation.

7. Ethics statement

The dataset provided does raise privacy concerns since there are real names in the reviews associated with sensitive information including medical diagnoses and critical comments. The intent is to use this data to derive general scientific insights of benefit to the community without examining people's personal lives or looking to sell this data to third parties. The investigators are morally obligated to treat people's personal information with dignity and will not unnecessarily disclose the data provided.

Bibliography

[1] Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. (2014). *A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews.* *J Am Med Inform Assoc*, 21(6):1098{103.

[2] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERTNetworks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro-cessing. Association for ComputationalLinguistics.Assoc*,21(6):1098{103. Accessed: July, 2022

[3] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., and Donaldson, L. (2013). *Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet* Res, 15(11):e239.

[4] Lopez, A., Detz, A., Ratanawongsa, N., and Sarkar, U. (2012). *What patients say about their doctors online: a qualitative content analysis. J Gen Intern Med*, 27(6):685{92.

[5] Georgiou D, MacFarlane A, Russell-Rose T. *Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. Science and Information Conference*. 2015:352–61.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

[7] Akber, M. A., Ferdousi, T., Ahmed, R., Asfara, R., & Rab, R. (2023). *Personality Prediction Based on Contextual Feature Embedding SBERT. 2023 IEEE Region 10 Symposium (TENSYMP), Region 10 Symposium (TENSYMP), 2023 IEEE*, 1–5. https://doi.org/10.1109/TENSYMP55890.2023.10223609

[8] Computer, P. H. D. of, Huang, P., Computer, D. of, Computer, H. Z. D. of, Zhu, H., Computer, L. Z. D. of, Zheng, L., Computer, Y. W. D. of, Wang, Y., & Metrics, O. M. A. (2021, December 1). *Text sentiment analysis based on Bert and Convolutional Neural Networks: Proceedings of the 2021 5th International Conference on Natural Language Processing and information retrieval.* ACM Other conferences. https://dl.acm.org/doi/abs/10.1145/3508230.3508231