Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis

3 authors, including:

Márcio Guia
Instituto Superior de Engenharia de Coimbra

2 PUBLICATIONS   **7** CITATIONS

Jorge Bernardino
Polytechnic of Coimbra - Portugal

251 PUBLICATIONS   **1,403** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project eAGORA? View project

Project Software Engineering Teaching View project

Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis

Márcio Guia[1], Rodrigo Rocha Silva[2,3] [a] and Jorge Bernardino[1,2] [b]

[1]Polytechnic of Coimbra – ISEC, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal
[2]CISUC – Centre of Informatics and Systems of University of Coimbra, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
[3]FATEC Mogi das Cruzes, São Paulo Technological College, 08773-600 Mogi das Cruzes, Brazil

Keywords: Data Mining, Sentiment Analysis, Text Classification, Naïve Bayes, Support Vector Machine, Random Forest, Decision Trees.

Abstract: Every day, we deal with a lot of information on the Internet. This information can have origin from many different places such as online review sites and social networks. In the midst of this messy data, arises the opportunity to understand the subjective opinion about a text, in particular, the polarity. Sentiment Analysis and Text Classification helps to extract precious information about data and assigning a text into one or more target categories according to its content. This paper proposes a comparison between four of the most popular Text Classification Algorithms - Naive Bayes, Support Vector Machine, Decision Trees and Random Forest - based on the Amazon Unlocked mobile phone reviews dataset. Moreover, we also study the impact of some attributes (Brand and Price) on the polarity of the review. Our results demonstrate that the Support Vector Machine is the most complete algorithm of this study and achieve the highest values in all the metrics such as accuracy, precision, recall, and F1 score.

1 INTRODUCTION

Text Mining is the process that can extract valuable information from a text (Mouthami, Devi and Bhaskaran, 2013). One of many applications of Text Mining is Sentiment Analysis, which is the process used to determine the opinion or the emotion that a person writes about an item or topic (Mouthami, Devi and Bhaskaran, 2013).

With the growth of the Internet, especially social networks, people can easily express their opinion about any topic in a few seconds, and valuable information can be extracted from this, not only about the person who wrote it but also about a particular subject.

There are three categories to classify Sentiment: Machine Learning, Lexicon-Based and an hybrid that combines Machine Learning and Lexicon- Based (Ahmad, Aftab and Muhammad, 2017). In literature, the Machine Learning categories to extract Sentiment are one of the most discussed areas and for this reason, in this paper, we propose to do a comparison between four of the most popular Machine Learning algorithms: Naive Bayes (Kononenko, 1993), Support Vector Machine (Cortes and Vapnik, 1995), Decision Trees (Quinlan, 1986) and Random Forest (Ho, 1995). In order to evaluate these classifiers, we use Amazon Reviews: Unlocked Mobile Phones dataset and our focus goes to the Polarity Review of a text, which can be Negative or Positive.

The main contributions of this work are the following:

- Compare Naive Bayes, Support Vector Machine, Decision Trees and Random Forest on Polarity Text Review based on Accuracy, Precision, Recall, and F1 score;
- Compare different types of each studied classifier models;
- Evaluate the impact of Brand and Price of the mobile phones on final Polarity Review.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 describes the experimental approach. Section 4 presents the

[a] https://orcid.org/0000-0002-5741-6897
[b] https://orcid.org/0000-0001-9660-2011

results and discussion. Finally, Section 5 concludes the paper and presents future work.

2 RELATED WORK

Sentiment Analysis has been utilized by many authors to classify documents, especially with machine learning approaches. However, the researches usually just focus on one of the most popular machine learning algorithms like the Support Vector Machine, Naïve Bayes or Random Forest classifier.

(Moe et al., 2018) compares Naïve Bayes with Support Vector Machine on Document Classification. The authors conclude that Support Vector Machine is more accurate than Naïve Bayes classifier.

