# Research Report

## 1. Introduction

Patient experiences and satisfaction serve as integral components in the healthcare delivery system. This makes understanding patients' sentiment pivotal for enhancing healthcare quality. In the modern digital age, online reviews on healthcare platforms have emerged as important sources of such sentiments.

This research aims to classify the sentiments of online physician reviews using machine learning models, specifically Naive Bayes (NB) [4][8], K-Nearest Neighbors (KNN) [8], and Multilayer Perceptron (MLP) [7][8]. The dataset employed in this research is sourced from ratemds.com and further transformed into a computational-friendly format [3][5]. The dataset comprises a total of 54,107 reviews, with 43,003 designated for training purposes. In addition to the raw review texts, two feature set files, generated from this dataset, are made available for our analysis: Term Frequency - Inverse Document Frequency (TFIDF), and Word-Embeddings [9]. Additionally, this research also attempts to explore whether simpler models such as NB and KNN, enhanced through feature selection and hyperparameter tuning, can outperform a more complex model such as MLP without such optimisations.

## 2. Literature Review

The existing literature on online reviews sentiment analysis in healthcare displays a diverse range of approaches to assessing patient sentiments. Each study brings forward unique methodologies and objectives within this domain.

Lopez et al. (2012) [3] conducted an explorative study to understand what patients articulate about their doctors online, focusing on internists and family medicine physicians. They employed a qualitative approach to derive context from reviews, sourced by mimicking user searches in an urban locale.

In a different approach, a probabilistic generative model is developed by Wallace et al. (2014) [5] to explore latent factors and sentiment in online doctor reviews, leveraging a small amount of expert-annotated dataset. The deployment of mixed qualitative and quantitative methods, alongside factorial LDA, yielded superior results over traditional bag-of-words or LDA models.

In more quantitative work, Greaves et al. (2013) [6] used numeric ratings on a Likert scale to label extracted comments. This method, compared to manual annotation, presented a more economical alternative for assessing patient satisfaction. The study conducted a comparative analysis on four machine learning models: Multinomial NB (MNB), Decision Tree (DT), Bagging and Support Vector Machines (SVM), with MNB outperforms others in accuracy and efficiency.

Recent work by Doing-Harris et al. (2016) [2], demonstrated an automated solution for annotating comments with common topics of patient satisfaction and dissatisfaction. They utilise a vocabulary-based classifier and a NB classifier for tagging topics, and a Natural Language Tool Kit (NLTK) NB classifier for sentiment prediction. The iterative approach of feature refinement enhanced overall model performance.

## 3. Method

### 3.1. Exploratory Data Analysis

Prior to implementing machine learning models, an explorative data analysis (EDA) was conducted to understand the dataset. This preliminary investigation aimed to explore various attributes such as the distribution of positive and negative reviews, feature correlations, and the potential decision boundaries. These insights guide the subsequent model and feature selections.

### 3.2. Comparative Analysis of Models

NB, KNN, and MLP were examined for their efficacy in sentiment classification. They were implemented by the Scikit-learn Python library [15].

NB was selected for its proven efficiency and effectiveness in text classification tasks [1][2][6][7][10][11], despite its assumptions of conditional and positional independence of features, which may not be well-suited for sentiment analysis tasks [4][14]. In this study, MNB is chosen as the value distribution of the combined feature set aligns more closely with a multinomial distribution, as evidenced in Figure 1.

KNN was included for its distance-based classification capabilities, making it well-suited for capturing semantic similarities in the Word-Embeddings [8]. Additionally, KNN's non-linear decision boundary aligns well with the data's complexities.

MLP, a feedforward neural network [8], is considered the most complex out of the three models. Considering the nature of the text dataset, it was chosen for its ability to better capture feature interactions and its robustness in forming non-linear decision boundaries.

A Zero-Rule model utilising a majority vote strategy was employed to serve as the baseline due to the dataset's skew towards positive ratings, which exceed 70% in both training and validation sets. (Figure 2).

