Star rating prediction from user reviews

Harshal Patel

hapt18@student.bth.se

Department of Computer Science
Blekinge Institute of Technology

Karlskrona, Sweden.

ABSTRACT

This report is about the classification of reviews using support vector machines (SVM) and Naïve Bayes classifier into categories of stars ranging from 1 to 5 as well as binary classification (positive and negative review) This review dataset is obtained from yelp datasets containing 1.5 million reviews over all with their star rating, data for specific region is used for classification in this project. The raw data is first preprocessed to remove all the meaningless words as well as stop words, then split into training and testing data. The performance of the algorithm is evaluated, compared based on accuracy and confusion matrix.

PROBLEM ADDRESSED

Users share reviews about almost anything, as a feedback for the owner. This review or bunch of reviews can contribute to the average rating of the product, restaurant, application, location, etc. which can be further used to analyze the quality of the product. For example, YouTube videos can have ratings on each video based on the comments received on it.

INTRODUCTION

In this democratic era, everyone is willing to share their opinion about everything and twitter and Facebook is a platform that allows one to do so, at an international level. Reviews on any locations, restaurants, products, movies, applications, etc. can be helpful for other users. These opinions or reviews generates a gigantic amount of data. This can be processed to find general likes and dislikes of people. These reviews with star ratings are being used for supervised learning in this project. However, working with textual data has always been a challenge due to constant use of slang words, sarcasm, and misspellings, this calls for data pre-processing before using Natural Language Processing (NLP) techniques. Hence, main part of text analysis or text classification is the processing done on the data to extract features from it. Here, we use reviews categorized according to their star ratings to train a model to be able to classify the class it belongs to and find the overall accuracy of the algorithm. To do so, we use dataset from yelp that has 1.5 million instances, data for Las Vegas is used in this project as it has most number of reviews [7] which will be enough for a descent training of model, this model can then be used in areas where finding out stars is not an option to judge the quality. Here, comparison between support vector machines (SVM) and Naïve Bayes algorithm is done. Performance index for each algorithm is accuracy and confusion matrix.

Also, binary classification is explored using two algorithms, to understand their performance in further depth. The dataset used is divided into 80% training and 20% testing, after pre-processing raw text reviews. Accuracy is then calculated using fit function to map the text review data with the target review stars.

RELATED WORK

In [1] Huang, J.; Rogers, S.; Joo, E. used restaurant rating data from yelp datasets to improve their business by gaining insights. They used Latent Dirichlet Allocation (LDA) algorithm to discover subtopics from reviews. Hidden topics in a review are broken down and stars are predicted for every hidden topic. This helped in finding the insights from the review that can be helpful for the vendors. They found out that the user area of interests are service, value, take out and décor of the restaurants. Also, it shows that the user is less likely to give high stars on peak times, that must be because of the waiting time for the customer.

In [2] Neethu, M.S., et al is to focus on creating an efficient feature vector after pre- processing, to deal with issues like tweets having multiple keywords. Using different classifiers (Naive Bayes, SVM, Maximum Entropy and ensemble classifier) to test the accuracy of feature vector in the electronic product domain. From this paper we know that how pre-processing is done and creating a feature vector in order to avoid multiple keywords to improve the accuracy.

In,[3] Tripathy et al, in this paper sentimental reviews, movie dataset is used to classify the comments, using SVM and naive Bayes algorithm. Firstly, a dataset containing 1000 positive and 1000 negative reviews are considered and preprocessed by removing stop words, blank spaces, numerical and special characters. Vectorization is done i.e. to convert textual data into numerical format. Training of model using sparse matrix. Finally, classification is done to get the result. Next, comparison of proposed work with existing literature is shown and how it is better. In conclusion, SVM was found to be the most efficient algorithm. from this paper we can conclude that comparison between the two machine learning algorithms and SVM found to be better when compared to Naive Bayes and how textual data is converted into numerical data. Which helps to improve the accuracy.