(Xu, Li and Zheng, 2017) defend that although Multinomial Naïve classifier is commonly used on Text Classification with good results, it's not a fully Bayesian Classifier. So, the authors propose a Bayesian Multinomial Naïve Bayes classifier and the results show that the new approach has similar performance when compared to the classic Multinomial Naïve Bayes classifier.

(Manikandan and Sivakumar, 2018) propose an overview of the most popular machine learning algorithms to deal with document classification. The authors provide the advantages and main applications of each algorithm. However, this paper does not provide any practical study about the algorithms and does not do a comparison between them.

(Rodrigues, Silva and Bernardino, 2018) propose a new ontology to deal with social event classification. Instead of label an event with just one category the authors propose a classification based on tags. So, an event can have more than one tag and this approach can more successfully achieve the interest of a user. To make the classification tests the authors use the Random Forest Classifier which achieve good results. However, to do the classification the authors have just use one algorithm.

(Parmar, Bhanderi and Shah, 2014) study Random Forest classifier on Sentiment Analysis. The authors proposed an approach that tunes the hyperparameters like number of trees to construct the Decision Forest, number of features to select at random and depth of each tree. They conclude that with optimized hyperparameters the Random Forest classifier can achieve better results. In (*Text Mining Amazon Mobile Phone Reviews: Interesting Insights*, no date) the authors of the dataset that we use in this paper provided a statistical study about the relationship between the attributes of the dataset and they also extract the sentiments that are present in the reviews.

In our paper we also added some statistical study to the one done initially by the authors of the database by study the impact of brand and price in the polarity review.

The main difference of these works with ours is that we don´t focus on just one machine learning algorithm. We propose a comparison between four algorithms: Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest. Besides none of these works studies the impact of the attributes of the dataset in the classification of documents.

3 EXPERIMENTAL APPROACH

This section presents the experimental approach used for the classification task, Fig.1 displays the overall architecture. The proposed architecture consists of five parts. The first one deals with cleaning the dataset, described in section 3.1. After cleaning dataset, we do Pre-Processing and Text Transformation, all these steps are described in section 3.2. In section 3.3 we describe the classification process. The Evaluation process and the compare of results are described in section 4.
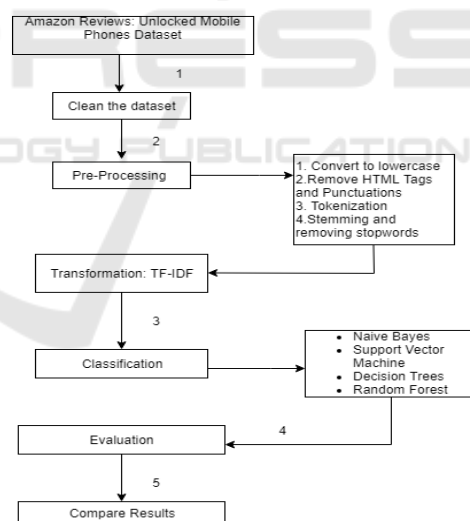

Figure 1: Overview of our approach.

3.1 Dataset

The dataset that we use for this study (*Amazon Reviews: Unlocked Mobile Phones | Kaggle*, 2016) consists of 400 000 reviews of unlocked mobile phones sold on Amazon.com and contains attributes such as Brand (string), Price (real number), Rating (integer number) and Review text (string).

For the classification task, we only select the Rating and Review text attributes. Rating is a numerical value from 1 to 5, and the Review text is a String which contains the opinion of the user. Before using the dataset, we apply a few steps to get better results. These steps are described as following:

1. Assign Rating value 1 and 2, to Negative;
2. Assign Rating value 4 and 5, to Positive
3. Remove all the instances that contain a Rating value equal to 3.

3.2 Pre-processing and Text Transformation

In order to improve results for the four algorithms that we study in this paper, it is necessary to do some pre-processing steps which will make it possible to reduce data dimension without affecting the classification task (Eler et al., 2018). The first step is to convert all the instances of the dataset into lowercase. Next, we remove some noisy formatting like HTML Tags and Punctuation. Tokenization, removal of stop words and stemming are described as follows:

- Tokenization: is the process that splits strings and text into small pieces called tokens (Mouthami, Devi and Bhaskaran, 2013). This process is widely used and popular in pre-processing tasks.