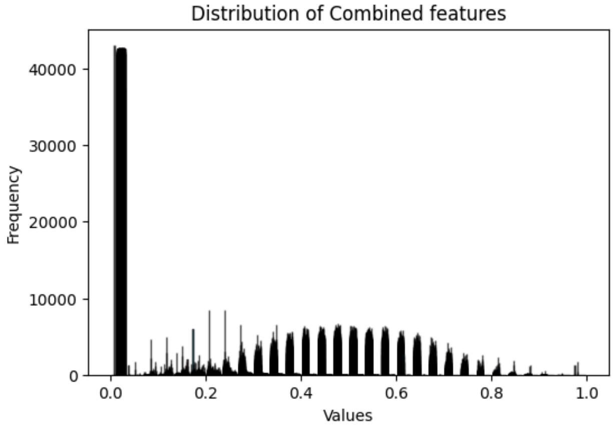

Figure 1: Value distributions for combined features

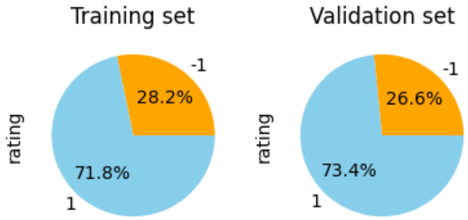

Figure 2: Rating distribution in dataset

### 3.3. Feature Selection and Hyperparameter Tuning

Word-Embeddings, TFIDF and the combination of both were examined. Each model was trained across these feature sets, and the performance was compared to choose the most effective feature set for subsequent phases.

To optimise MNB and KNN, we used Random Forest with entropy-based feature selection. This ensemble method, built upon decision trees, is capable of capturing non-linear feature interactions [8]. We aim to select the most informative features, and improve the decision boundaries for linear models like MNB. Coupled with an exhaustive grid search, we aim to find the optimal hyperparameters for the Random Forest, and the optimal feature subsets.

Hyperparameter tuning was specifically focused on the KNN model, as MNB offered limited room for hyperparameters adjustments. In this phase, we employed the inverse of the distance as a weighting factor to diminish the effects of random noise and outliers. Furthermore, we ensured an odd value for k to eliminate the possibility of classification ties, thereby ensuring decisive classification.

MLP, on the other hand, operated on default parameters as defined by Scikit-learn [15].

### 3.4. Evaluation

Given the skewed distribution of the dataset and the equally critical role of both positive and negative sentiments in healthcare reviews, the macro average F1 score [8] was chosen as the primary performance metric. This metric offers a balanced view of precision and recall across both classes, thereby providing a comprehensive measure of model effectiveness. Additionally, stratified cross-validation [8], a technique renowned for maintaining class distribution across folds—a critical factor given the dataset's skewness, was carried out to evaluate the model bias and variance using training dataset. Specifically, model bias was quantified by the average training score, and variance was determined by the gap between the training and validation scores. Confusion matrix was employed to pinpoint the sources of classification errors [8].

# 4. Results

## 4.1. Exploratory Data Analysis

The EDA unveiled key insights. A focused examination of the TFIDF feature set revealed strong correlations between specific word choices and sentiment ratings as illustrated in Figure 3.

