Paper Template for COMP90049 Report

Anonymous

1. Introduction

In recent years, individual looking to seek information about healthcare practitioners has utilized online review websites such as ratemds.com. The amount of user feedback and comments on these websites allowed them to be a useful source of data to identify and study emerging problems in healthcare, as well as assess quality of care and patient safety and satisfaction. Due to the large amount of user-generated comments, data from websites such as ratemds.com have been proven to be useful in evaluating the performance and appropriateness of machine learning and natural language processing models in classifying trends and patterns in healthcare services provided by clinicians.

Data used in this report are derived from Wallace et al. (2014) and López et al. (2012). Reviews are obtained from ratemds.com. The dataset provides information about clinician ID for each distinct clinician, their gender, patient's comments and rating.

In this report, various feature selection methods are used to select input features for KNN, Gaussian Naïve Bayes and Logistic Regression classifiers to investigate whether the effect of different features set on the performance of classifiers is observable. Classifiers with embedded feature selection mechanisms like Decision Tree will also be used to discuss the impact of feature engineering. Feature selection methods and word representations such as TFIDF, word embeddings using SBERT; and filtering methods such as Chi-Square and Mutual Information are explored in this report.

2. Literature review

2.1. Dataset

The patient review dataset was derived from Wallace et al. (2014) and López et al. (2012). It aimed to replicate the process of looking up healthcare information online by people who are seeking care. The website used for data retrieval is ratemds.com, "RateMDs is a platform for patients to review doctors across four dimensions of care: helpful, knowledge, staff, and punctual." (Wallace et al., 2014, p.1098). The dataset is highly dimensional, with its features being words from the patient's comment regarding care received. The original dataset included doctor ID, gender, comment field, and rating. Patient's comments are presented as raw text, with whitespace, punctuation and misspellings not removed. Rating is the class label, and all comments can be classified as either -1 for negative sentiment or 1 for positive sentiment.

The comments are pre-processed by vectorizing using count vectorizer to obtain a dictionary of words (with uninformative words such as articles and pronouns being removed) in the comment and their frequency.

2.2 Related work

In Pintas et al. (2021), common methods of feature selection used for text classification are discussed. Feature selection methods are usually categorized by strategy (Pintas et al., 2021), which are filter, wrapper, embedded, hybrid and ensemble methods (Pudjihartono et al., 2022; Pintas et al., 2021). Filter methods are independent of classifier, whereas wrapper and embedded methods are model-dependent. Hybrid methods attempt to apply more than one feature selection method on the dataset; most commonly filter (or univariate filtering) followed by embedded or wrapper methods (Pudjihartono et al., 2022; Pintas et al., 2021). Ensemble methods work by aggregating outputs of different feature selection methods to derive selected features (Pudjihartono et al., 2022).

Pintas et al. (2021) found that filter is the most common method for feature selection due to its model-independent characteristic and to reduce data

sparsity in text classification. Additionally, Pintas et al. (2021) presented that Naïve Bayes, K-Nearest Neighbours and Decision Tree were among the most used classifiers to investigate the effect of feature engineer on text classification.

3. Methods

3.1. Feature selection methods

Feature selection is crucial in machine learning to select relevant and highly correlated features when using high-dimensional datasets (Bommert et al., 2020). Selecting appropriate features helps to avoid overfitting and allows the model to have better generalizability. There are various approaches to feature selection, namely filter method, wrapper method and embedded method. In natural language processing, methods such as word representations using TFIDF and word embeddings using SBERT are also applied in an attempt to derive highly correlated words in document. This report will discuss filter methods (Chi-Square, Mutual Information), TFIDF and word embeddings using SBERT, embedded method via Decision Tree; and the performance of three machine learning models KNN, Gaussian Naïve Bayes and Logistic Regression. Wrapper method is not explored in this report due to the dataset being large and highly dimensional; therefore, leads to the method being computationally heavy and not suitable.

3.1.1 Filter method (Univariate filtering)

Filter methods work by ranking features based on calculated scores independent of model; and selecting features with the highest score or based on threshold (Bommert et al., 2020). Filter methods are model-agnostic and can help reduce run time of machine learning algorithms (Bommert et al., 2020). Two filter methods that are explored in this report are Mutual Information and the Chi-Square test.

Chi-Square test is a statistical technique to determine relationships between different features and the label. Mutual Information is a method of selecting features based on the reduction of entropy, i.e. random information or uncertainty in the class variable (Zhou, H., Wang, X. & Zhu, R, 2022). However, Mutual Information has a bias toward rare and uninformative features that occurred highly with a class as it is calculated based on probabilities and not the frequency of such events occurring.

In Bommert et al., (2020), it is shown that there is no clear filter method that outperforms across datasets as the best method depends on the dataset. It also reaffirmed the belief that filtering features is better than no filtering at all.

Both Mutual Information and Chi-Square tests are used on the vectorized raw text comment to select highly relevant features as input for the classifiers based on the highest 50 scores.