- Removal of Stop Words: A stop word is a commonly used word that appears frequently in any document. These words are usually articles and prepositions. An example of these terms is "the"," is"," are", "I" and "of" (Eler et al., 2018). Hence, we can say that these terms do not add meaning to a sentence, and for this reason, we can retrieve them from the text before doing the classification task. For this study, we use a list of common words of the English Language which includes about 150 words.

- Stemming: is the process that reduces a word to their base or root form. For example, the words "swimmer" and "swimming" after the stemming process are transformed into "swim". In this study, we use the Porter Stemmer because is one of the most popular English rule-based stemmers (Jasmeet and Gupta, 2016) and compared with Lovins Stemmer it´s a more light stemmer. Moreover, produces the best output as compared to other stemmers (Ganesh Jivani, 2011).

Text Transformation: Machine learning algorithms do not work with text features, so, for this reason, we need to convert text into numerical features. To deal with that, we use the TF-IDF (Term Frequency-Inverse Document Frequency). This algorithm assigns to each word of the sentence a weight based on the TF and IDF (Yang and Salton, 1973).

The TF (term frequency) of a word is defined as the number of times that the word appears in a document.

The IDF (inverse document frequency) of a term is defined as how important a term is (Salton and Buckley, 1988) (Yang and Salton, 1973).

3.3 Classification Process

After cleaning the dataset and apply pre-processing and text transformation steps, we split the data into training and test. The percentage used for training is 80% and the remaining 20% are used for test. It is necessary to feed the classification algorithms, so the train data will be used for training the classifiers and the test data will be used to evaluate them. The four classifiers that we use are described in the following:

- Random Forest: is defined as a classifier with a collection of tree-structured classifiers {h(x, k), k = 1,...} where the {k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. When a large number of trees is generated each one of them will vote for a class, and the winner is the class that has more votes (Breiman, 2001). For this study we evaluate the Random Forest classifier with a different number of trees to construct the Decision Forest, in particular, we test the classifier with 50,100,200 and 400 trees.

-Naive Bayes: is a probabilistic machine learning classifier based on the Bayes Theorem with an assumption of independence among predictors, in other words, this algorithm considers that a presence of a feature in a class is independent of any other features (Ahmad, Aftab and Muhammad, 2017). For this study we evaluate two types: Multinomial and Bernoulli.

Support Vector Machine: is a supervised learning model which can achieve good results in text categorization. Basically this classifier locates the best possible boundaries to separate between positive and negative training samples (Ahmad, Aftab and Muhammad, 2017) For this study, we evaluate two distinct kernel models for Support Vector Machine: RBF and Linear (Minzenmayer et al., 2014) .

Decision Trees: is an algorithm that use trees to predict the outcome of an instance. Essentially, a test node computes an outcome based on the attribute values of an instance, where each possible outcome is associated with one of the subtrees. The process of classify an instance starts on the root node of the tree. If the root node is a test, the outcome for the instance

it is predicted to one of the subtrees and the process continues until a leaf node it is encountered, in this situation the label of the leaf node gives the predicted class of the instance (Quinlan and Quinlan J. R., 1996).

4 EXPERIMENTAL EVALUATION

We use the Amazon Reviews: Unlocked Mobile Phones dataset (*Amazon Reviews: Unlocked Mobile Phones | Kaggle*,2016) and we split the dataset into 80 % for train and 20% for the test. As mentioned, before we provide a comparison between four algorithms and also offer a statistical about the impact of the brand and the price in the final polarity review. These experiments are described as follows:

4.1 Algorithms Classification

In order to evaluate the results of the four algorithms we use four of the most popular measures: Accuracy, Precision, Recall, and F1 score. These four metrics are explained in the following:

- Accuracy: is the most popular measure and also very easy to understand because is a simple ratio between the number of instances correctly predicted to the total number of instances used in the observation, in other words, accuracy gives the percentage of correctly predicted instances (Mouthami, Devi and Bhaskaran, 2013).
- Precision: is a measure that provides for each class the ratio between correctly positive predicted instances and total of positive instances predicted (Mouthami, Devi and Bhaskaran, 2013).
- Recall: is a measure that provides for each class the ratio between the true positive instances predicted and the sum of true positives and false negatives in the observation (Mouthami, Devi and Bhaskaran, 2013).
- Fl score: is the weighted average of Precision and Recall (Mouthami, Devi and Bhaskaran, 2013), and it's considered perfect when it´s 1.0 and the worst possible value is 0.0, so a good F1 score means that we have low false positives and low false negatives.

4.2 Naive Bayes

Table 1 shows the results of application Naive Bayes on the dataset. The first experimental for the Naive Bayes classifier was the Multinomial variant. The results demonstrated that the classifier obtains 0.83 which means that in 83% of times the polarity reviews

was correctly predicted. Precision and Recall obtain similar values, 0.84 and 0.83 respectively, F1 score obtains 0.80. The second experimental was with Bernoulli variant and the results show an improvement of 2% for Accuracy and Recall and 4% for F1 score.

In conclusion, the two variants of Naive Bayes can both achieve good results in Sentiment Analysis especially the Bernoulli Variant. However, the Naive Bayes classifier when compared to Random Forest and especially Support Vector Machine obtain modest results.

Table 1: Results for the measures of application Naïve Bayes on the dataset.

	Accuracy	Precision	Recall	F1 score
Multinomial	0.83	0.84	0.83	0.80
Bernoulli	0.85	0.84	0.85	0.84

4.3 Random Forest

Table 2 shows the results of application Random Forest on the dataset. When the number of estimators was 50 the classifier obtains 0.87 for Accuracy, Precision, Recall and F1 score, which can be considered a good result considering the small number of estimators. When the numbers of estimators were 100 the results demonstrate an increment of 1% for Accuracy and Recall, and the Precision and F1 score remained the same values. The results for the third experimental test with 200 estimators for the Random Forest classifier demonstrate that Precision achieves 0.88 which is more 0.1% than the experimental with 100. Finally, in the last experimental, the number of estimators was 400 and the results show that with this high number of estimators the results for all the measures are still equal to the experiment with 200 estimators.

In conclusion, the results for the application of Random Forest classifier show that this algorithm can achieve high values for all the measures even when the number of estimators is low, it means that Random Forest can be used with success on text classification tasks. It is also possible to conclude that when the number of estimators increases the Precision, Recall and Accuracy also increases. However, the best result of Random Forest was with 200 estimators. Increasing the number of estimators did not achieve better results.

Table 2: Results for the measures of application Random Forest on the dataset.

	Accuracy	Precision	Recall	F1 score
50 estimators	0.87	0.87	0.87	0.87
100 estimators	0.88	0.87	0.88	0.87
200 estimators	0.88	0.88	0.88	0.87
400 estimators	0.88	0.88	0.88	0.87

4.4 Support Vector Machine

Table 3 shows the results of application Support Vector Machine on the dataset. As mentioned before we use two types of kernel models to evaluate the Support Vector Machine. The first experimental evaluation demonstrates that with Linear kernel, the classifier obtains 0.89 for Accuracy, Precision, Recall and F1 score which means that 89% of the times the classifier predicted correctly the polarity of a review. The second experimental demonstrates that with RBF Kernel the results obtained are significantly lower than the results with Linear Kernel, namely, the results for Accuracy and Recall decrease 16 %, the value of Precision drastically decreases 36 % and the value of F1 score decreases 28%.

In conclusion, the Support Vector Machine with Linear Kernel achieves the best results of this study and proves that it is one of the best algorithms to deal with Sentiment Analysis. However, the poor results of the application of Support Vector Machine with RBF kernel demonstrate that the latter it is not a good classifier for Sentiment Analysis.

Table 3: Results for the measures of application Support Vector Machine on the dataset.