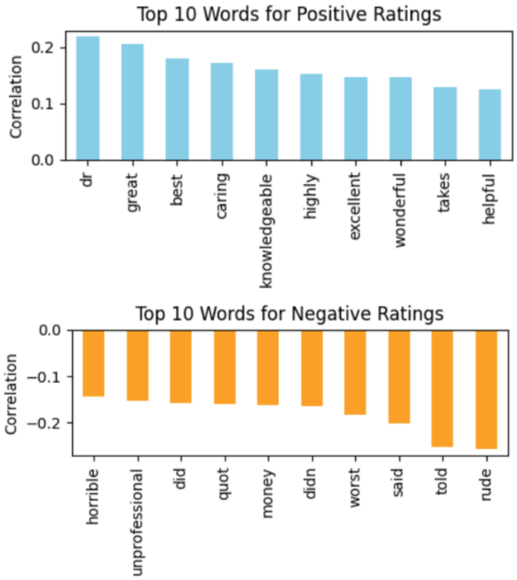

Figure 3: Words correlated to positive and negative ratings

Moreover, Principal Component Analysis (PCA) [8] applied to the combined feature set failed to reveal clear decision boundaries (Figure 4). This absence suggests that the underlying structures separating rating sentiments are likely inherently non-linear.

## 4.2. Model Performance

MNB preferred TFIDF, while KNN and MLP preferred Word-Embeddings. Nonetheless, combining both features led to enhanced performance.

Figure 5 provides a comprehensive view of these models' performance metrics, benchmarked against the Zero-Rule baseline. Evidently, each model surpassed the baseline in all metrics. Despite the effort, MLP outperformed both MNB and KNN. Optimisation had a marginal impact on MNB's performance but notably enhanced KNN's.

## 4.3. Model Bias and Variance

Predictably, the Zero-Rule model exhibited the highest bias but a low variance, attributed to the use of stratified cross-validation.

As captured in Figure 6, MNB demonstrated the highest bias yet the lowest variance. Conversely, KNN displayed the highest variance, flagging the potential overfitting. MLP, on the other hand, stood out for its lowest bias and moderate variance.

## 4.4. Confusion Matrix

Figure 7 reveals model performances in false positives and negatives. Zero-Rule defaulted to the majority label—positive—generated 1,462 false positives. MLP demonstrated the least susceptibility to false positives, with only 114 cases, but highest false negatives. In contrast, KNN and MLP showed similar rates of both false positives and negatives, with KNN slightly outperforming MLP in reducing false positives by 22 cases but lagging in minimising false negatives by 21 cases.

In a complementary vein, the examination of classification errors, as depicted in Table 8, suggests the presence of mislabeled reviews in the dataset, thereby hinting at the presence of random noise in the dataset.

| Review | Label |
|---|---|
| *I have been seeing Dr. Fahrendorf for years. He always seems to help me with my back problems. He really is a great chiropractor* | -1 |
| *Doesn't have an on call Dr. number. I was in pain and just wanted to speak to a dentist that knew me, and let them know what was going on. I had to go to an urgent care to get antibiotics sbd pain meds that didn't work.* | 1 |