In [4], N. Bindal et al argue as the tweets are not in formal languages it is difficult to classify using the model which was trained by formal English language. So, 2 approaches were identified. Firstly, a sentiment lexicon was built having sentiment polarity and numeric value based on its strength with the help of emoticons which were used as labels. Secondly various machine learning algorithms and data preprocessing steps were used to improve the performance of the classification system. From this paper we learn how sentiment polarity is built and how emoticons are removed will helps to improve the accuracy.

IMPLEMENTATION

A. Dataset

This experiment is performed on yelp dataset challenge [5]. This dataset includes business, review, user, and check-in data in the form of separate JSON objects [1]. Data used has rating, review text, user id and date of review. This data is used by object creation having above mentioned attributes. This data will help us identify the attributes that belong to each class, these can be further used to identify the rating based on a given comment. The dataset as a whole is vast, so reduction of data is performed by filtering data to a particular state area and using reviews ranging from 100-200 words [7].

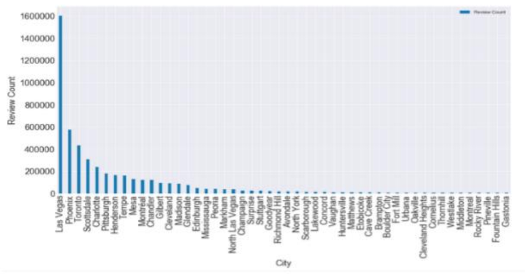


Fig 1. Distribution of reviews based on cities [7]

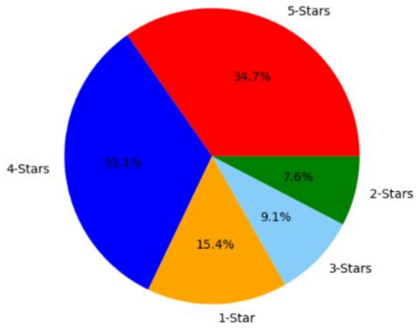


Fig 2. Shows distribution of classes in yelp dataset.

A. Tools

We used python as development environment, including appropriate modules for text cleaning and pre-processing as follows: nltk (natural language toolkit), sklearn, pandas, re, json, matplotlib and NumPy.

B. Data- preprocessing

Raw data is processed to make it easy for the algorithm to understand it and extract

the features related to each class by going through every instance.

For Naïve Bayes algorithm, raw text data is first filtered by removing stopping words and removing conjunctions from each review stop words, as they contribute nothing when it comes to polarity of a text. Using bag-of-words approach, each review is converted into vector format. Data is now split into training and testing (80% and 20% respectively).

For SVM algorithm, for each review text, case normalization is performed, punctuation and stop words are removed to make a simple string of words. This filtered data is now split into training and testing.

C. SVM

Support vector machines is one of the supervised learning algorithms mostly used for classification purposes. This algorithm separates different classes using a hyperplane having maximum possible distance between the two vectors called support vectors as shown in fig 2.

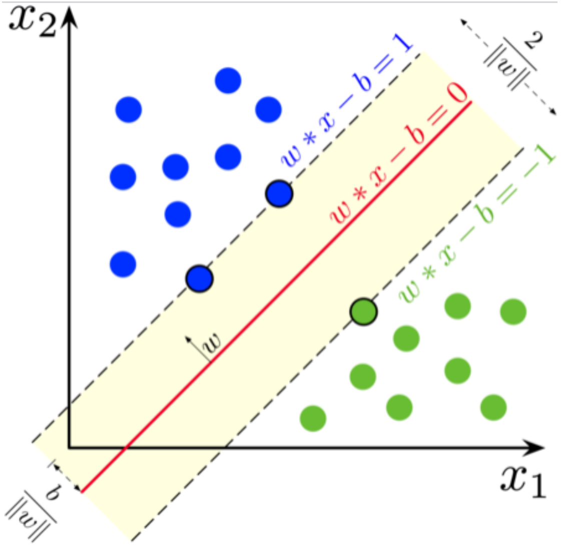


Fig 3. Hyperplane 'w' separating 2 classes.

