



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Module Title:
Application of Machine Learning

Module Code:
ACCA7025

Assignment Title:
Image classification using ML algorithms.

Student Name:
MD Mahadi Hasan

Student ID:
2505591

Programme:
MSc in Software Engineering and Artificial Intelligence

University Name:
University of Wales Trinity Saint David (Swansea Campus)

Lecturer:
Mark Huntly

Submission Date:
17 April 2026

Table of Contents

Abstract

1.Introduction.....	4
2.Related Work.....	4
3.Methodology.....	5
3.1 Dataset Selection	5
3.2 Data Preprocessing	6
3.3 Feature Extraction	7
3.4 Model Selection.....	7
3.5 Model Training	7
3.6 Model Evaluation	8
 4.Discussion and Recommendations	 8
4.1 Model Comparison	8
4.2 Model Analysis	9
4.3 Recommendation	9
4.4 Proposed Workflow	10
4.5 Evaluation Metrics	10
 5.Conclusion.....	 11

References

Abstract

Sentiment analysis has become an essential analysis tool which is used by companies to understand better what customers think and improve goods and services. Nowadays, online reviews, social media and electronic commerce create huge volumes of unstructured textual data. This information cannot be manipulated in an effective manner by hand, and it is in this regard that automated sentiment classification techniques have been employed. This report discusses the application of traditional machine learning algorithms to sentiment analysis and how it would be applied in the context of computational efficiency and applicability to small marketing organizations with limited resources.

The study tests three of the most popular algorithms: Naive Bayes, Support Vector Machines (SVM), and Logistic Regression. It offers a complete machine learning workflow, including data preprocessing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), model training and model assessment using performance measures. Naive Bayes is believed to be the most appropriate algorithm according to the literature review and analytical comparison, being the simplest, cost-effective and competent in terms of text classification tasks.

The report ends by recommending a smaller, resource-saving sentiment analysis system that can be scaled to small organizations. The results indicate that the conventional machine learning methods are still very effective in sentiment classification in cases where there are computational constraints.

1. Introduction

Customer feedback constitutes a major aspect of business strategies, particularly in the new digital age when customers can easily give their feedback via the internet. Most textual data is generated on websites such as e-commerce sites, social media platforms, and review portals. This information gives some valuable data concerning customer satisfaction, preferences as well as expectations. Nonetheless, it is unfeasible and time-consuming to analyse such large-scale data manually.

One approach to solving this challenge is sentiment analysis also known as opinion mining, which is the automated classification of textual data into groups of sentiment: positive, negative and neutral [2]. It helps organizations to derive actionable insights and make data-driven decisions to enhance their services and products.

Machine learning (ML) has been much more useful in performing sentiment analysis. Neural networks and other deep learning methods have gained popularity due to their high accuracy; however, they are highly computationally intensive, data volumes and implementation. The requirements make them unsuitable in small or new marketing companies that have poor infrastructures [3].

On the other hand, the conventional machine learning algorithms are more convenient and cost feasible. Naive Bayes, Support Vector Machines, and Logistic Regression algorithms are trade-offs between performance and calculation speed. These are easier to operate, less resource consuming and can also offer competitive performance in most sentiment analysis tasks.

2. Related Work

Sentiment analysis is a studied area in the field of natural language processing and a few machine learning approaches have been proposed to accomplish a text classification task. The first studies by Pang et al. (2002) have already determined that machine learning algorithms such as Naive Bayes, Maximum Entropy, and Support Vector Machines could rank sentiment in movie reviews [1]. They showed that SVM was more precise as compared to other techniques and this is partly because it can process high dimensional feature space.

In his subsequent work, Liu (2012) minimised the importance of preprocessing of text and feature engineering with the aim of improving the performance of the classification process [2]. The use of tokenization, stop-word removal and stemming methods were all found to be significant in transforming raw text to useful input to the machine learning models. Noise and irrelevant data can greatly decrease the model accuracy without proper preprocessing.

Support Vector Machines were mentioned to be effective when it came to the classification of text (Joachims, 1998) [3]. SVM is effective particularly when the data is

many dimensional; a common occurrence in text analysis since there is a large word set. Nonetheless, the paper also pointed out that SVM uses more computational resources and parameter adjustments than simpler models.