The list of 50 features selected using Mutual Information is as follows:

63, amazing, arrogant, asked, bad, best, called, caring, comfortable, compassionate, did, didn, dr, excellent, friendly, got, great, helpful, highly, horrible, kind, knowledgable, knowledgeable, later, left, listens, love, money, pain, poor, professional, quot, refused, room, rude, said, takes, terrible, thorough, told, tried, uncaring, unprofessional, wanted, waste, went, wonderful, worse, worst, wrong.

The list of 50 features selected using Chi-Square is as follows:

63, appointment, arrogant, asked, bad, best, blood, called, caring, did, didn, dr, excellent, friendly, got, great, helpful, highly, horrible, insurance, knowledgeable, later, left, listens, love, medication, money, pain, poor, quot, records, refused, room, rude, said, takes, terrible, test, thorough, told, tried, uncaring, unprofessional, wanted, waste, went, wonderful, worse, worst, wrong

The features selected using Mutual Information and Chi-Square tend to relate to physicians attitudes and facilities.

3.1.2 TFIDF

Term frequency-inverse document frequency is a method of representing a term with regard to the number of times it appears in a document and how many documents it appears in. It is an attempt to measure the relevancy of words. In this report, each patient comment is represented as a list of 500 words with the highest TFIDF scores.

3.1.3 Word embeddings

Comments from the original dataset are mapped to a 384-dimensional embedding using SBERT to reflect the closely related comment semantically in the 384-dimensional space. SBERT is a sentence transformer that "derive semantically meaningful sentence embeddings that can be compared using cosine-similarity" (Reimers, N., & Gurevych, I., 2019).

3.1.4 Embedded method

Certain machine learning models such as Decision Tree or Random Forest have inherent feature selection mechanisms built into their algorithms. Embedded method differs from filter and wrapper method in that it is model-dependent, therefore the selected features are also highly dependent on the model selected (Pudjihartono et al., 2022). With Decision Tree, features are already split based on information gain.

In this report, the performance of classifier with embedded feature selection method via Decision Tree with count vectorized raw text comment as input as this would show whether feature engineering using embedded method would affect classifier performance.

3.2. Classification methods

As the class label for sentiment analysis is binary, several classification models are chosen to be compared in this report.

K-Nearest Neighbours, Gaussian Naïve Bayes and Logistic Regression are selected to investigate the effect of feature engineering on classifiers performance as they do not have embedded feature selection; hence it would demonstrate the difference in performance. K-Nearest Neighbours represents lazy learners, Gaussian Naïve Bayes represents generative models, and Logistic Regression represents discriminative models.

For embedded methods, Decision Tree is used to classify the count vectorized raw text comment as it has built-in feature selection mechanism.

3.3. Evaluation metrics

The performances of all classifiers are evaluated using common evaluation metrics such as accuracy, precision, recall and F-score.

Accuracy can be calculated as the number of true positives and true negatives divided by the number of instances.

Precision is the rate of which true positive instances have been predicted correctly from all instances predicted to be positive.

Recall is the rate of classifying positive instances correctly among the actual number of positive instances.

F-Score is the harmonic mean of recall and precision scores.

Classifiers are trained on the training dataset and tested on a separate validation dataset.

4. Results

In this section, evaluation metrics between each feature selection method for one classifier are compared.

Table 1. Baseline using Zero R

	Accuracy
Baseline	0.73

Table 2. K-Nearest Neighbours with k = 5

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**
None	0.81	0.71	0.49	0.58
Mutual Information	0.85	0.79	0.61	0.69
Chi-Square	0.85	0.79	0.58	0.67
TFIDF	0.78	0.68	0.31	0.42
Word Embedding	0.88	0.84	0.69	0.76

From the table, compared to the benchmark model of no feature selection applied, only TFIDF has lower accuracy, precision, recall and F-Score. In contrast, word embedding method has the highest score across all metrics.

Table 3. Gaussian Naïve Bayes

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**
None	0.41	0.29	0.84	0.43
Mutual Information	0.86	0.76	0.67	0.71
Chi-Square	0.85	0.75	0.65	0.69
TFIDF	0.86	0.67	0.89	0.77

Word Embedding	0.87	0.69	0.89	0.78

Using Gaussian Naïve Bayes as classifier, there are observable improvements in performance when any method of feature selection is applied; this is reflected in the increase in score for all metrics across methods.

Table 4. Logistic Regression

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**
None	0.91	0.87	0.79	0.83
Mutual Information	0.87	0.83	0.65	0.73
Chi-Square	0.87	0.82	0.64	0.72
TFIDF	0.91	0.81	0.85	0.83
Word Embedding	0.92	0.86	0.83	0.85

In Logistic Regression models, both filter methods of Mutual Information and Chi-Square have lower evaluation metrics than the benchmark of no feature selection applied. In contrast, both TFIDF and Word Embeddings using SBERT; which are word representation methods.