SVM is used in the project because it is likely to perform more accurately in classification than other algorithms, as previous studies have proven the same. Also, SVM is convenient to use when working with a defined dataset. Finally, using fit function on training data to train the model and test on testing data that has been processed. This labeled data is useful to measure the performance of algorithm using accuracy and confusion matrix. SVM is also used for binary classification in this project. Using

SVM to find the polarity of a text (positive or negative) and accuracy is calculated.

D. Naïve Bayes

Naïve Bayes is a probability-based classifier based on Bayes theorem; this is a supervised learning algorithm [8]. This model can create new pieces of data form underlying probability distribution [7]. This model functions by storing probability of classification gained from training data.

For a new data (test data), classification probability for each class is calculated and the class having highest probability is the data classified as. To do so, each class has a specific word set (in case of text data analysis), which is collected during the training phase using tokenization technique.

In this experiment, bag-of-words is used to convert text reviews into feature vector.

These algorithms are also analyzed, by considering only 2 classes. i.e. all the reviews containing star rating 1, 2 or 3 are given rating 1 and those above 4 are given 5. This makes it a binary classification problem. This technique is an attempt to improve the accuracy, since, the reviews of star rating 2 and 3 are similar. Hence, 1 and 5 makes a huge difference when classifying.

MEASUREMENT METRICS

Accuracy: ratio of number of correct predictions to all the predictions made.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Precision: number of correct documents returned by our ML model

Precision=TP/TP+FN

Recall: number of negatives returned by the model

Recall=TN/TN+FP

Support: it is the number of samples of the true response belonging to each class.

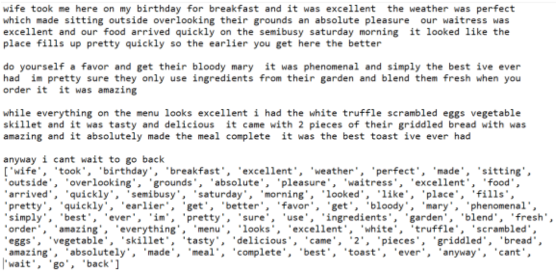
F1 score: harmonic mean of precision and recall.

F1=2*(precision*recall) / (precision + recall)

Confusion matrix to compare actual vs predicted outcome.

RESULTS

A snippet of outcome by using bag-of-words technique is shown below, stop words and punctuations are removed to extract meaningful words.

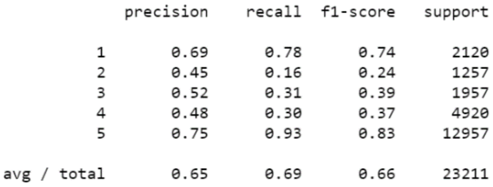


```
wife took me here on my birthday for breakfast and it was excellent  the weather was perfect
which made sitting outside overlooking their grounds an absolute pleasure  our waitress was
excellent and our food arrived quickly on the semibusy saturday morning  it looked like the
place fills up pretty quickly so the earlier you get here the better

do yourself a favor and get their bloody mary  it was phenomenal and simply the best ive ever
had  im pretty sure they only use ingredients from their garden and blend them fresh when you
order it  it was amazing

while everything on the menu looks excellent i had the white truffle scrambled eggs vegetable
skillet and it was tasty and delicious  it came with 2 pieces of their griddled bread with was
amazing and it absolutely made the meal complete  it was the best toast ive ever had

anyway i cant wait to go back
['wife', 'took', 'birthday', 'breakfast', 'excellent', 'weather', 'perfect', 'made', 'sitting',
'outside', 'overlooking', 'grounds', 'absolute', 'pleasure', 'waitress', 'excellent', 'food',
'arrived', 'quickly', 'semibusy', 'saturday', 'morning', 'looked', 'like', 'place', 'fills',
'pretty', 'quickly', 'earlier', 'get', 'better', 'favor', 'get', 'bloody', 'mary', 'phenomenal',
'simply', 'best', 'ever', 'im', 'pretty', 'sure', 'use', 'ingredients', 'garden', 'blend', 'fresh',
'order', 'amazing', 'everything', 'menu', 'looks', 'excellent', 'white', 'truffle', 'scrambled',
'eggs', 'vegetable', 'skillet', 'tasty', 'delicious', 'came', '2', 'pieces', 'griddled', 'bread',
'amazing', 'absolutely', 'made', 'meal', 'complete', 'best', 'toast', 'ever', 'anyway', 'cant',
'wait', 'go', 'back']
```