Recent research has investigated the use of logistic regression as a powerful baseline model of sentiment analysis. Zhang et al. (2010) demonstrated that logistic regression not only provides a competitive performance but also is easy to understand and is simple [4]. The logistic regression is simpler to use and less computationally intensive compared to SVM and thus can be applied in practice.

The feature extraction methods are also important in the sentiment's classification. Manning et al. (2008) outlined the advantages of TF-IDF as compared to the original bag-of-words [5]. Using TF-IDF gives greater weight to significant words and decreases the effect of the common words, resulting in better classification. As it has been demonstrated in comparative studies, TF-IDF boosts the accuracy of the models on sentiment analysis assignments.

Additionally, recent literature indicates that even though deep learning models can be more efficient than traditional algorithms in large-scale tasks, traditional algorithms can be quite helpful in situations where resources are scarce and the size of the data is small. In such cases, naive bayes is typically the option due to the ease of computation and its simplicity of implementation [1][5].

The literature overall indicates that SVM is a very accurate model but the cost of calculations might be too high to be applicable to small organizations. The Logistic Regression is a moderate solution and Naive bayes a good solution in low resources environment where its simplicity and rapid nature are crucial.

3. Methodology

This section outlines a machine learning pipeline sentiment analysis and outlines all the steps involved in data collection to model evaluation.

3.1 Dataset Selection

To train and test machine learning models a labelled dataset is needed. Such that the supervised learning algorithms might learn meaningful patterns between a given input text and sentiment classes. IMDB movie reviews and the Amazon product reviews are popular benchmark data used in sentiment analysis studies. They are large datasets of user-generated text already annotated with sentiment labels (good and bad) and are therefore very well suited to train classification models [6]. The IMDB dataset, such as exemplified by the IMDB dataset, is typically employed in natural language processing research since it is balanced and sentiment annotations are of high quality, whereas Amazon product reviews are domain-specific and thus, provide a wider range of real-world sentiment expressions.

The choice of the right dataset is very important in deciding the overall performance and reliability of the model. An organized dataset is supposed to be as close to real-world data distributions as possible to be able to guarantee that the trained model will be generalized to unseen data. Also, dataset size will affect the learning ability of models, larger datasets generally enable models to learn more complex linguistic patterns and contextual relationships.

Moreover, balance of data sets is also a factor to consider in classification tasks. A well-balanced dataset with equalized positive and negative samples will help to avoid model bias on most of the samples. Where the dataset is unbalanced, it is possible that the model is biased and can obtain a falsely high accuracy by scoring high on most of the classes and low on the minority classes. As such, it is important that data sets are balanced or that methods are employed like resampling or weighting to enhance the fairness and strength of classification [6].

3.2 Data Preprocessing

Preprocessing of text is an essential stage of the machine learning pipeline, particularly within natural language processing (NLP) tasks, like sentiment analysis. Raw text data is normally unstructured, and it has noise, irregularities, spelling, punctuations and other irrelevant symbols that can adversely influence the model performance. As such, preprocessing is also critical to convert raw text into a clean and structured format that can be efficiently processed through machine learning algorithms [2]. Effective preprocessing leads to accuracy in the model, decreased complexity and the quality of the extracted features.

There are a few significant preprocessing methods that are used in machine learning tasks with text. Lowercasing converts all the text into lower cases to make all the text standard so that words like Good and good become the same feature. This removes overlaps of the data and improves uniformity in the extraction of features.

This splitting text into small units such as words, phrases, or tokens, is called tokenization. This is an important step since machine learning models do not operate directly on raw sentences, they need to be fed with structured representations of individual words or tokens to discern patterns in the data.

Stop-word removal is done to remove words that are frequent like the, is, and, and a which in most of the cases do not have much semantic value. The elimination of these words aids in lowering noise in dataset and is more efficient in computing since attention to meaningful terms is given.

A reduction in words to their base or root form can be achieved by stemming and lemmatization. Stemming yields the suffixes in a rule-based fashion, whereas lemmatization takes the contextual meaning of words, to transform them into their

dictionary forms. The methods can be used to normalize the dataset and dimensionality of features, thereby enhancing the performance of the model and its capacity to generalize [2][5].

In general, these preprocessing procedures are critical to enhance the quality of input data allowing machine learning models to pay attention to the pertinent linguistic patterns and attain higher classification results.

3.3 Feature Extraction

Machine learning models require numerical data. Therefore, the textual data should be converted into feature vectors. There are two Bag-of-Words and TF-IDF.