Table 5. Decision Tree

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**
Embedded	0.85	0.71	0.76	0.74

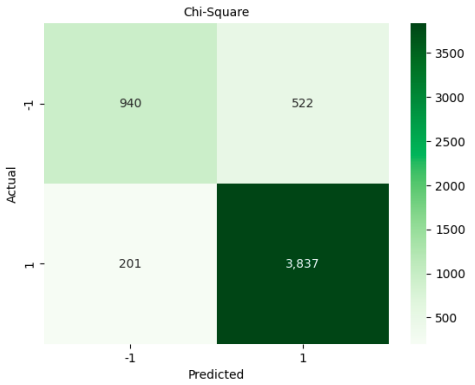
5. Discussion

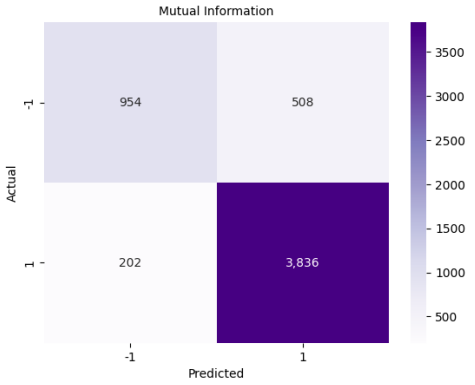
Compared to the Zero R baseline, only the Gaussian Naïve Bayes benchmark classifier has lower accuracy; this means that all other models have learned to predict both class "-1" and "-1" than just assigning the label of the majority class ("1").

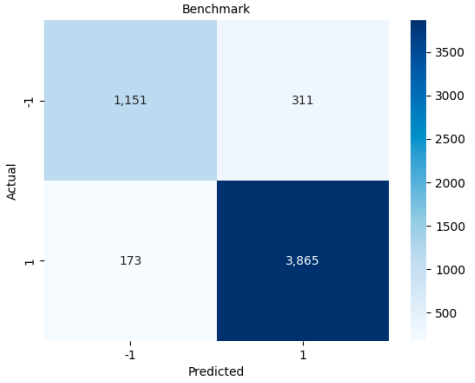
It can be observed that across multiple classifiers, both Mutual Information and Chi-Square, which are filter methods of feature selection, lead to similar performance across different evaluation metrics. This is consistent with findings in Bommert et al., (2020).

With regards to performance in term of accuracy, Gaussian Naïve Bayes classifier benefited the most from utilizing feature selection as there is a clear improvement in all evaluation metrics compared to the benchmark. Due to the assumption of variable independence, it is possible that Gaussian Naïve Bayes performed better after feature selection methods have pruned some features that correlated highly with each other.

Among the Logistic Regression classifiers, models using Mutual Information and Chi-Square have worse performance than the benchmark model, especially concerning recall and precision. This indicates that using filter methods to select the highest-rank features for Logistic Regression led to both higher false negatives and false positives, which can be observed in the respective confusing matrices.







In K-Nearest Neighbours classifiers, all except TFIDF performed better than the benchmark as it has an increase in the number of false negatives which leads to lower accuracy and recall. This could be due to the differences in TFIDF weight of the training and test dataset.

On average, using feature methods led to better evaluation metrics than benchmark classifiers. Comparing different feature selection and engineer methods, filter and word presentations led to better evaluation metrics than embedded method using Decision Tree; however, Decision Tree has higher accuracy, precision, recall and F-score than benchmark KNN and Gaussian Naïve Bayes classifiers.

6. Conclusion

Feature selection and engineering is an integral part of machine learning and is essential in removing data sparsity in text classification. On average, using feature selection led to higher accuracy compared to no feature selection at all. Although there is no clear method that led to improvement in performance metrics across all classifiers used in this report, filter methods work well for models with little to no inherent feature selection and assumption such as KNN and Gaussian Naïve Bayes. Additionally, compared to benchmark KNN and GNB models, embedded method using Decision Tree shows higher evaluation metrics. Feature engineering via word representation methods such as TFIDF and Word Embedding using SBERT led to better improvement in Logistic Regression; with Word Embedding led to improvement in performance for all models used.

7. Ethics Statement

The dataset used in this report is derived from Wallace et al. (2014) and López et al. (2012), which was publicly available when obtained from ratemds.com. The dataset contains no identifiable information except for the gender of the clinicians, therefore, there is a limited misuse potential against the writers of review comments and the clinicians mentioned.

References

[1] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839. doi: https://doi.org/10.1016/j.csda.2019.106839

[2] López, 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, 685-692.

[3] Pintas, J.T., Fernandes, L.A.F. & Garcia, A.C.B. (2021). Feature selection methods for text classification: a systematic literature review. Artif Intell Rev 54, 6149–6200 doi: https://doi.org/10.1007/s10462-021-09970-6

[4] Pudjihartono, N., Fadason, T., W., A., & M., J. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2, 927312. doi: https://doi.org/10.3389/fbinf.2022.927312

[5] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv. /abs/1908.10084. doi: https://doi.org/10.48550/arXiv.1908.10084

[6] 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):1098–1103.

[7] Zhou, H., Wang, X., & Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. Applied Intelligence, 1-18. doi: https://doi.org/10.1007/s10489-021-02524-x