Fig 4. Snippet of review text and words extracted from it.

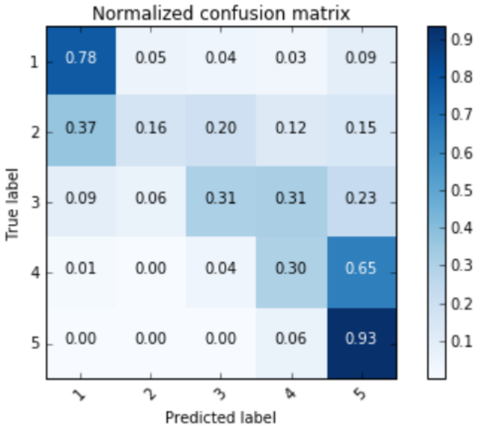
The SVM algorithm is working well enough to classify the texts into defined classes, and it is found to have an accuracy of 69.1%.

In this project, the classes do not differ much when you compare 1-star rating comment to a 2-star rating comment, considering this, SVM has done reasonable job in classification.



```
             precision    recall  f1-score   support

          1       0.69      0.78      0.74      2120
          2       0.45      0.16      0.24      1257
          3       0.52      0.31      0.39      1957
          4       0.48      0.30      0.37      4920
          5       0.75      0.93      0.83     12957

avg / total       0.65      0.69      0.66     23211
```

Fig 5. Classification report for SVM.



	1	2	3	4	5
1	0.78	0.05	0.04	0.03	0.09
2	0.37	0.16	0.20	0.12	0.15
3	0.09	0.06	0.31	0.31	0.23
4	0.01	0.00	0.04	0.30	0.65
5	0.00	0.00	0.00	0.06	0.93

Fig 6. Confusion matrix for SVM

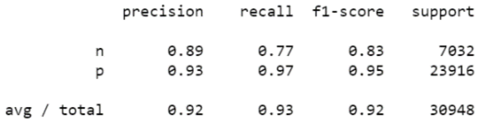
Above table clearly displays precision, recall, f-measure, support and accuracy for the text classification. This report can be useful to understand how accurate the algorithm works on this data [6].

Low precision means that the ratio of correct prediction made to actual total true outcomes is less which tells us the algorithm is less likely to guess correctly. As mentioned previously, this classification metrics is low in percentage because of similarities in target classes.

Recall-number of negatives returned for class 1,4 and 5 is more than that of 2 and 3 which means, that the model is more accurate when stars rating is 2 or 3.

Support for classes 4 and 5 is more than that of 1,2 and 3 which means, correct true sample of 4 and 5 are more in number.

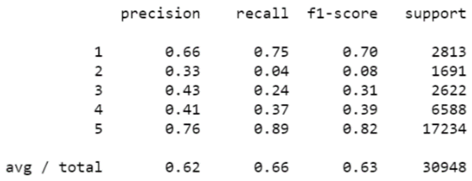
The algorithm is further evaluated for binary classification, and accuracy is found to be 92.6%.



```
             precision    recall  f1-score   support

          n       0.89      0.77      0.83      7032
          p       0.93      0.97      0.95     23916

avg / total       0.92      0.93      0.92     30948
```

Fig 7. Classification report for binary class using SVM.