	Accuracy	Precision	Recall	F1 score
Linear	0.89	0.89	0.89	0.89
RBF	0.73	0.53	0.73	0.61

4.5 Decision Trees

Table 4 shows the results of the application of Decision Trees on the dataset. The results show that the Decision Trees classifier obtains the same value (0.82) for all the four measures: Accuracy, Precision, Recall, and F1 score. These results are similar to the Multinomial Naive Bayes and we can conclude that Naive Bayes and Decision Trees achieve similar values in the Sentiment Analysis task which can be

explained by the lower complexity of these two algorithms when compared to Random Forest and Support Vector Machine.

Table 4: Results for the measures of application Decision Trees on the dataset.

	Accuracy	Precision	Recall	F1 score
Decision Trees	0.82	0.82	0.82	0.82

4.6 Impact of Brand and Price

In this study, we also make a statistical comparison of the impact of attributes (brand and price) in the final polarity review. For brand, we study the most popular brands of phones that are present in the dataset and for price we provide an overview of all the prices that are presented in the dataset.

4.6.1 Brand

Table 5 shows the impact of the brand in the polarity review. After having analyzed these results we conclude that the impact of the brands is similar and is in a range of 77% to 79%. However, there are two brands which stand out from the rest. The first one is the BlackBerry with only 74.3 % positive reviews. The second one is ZTE which has the best results with 82.9% positive reviews. We think that the significant difference in the percentage of positive reviews between BlackBerry and ZTE could be explained by a phone model from BlackBerry that has the potential to give problems or does not match customer expectations and the high results of ZTE can be explained by the fewer models that are present in the dataset.

Table 5: Results for the impact of the brand on polarity review.

Brand	% of reviews	
	Positive	Negative
Samsung	79.94	20.06
Apple	77.3	22.7
Nokia	78.01	21.99
BlackBerry	74.3	25.7
Asus	77.41	22.59
LG	77.2	22.8
Sony	79.86	20.14
ZTE	82.9	17.1

4.6.2 Price

Table 6 shows the impact of the price in the polarity review. After having analyzed these results we conclude that there's a significant difference between the range of fewer than 100 dollars (73.2 % of positive reviews) and the range of 1000 to 1500 dollars (84.3% of positive reviews). It's also possible to conclude that as the price range increase the percentage of positive reviews also increases reaching the maximum in the range of 1000 to 1500 after that the percentage of positive reviews falls by one percentage point to 83.3 %. These results can be explained by the quality of the phones, it means that products with a lower price may have less quality than products with high price, which have more features and also more quality. Hence it is expected that as the price increases the percentage of positive reviews also increases.

Table 6: Results for the impact of price on polarity review.

Price (Dollars)	% of reviews	
	Positive	Negative
Less than 100	73.2	26.8
100 to 200	76.8	23.2
200 to 300	79.1	20.9
300 to 400	79.2	20.8
400 to 500	81.4	18.6
500 to 1000	81.4	18.6
1000 to 1500	84.3	15.7
1500 to 2000	83.3	16.7
Above 2000	83.3	16.7

5 CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed four of the most popular machine learning algorithms to deal with Sentiment Analysis, based on four measures: Accuracy, Precision, Recall, and F1 score. We found that the Support Vector Machine classifier is not only the most accurate of this study but also the most complete classifier with high values to all the measures. Our results show that Random Forest is also a classifier to take into account and can achieve high values to all the measures being just slightly worse than the Support Vector Machine classifier.

This study also proposes a statistical study about the impact of brand and price in the polarity review and concludes with some interesting facts about each one of these attributes. For the brand, we can have an overview of the impact of each brand in the polarity review and concluded that ZTE is the brand with the most positive reviews with 82.9 %, as opposed to BlackBerry with just only 74.3 %. For the price, we can conclude that as the price increases the percentage of positive reviews also increases, reaching a maximum of positive reviews in the range of 1000 to 1500 dollars after that the percentage of positive reviews falls from 84.3% to 83.3 %.