The TF-IDF is selected in this study due to its ability to provide words with their meaning in respect to their frequency and relevance. It reduces the effect of common words and emphasizes on exceptional words that assist in the categorization of sentiment [5].

3.4 Model Selection

Three classical machine learning algorithms are experimented:

Naive Bayes

Naive Bayes is a probabilistic classifier which is based on the Bayes theorem. It is also feature-independent and therefore highly efficient in computing as it is very easy to compute and classify text [1].

Support Vector Machine (SVM)

SVM is a supervised learning algorithm which identifies an optimal decision boundary between classes. It is also typified with high accuracy and is more challenging about computational resources and parameter tuning [3].

Logistic Regression

Logistic regression is a binary statistical model of classification. It is simple, readable and works on large number of sentiment analysis tasks [4].

3.5 Model Training

The data is frequently divided into a training and testing subset in 80/20 split. The model is fed with the training data and the testing data is fed with the model to determine the generalization performance of the model [6]. Cross-validation techniques can also be used to assure reliability and overfitting.

3.6 Model Evaluation

Standard evaluation measures are also commonly applied to determine the performance of classification models and give a rich insight into the performance of a model in various dimensions of prediction. The metrics are needed because it can be deceptive to solely rely on accuracy, particularly when dealing with imbalanced datasets as one category can be dominant over the rest [5].

Accuracy is the total accuracy of the model predictions. It is computed as the number of correctly predicted instances, whether positive or negative, to the total number of instances. Although accuracy gives a rough picture of the performance, it is not necessarily a measure of the real performance of a model in case the data distribution is skewed.

Precision is a measure of the fraction of positive instances correctly predicted amongst all positive instances predicted. It is concerned with the accuracy of positive predictions and is especially significant in those applications where false positives are highly expensive such as spam detection or fraud detection.

Recall measures how well the model identifies all real positive cases correctly. It quantifies the degree to which the model can capture positive examples in the dataset. High recall is essential in situations where false negative is not as important as false negative.

The F1-score is a balanced estimate of precision and recall as it is the harmonic mean of the two. It is particularly handy when precision and recall have a trade-off because it guarantees that both measures are considered. F1-score is commonly desirable in text classification and sentiment analysis problems where the imbalance of the data is frequent [5].

All in all, this evaluation metrics combined can give a comprehensive and accurate analysis of classification performance which can be more readily compared and selected machine learning models to apply in practice.

4. Discussion and Recommendations

4.1 Model Comparison

Model	Accuracy	Speed	Complexity
Naive Bayes	Good	Very Fast	Low
SVM	Very High	Slow	High
Logistic Regression	Good	Medium	Medium

4.2 Model Analysis

Naive Bayes has always displayed good performance in text classification activities though it has a simplifying assumption of independence of features. Practically, the assumption is entirely accurate in natural language data; nevertheless, the algorithm still works well since the presence of words in sentiment analysis tend to give good probabilistic hints. This is because its primary strength is that it has a very low computational cost, it can be trained very quickly and scales to very large datasets, which makes it very applicable to real-time processes for example spam detection systems and online sentiment monitoring systems [1]. Also, Naive Bayes excels in high-dimensional but sparse data set and this is typical of text representation algorithms like Bag-of-Words and TF-IDF.

In general, Support Vector Machines (SVM) can have a higher classification accuracy than the simpler probabilistic models since they are able to build optimal separating hyperplanes in feature spaces with many dimensions. This enables SVM to deal with non-linear and complex decision boundaries particularly when there is the use of kernel functions. This improved performance is however at the expense of greater computational complexity and increased training time. Consequently, SVM may not be practical in cases where large datasets or systems with limited computing power are used where efficiency is a very important aspect [3].

Logistic Regression is an alternative to Naive Bayes and SVM which offers a middle ground. Although its predictive accuracy is sometimes a bit less than that of SVM in complicated classification issues, it has a powerful interpretability and comparatively easy computation. Logistic Regression is easier to run and understand because the results are probabilistic that can be easily translated to the probability of each class. This is especially handy in scenarios where transparency and explainability of models are relevant like in marketing analytics and customer sentiment interpretation. In general, a pragmatic trade-off between performance and computational efficiency in a wide range of machine learning applications in the real world [4].