Naïve Bayes algorithm is evaluated to have accuracy of 66% for multi-class classification (5-classes) and 88.3% for binary classification. Classification report and confusion matrix are shown below.



```
             precision    recall  f1-score   support

          1       0.66      0.75      0.70      2813
          2       0.33      0.04      0.08      1691
          3       0.43      0.24      0.31      2622
          4       0.41      0.37      0.39      6588
          5       0.76      0.89      0.82     17234

avg / total       0.62      0.66      0.63     30948
```

Fig 8. Represents classification report for Naïve Bayes

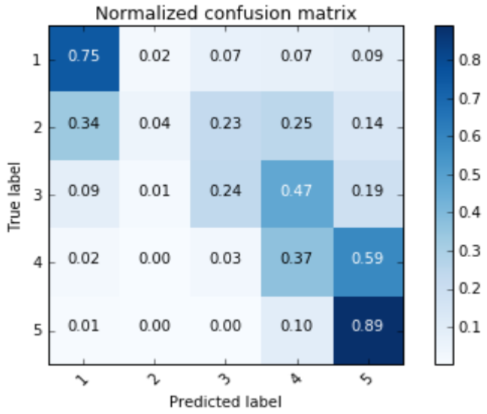
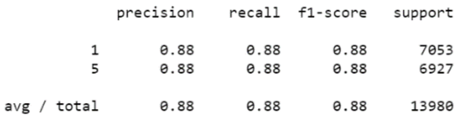


Fig 9. Confusion matrix for Naïve Bayes



```
             precision    recall  f1-score   support

          1       0.88      0.88      0.88      7053
          5       0.88      0.88      0.88      6927

avg / total       0.88      0.88      0.88     13980
```

Fig 9. Classification report for binary class using Naïve Bayes

LIMITATION OF THE PROJECT

As stated, the accuracy of the algorithm can be improved by using some more sophisticated data processing techniques. Also, the features collected for each class are not entirely correct for English language, thus the pre-processing of data should be improved in such a way that exact words are extracted for classification. This will also affect the accuracy, precision, F1 and support for the algorithm.

When running Naïve Bayes algorithm, training time is observed to be higher than usual i.e. 68 minutes (on i5 7th generation intel processor).

For SVM algorithm, vectorization process may take up some execution time (5 mins).

CONCLUSION

SVM is found to be performing better than Naïve Bayes algorithm, for both multi-class as well as binary class classification.

	Multi-class	binary
SVM	69.1%	92.6%
Naïve Bayes	66%	88.3%

Above table shows the accuracy of both algorithms in multi-class and binary classification. This can be justified as, SVM performs classification by constructing hyperplane in a multidimensional space [7], hence it maps the data in a higher dimension, which in turn gives better performance. Whereas, in Naïve Bayes, classifier assumes that value of a feature is not affected by presence of another feature [8].

REFRENCES

1. Huang, J.; Rogers, S.; Joo, E. (2014). Improving Restaurants by Extracting Subtopics from Yelp Reviews. In iConference 2014 (Social Media Expo)

2. MS Neethu and R Rajasree. Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pages 1–5. IEEE, 2013.

3. Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. Classification of sentimental reviews using machine learning techniques. Procedia Computer Science, 57:821–829, 2015.

4. Nimit Bindal and Niladri Chatterjee. A two-step method for sentiment analysis of tweets. In 2016 International Conference on Information Technology (ICIT), pages 218–224. IEEE, 2016.

5. https://www.yelp.com/academic dataset.

6. Flach, P., 2012. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press

7. Karan Tyagi and Shraddha Shah. Restaurant recommendation with yelp rating reviews in 2018, Northeastern University.

8. Wikipedia contributors. "Naive Bayes classifier." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 29 Aug. 2020. Web. 21 Aug. 2020.