Tabel 8: Examples of misclassification

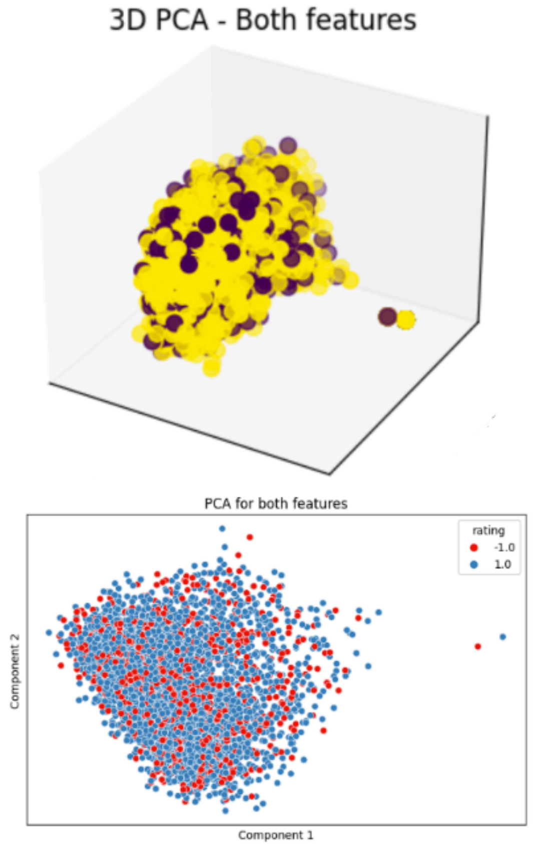

Figure 4: PCA projections for the combined feature set

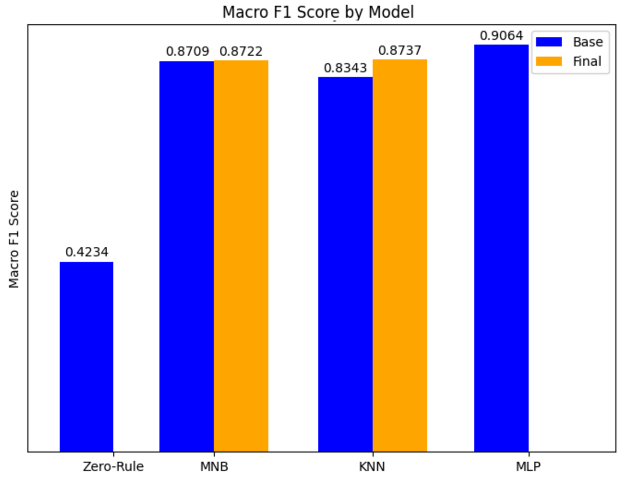

Figure 5: Model performance

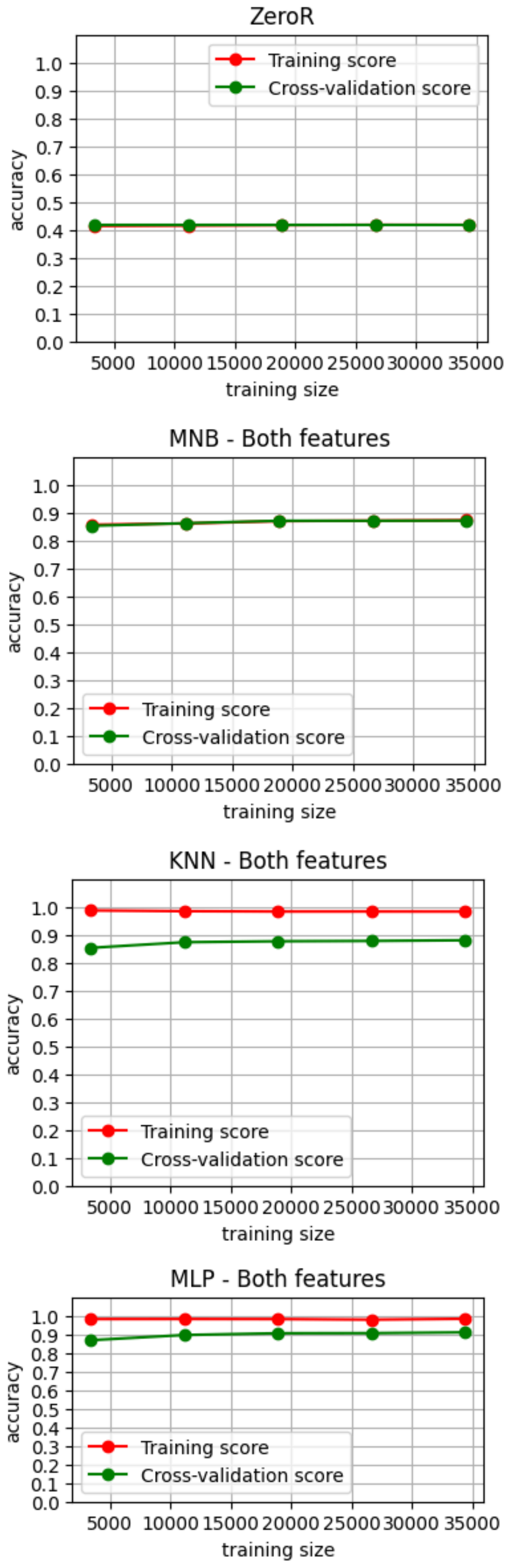

Figure 6: Model bias and variance

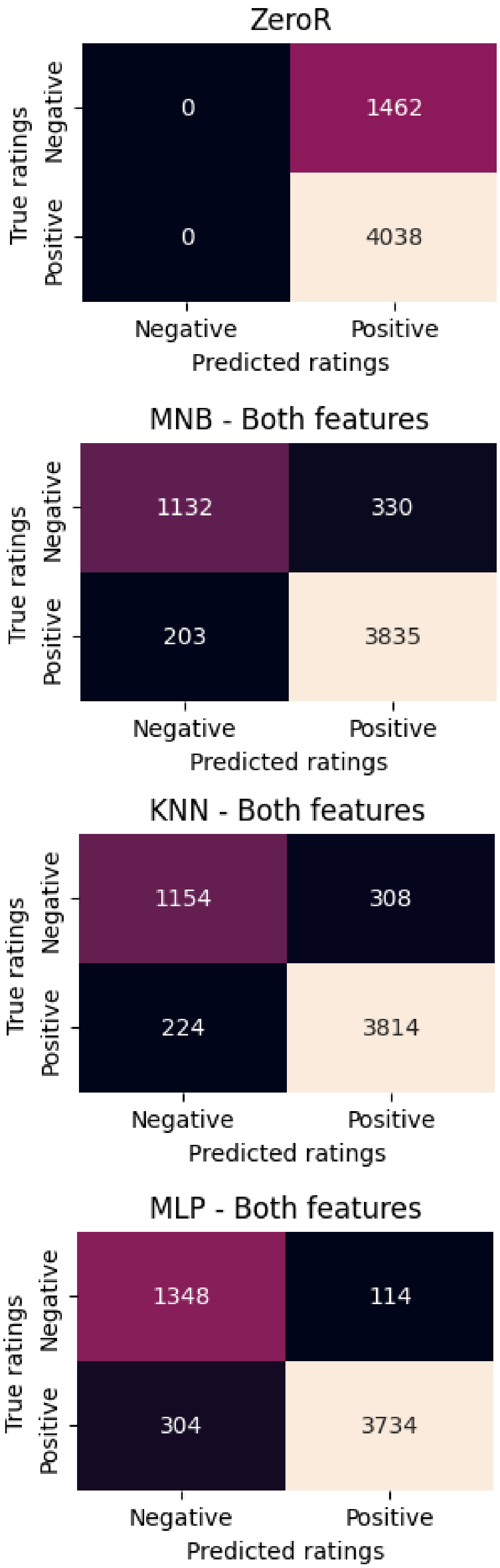

Figure 7: Confusion matrix

## 5. Discussion

Our research explores whether simpler models like MNB and KNN, when optimised via feature selection and model parameter adjustments, can outperform an unoptimised MLP. The performance metrics unequivocally favour MLP, even when it operates with default parameters that failed to converge. Additionally, the outperformance of both KNN and MLP over MNB aligns with the likely non-linear decision boundaries evident in Figure 4. This discovery underscores the importance of matching model assumptions with dataset characteristics.

Another layer of complexity is introduced by the presence of seemingly mislabeled reviews, which challenge the integrity of the "ground truth" used for training and evaluation. This random noise might partially explain why a robust model like MLP didn't converge and experienced elevated false negatives.

The consistent preference for the combined feature set across models suggests that each feature set captures complementary and unique aspects of the data. This calls for a deeper exploration of feature importance and signals the potential benefits of additional feature engineering

The MNB model demonstrated commendable generalisation, as evidenced by its low variance. However, its inherently linear nature limits its capacity to grasp non-linear complexities [14]. This limitation, compounded by its inability to select proper weights for imbalance dataset [14], resulted in the lowest F1 score and the highest false positives.