4.3 Recommendation

Based on the analysis, the most suitable algorithm to be implemented in the marketing firm is the Naive Bayes classifier. The model provides a good compromise between the performance in classification and the computation efficiency and thus can be well applied in environments that are resource constrained. Naive Bayes compared to more complicated algorithms uses relatively low processing power and memory to train and make predictions and thus it takes much less time to train and predict without much accuracy loss [1].

Moreover, its probabilistic nature enables it to be very effective when it comes to text classification like sentiment analysis even with high-dimensional data. Research has

revealed that the Naive Bayes algorithm with its simplifying assumption of feature independence can be used to achieve high-performance when used alongside good preprocessing and feature extraction methods like TF-IDF [2][5]. Thus, given the limitations of a small marketing organisation, the Naive Bayes algorithm offers an effective, scalable and affordable sentiment analysis solution, without compromising the quality and reliability of its performance.

4.4 Proposed Workflow

The suggested workflow comprises:

1. Data collection
2. Text preprocessing
3. TF-IDF feature extraction.
4. Train Naive Bayes model.
5. Model evaluation
6. Deployment

This is a scalable, efficient and simple workflow.

4.5 Evaluation Metrics

Evaluation metrics are very important in determining the performance and reliability of the machine learning model in sentiment analysis tasks. One of the most common metrics is accuracy which is the ratio of rightly classified cases of the overall number of predictions. Although accuracy is a general measure of the performance of a model, it is not always the case especially when using an imbalanced dataset whereby a single type of data greatly dominates others [5].

Precision and recall provide more information on the quality of classification. Precision is the measurement of the ratio of the number of predicted positive instances that are correctly predicted to the total number of instances that are predicted to be positive; thus, it is a measure of how the model can avoid false positives. Recall measures the percentage of the actual positive cases that the model correctly recognises which is a measure of how well it reduces false negatives [6]. Such measures are particularly crucial to sentiment analysis where false classification of sentiments may result in erroneous business results.

The F1-score is a score that integrates both precision and recall in one score by taking their harmonic mean to give a balanced score on the model performance. This is especially effective in situations that have unequal distribution of classes since it takes into consideration both false positives and false negatives [5]. As a result, the F1-score

is generally considered a strong and valid indicator of the sentiment classification models.

5. Conclusion

This report has given a detailed and systematic study of machine learning methods of sentiment analysis with specific focus on whether it is suitable in a low resource setting. The entire machine learning process was designed as an end-to-end system which included major steps, like data preprocessing, feature extraction with TF-IDF, model selection, training and evaluation. All parts of the pipeline were critically to achieve both computational efficiency and classification effectiveness, a reflection of best practices that were found in existing literature [2][5].

The relative comparison of the various algorithms brings out significant trade-offs in the accuracy and computational cost. Support Vector Machines (SVM) has been known to have a high classification accuracy, especially in high-dimensional text data. Nevertheless, their resource-consumption and computing complexity limits their applicability to small-scale organisations with limited infrastructure [3]. The Logistic Regression offers a more moderate way out still with a moderate accuracy but with better interpretability and less computational cost; recent research confirms this [4].

In contrast, the Naive Bayes algorithm comes out as the right solution to the given situation. Its probabilistic model limited computational needs and speedy processing functionalities render it remarkably useful in text classification on massive scale. Naive Bayes has been demonstrated to provide competitive results in sentiment analysis applications, even with its simplifying assumption of feature independence, especially when used in conjunction with powerful preprocessing and feature extraction methods [1][5].

Moreover, the suggested workflow provides a viable and scalable model that could be easily adopted by marketing organisations that aim to analyse customer feedback effectively. Through this strategy, companies can turn the unstructured textual data into valuable insights which will support data-driven decision-making and strategic enhancements.

Overall, this paper shows that machine learning approaches are still very useful and applicable to the modern sentiment analysis, particularly in the situation when computational sources are limited. This makes them a good base in the real-life applications due to their simplicity, efficiency and effectiveness as demonstrated in actual practice [2][5].

References

- [1] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” EMNLP, 2002.
- [2] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool, 2012.
- [3] T. Joachims, “Text categorization with Support Vector Machines,” ECML, 1998.
- [4] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: a statistical framework,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 4, pp. 550–561, Apr. 2010.
- [5] C. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
- [6] J. Jurafsky and J. Martin, Speech and Language Processing, 3rd ed., 2020.
- [7] Scikit-learn, “Text feature extraction,” [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html.