As future work, we plan to continue the study of other algorithms that are usually applied to Sentiment Analysis and evaluate them with the measures that we used in this study. We also plan to propose an architecture to improve the results of each one of the four algorithms that we evaluated and compared in this study.

REFERENCES

Ahmad, M., Aftab, S. and Muhammad, S. S. (2017) 'Machine Learning Techniques for Sentiment Analysis: A Review', *International Journal of Multidisciplinary Sciences and Engineering*, 8(3), p. 27.

Amazon Reviews: Unlocked Mobile Phones | Kaggle (no date). Available at: https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones (Accessed: 21 March 2019).

Breiman, L. E. O. (2001) '18_Breiman.pdf', pp. 5–32. doi: 10.1023/A:1010933404324.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(3), pp. 273–297. doi: 10.1007/BF00994018.

Eler, M. D. et al., (2018) 'Analysis of Document Pre-Processing Effects in Text and Opinion Mining', *Information* . doi: 10.3390/info9040100.

Ganesh Jivani, A. (2011) 'A Comparative Study of Stemming Algorithms', *International Journal of Computer Technology and Applications*, 2(6), pp. 1930–1938. doi: 10.1.1.642.7100.

Ho, T. K. (1995) 'Random decision forests', in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.

Jasmeet, S. and Gupta, V. (2016) 'Text Stemming: Approches', *ACM Computing Surveys*, 69(3), pp. 633–636.

Kononenko, I. (1993) 'Successive Naive Bayesian Classifier.', *Informatica (Slovenia)*, 17(2).

Manikandan, R. and Sivakumar, D. R. (2018) 'Machine learning algorithms for text-documents classification: A review', *International Journal of Academic Research and Development*, 3(2), pp. 384–389. Available at: www.academicsjournal.com.

Minzenmayer, R. R. et al., (2014) 'Evaluating unsupervised and supervised image classification methods for mapping cotton root rot', *Precision Agriculture*, 16(2), pp. 201–215. doi: 10.1007/s11119-014-9370-9.

Moe, Z. H. et al., (2018) 'Comparison Of Naive Bayes And Support Vector Machine Classifiers On Document Classification', in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pp. 466–467. doi: 10.1109/GCCE.2018.8574785.

Mouthami, K., Devi, K. N. and Bhaskaran, V. M. (2013) 'Sentiment analysis and classification based on textual reviews', *2013 International Conference on Information Communication and Embedded Systems, ICICES 2013*. IEEE, pp. 271–276. doi: 10.1109/ICICES.2013.6508366.

Parmar, H., Bhanderi, S. and Shah, G. (2014) *Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters*.

Quinlan, J. and Quinlan J. R. (1996) 'Learning decision tree classifiers', *ACM Computing Surveys (CSUR)*, 28(1), pp. 2–3. Available at: http://dl.acm.org/ citation.cfm?id=234346.

Quinlan, J. R. (1986) 'Induction of Decision Trees', *Machine Learning*, 1(1), pp. 81–106. doi: 10.1023/A:1022643204877.

Rodrigues, M., Silva, R. R. and Bernardino, J. (2018) 'Linking Open Descriptions of Social Events (LODSE): A new ontology for social event classification', *Information (Switzerland)*, 9(7). doi: 10.3390/info9070164.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management*. Pergamon, 24(5), pp. 513–523. doi: 10.1016/0306-4573(88)90021-0.

Text Mining Amazon Mobile Phone Reviews: Interesting Insights (no date). Available at: https://www. kdnuggets.com/2017/01/data-mining-amazon-mobile-phone-reviews-interesting-insights.html (Accessed: 19 May 2019).

Xu, S., Li, Y. and Zheng, W. (2017) *Bayesian Multinomial Naïve Bayes Classifier to Text Classification*. doi: 10.1007/978-981-10-5041-1_57.

Yang, C. S. and Salton, G. (1973) 'On the specification of term values in automatic indexing', *Journal of Documentation*. Emerald, 29(4), pp. 351–372. doi: 10.1108/eb026562.