KNN presented a unique set of strengths and weaknesses. Its high variance indicates susceptibility to noise and a risk of overfitting. The model's local learning and majority vote strategy [13] act as a double-edged sword in imbalance datasets: while effective in reducing false negatives, they inflate false positives simultaneously. Notably, Euclidean distance surprisingly outperformed cosine similarity, a metric often preferred for document similarity [7][8]. This can be attributed to feature normalisation, which enables Euclidean distance to capture both angular differences and point closeness. However, the computational cost of KNN [4][8], particularly when dealing with large datasets, presents a trade-off that may not be justified given the marginal improvements over MNB.

MLP is less sensitive to noise than KNN because it evaluates all instances rather than just those in close proximity. While it minimised false positives, it did so at the expense of increased false negatives. This is likely due to its capability to capture non-linear relationships, thereby increasing the risk of overfitting to the minority class, especially when the loss function is not tailored for class imbalance [12]. In the context of healthcare reviews, such trade-offs could carry ethical implications as it overlooks critical positive feedback on healthcare providers.

## 6. Limitations

This research adopted only Random Forest for feature selection, and did not exhaustively test all parameters. Future studies might consider alternative feature selection techniques, such as Pointwise Mutual Information (PMI) [8], given that the choice of algorithm can significantly influence model performance. Additionally, the research was limited to using only Word-Embeddings and TFIDF, leaving room for further feature engineering to potentially enhance model efficacy.

## 7. Conclusion

While simpler models like MNB and KNN show promise after optimisation, they failed to outperform an untuned MLP model in handling non-linearities and imbalanced classes. Each model exhibits unique merits and demerits, and they offer valuable insights for sentiment analysis in healthcare reviews.

## 8. Ethics Statement

The dataset used in this research originates from qualitative work by López et al. (2012) [3], with feature sets sourced from Wallace et al. (2014) [5]. Both datasets come without explicit terms of use and are utilised as provided. Importantly, the researcher has not independently validated their accuracy. The existence of seemingly mislabeled reviews in the dataset raises ethical concerns, warranting caution in using the prediction results for decision-making. Furthermore, the researcher cannot guarantee the absence of confidential or sensitive information in the dataset, although no such information has been disclosed in this study.

## Bibliography

[1] Didi, Y., Walha, A., & Wali, A. (2022). COVID-19 tweets classification based on a hybrid word embedding method. Big Data and Cognitive Computing, 6(2), 58.

[2] Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., & Conway, M. (2016). Understanding patient satisfaction with received healthcare services: A natural language processing approach. PubMed, 2016, 524–533.

[3] Lopez, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: A Qualitative content analysis. Journal of General Internal Medicine, 27(6), 685–692.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008b). Introduction to information retrieval. Cambridge University Press, Cambridge, UK.

[5] Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. Journal of the American Medical Informatics Association, 21(6), 1098–1103.

[6] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from Free-Text comments posted online. Journal of Medical Internet Research, 15(11), e239.

[7] Eisenstein, J. (2019). Introduction to natural language processing. MIT Press.

[8] Han, J., Pei, J., & Tong, H. (2022b). Data mining: Concepts and Techniques. Morgan Kaufmann.

[9] 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 Processing. Association for Computational Linguistics.

[10] Samuel, J.; Ali, G.; Rahman, M.; Esawi, E.; Samuel, Y. COVID-19 public sentiment insights and machine learning for tweets classification. Information 2020, 11, 314.

[11]Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. Information 2019, 10, 150.

[12] Aurelio, Y. S., De Almeida, G. M., De Castro, C. L., & De Pádua Braga, A. (2019). Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. Neural Processing Letters, 50(2), 1937–1949.

[13] Hastie, T., Tibshirani, R., & Friedman, J. H. (2013). The elements of statistical learning: data mining, inference, and prediction.

[14] Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. International Conference on Machine Learning, 616–623.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.