ASPECT BASED OPINION MINING IN YELP REVIEWS USING SUPERVISED LEARNING TECHNIQUE

Mini Project(Semester-II)

Fathima Nishad
Registration No: 31019005
Nasila
Registration No: 31019013
Vishnupriya A.V
Registration No: 31019029

M.Sc. Computer Science (Specialisation in Machine Intelligence) 2019-21
Indian Institute of Information Technology and Management - Kerala,
Thiruvananthapuram
April 2020

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT -KERALA

A MINI PROJECT REPORT
ON
ASPECT BASED OPINION MINING IN YELP REVIEWS

Submitted by,

Fathima.Nishad
Registration No: 31019005
Nasila
Registration No: 31019013
Vishnupriya.A.V
Registration No: 31019029

Under the guidance of
Dr. Asharaf S

DECLARATION

We, the students of course Master of Science in Computer Science specialising in Machine Intelligence, hereby declare that this report is substantially the result of our own work, except, where explicitly indicated in the text and has been carried out during the period.

Place: Trivandrum
Date: Student's signature

ACKNOWLEDGEMENT

We are extremely grateful to Dr. Saji Gopinath, Director of IIITM-K for providing us with good facilities and a proper environment which helped us in enhancing our ability for undertaking a project of this scale. We thank Dr. Asharaf.S Professor, IIITM-K, Thiruvananthapuram, our guide for his valuable suggestion, appraisal, guidance and his valuable support towards the successful completion of our project, for last, but not the least, we'd also like to take this opportunity to thank all the teaching and non-teaching staff here at this institution for their cooperation and support.

ABSTRACT

The project aims to use machine learning models to analyze consumer opinions, sentiments, evaluations, attitudes, and emotions from written language using machine learning techniques. Aspect-based opinion mining aims to extract major aspects of an review and with basis of rating we are predicting the polarities of each reviews. Ratings are the intended interpretation of user satisfaction in terms of numerical values. With the basis of this numerical value we are detecting the polarity either positive,negative or neutral.

In this project we will take various Machine Learning Classifiers ranging from simple classifiers like Naive Bias, SVM and KNN then we are comparing their performances. The algorithm giving the highest prediction accuracy will be considered for the development of the prediction tool.

CONTENT

INTRODUCTION

The emergence of user-generated content via social media has had an undeniable impact on the commercial environments. In fact, social media has shifted the content publishing from businesses towards the customers . While there are several sources of user-generated content (e.g., discussion forums, Tweets, Blogs, News and reports, consumer feedback from emails, call centers etc.), none of them is as focused as online reviews. That is why consumers mainly read and consult online reviews before purchasing products or using services. However, the growing volume of online opinions makes it harder and harder to make informed decisions. On the other hand, online reviews provide businesses with a rich source of consumer opinions for free. Marketing studies show that online reviews influence consumer shopping behaviour significantly . Many businesses are now tracking customer feedbacks through online sources. Amazon, Yelp and TripAdvisor are examples of these Web resources containing consumer opinions. Aspect-based opinion mining aims to extract major aspects of an review and with basis of rating we are predicting the polarities of each reviews. Aspects are attributes or components of items (e.g., ‘LCD’, ‘battery life’, etc. for a digital camera) and ratings are the intended interpretation of user satisfaction in terms of numerical values.

Names of Selected Classifiers

1. SVM

2. Naive Bias

3. KNN

PROBLEM STATEMENT

In this, we have to predict positive ,negative and neutral reviews classifying sentiments based on polarity.Mining aspects from text.The machine learning models we used to make predictions using different classification models.

METHODOLOGY

All operations are performed in Python and libraries used are numpy, for mathematical operations, pandas for handling data, and sklearn for applying the machine learning models.

Data

yelp reviews:

SLNO	ATTRIBUTES
1	review_id
2	user_id
3	business_id
4	stars
5	date
6	text
7	useful
8	funny
9	cool

Pre-processing:

After reducing the attributes of the data set, the next step was to assign labels to all the data points, which was done by the means of a predetermined threshold. The categorical data is converted to numerical data by integer encoding. The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship. Once the labels obtained, they had to be converted into numerical values for the purpose of training, where 1 denoted predicted disease and 0 denoted not predicted disease. Finally the data and labels had to be split into two parts, training set and testing set, for the purpose of training and testing the models. It should be noted that labelling, converting labels to numerical value, and dropping a column was done using methods from pandas, while methods for calculating feature importance and train-test splitting are available in scikit learn.

Prediction:

The machine learning models we used to make predictions are support vector machines (SVM), Naive Bayes classification, and KNN. All of these packages are from the scikit learn package. First a model was initialized, and then it was trained using the training data and training labels. Then a prediction was made, using the testing set on the model, and the predicted values are stored separately. Finally, in order to evaluate the performance of the model, we must compare the predicted results with the test labels and see how efficiently the model worked. So we compute the accuracy score of the model using the predicted outputs and the test labels. We also obtain the confusion matrix of the model with respect to the test labels. A confusion matrix is a technique for summarizing the performance of a classification model. Accuracy alone can be misleading during classification problems so a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making. Methods for computing accuracy and confusion matrix are all available in the sklearn.metrics library of the scikit learn package. The purpose behind using different machine learning models is to do a comparative study of their performances for the given dataset and infer which model is most suitable for such types of problems.5

ALGORITHMS USED

Naive bayes classifier:

This classifier is a powerful probabilistic representation, and its use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real world datasets because it can give better predictive accuracy.

Support Vector Machine:

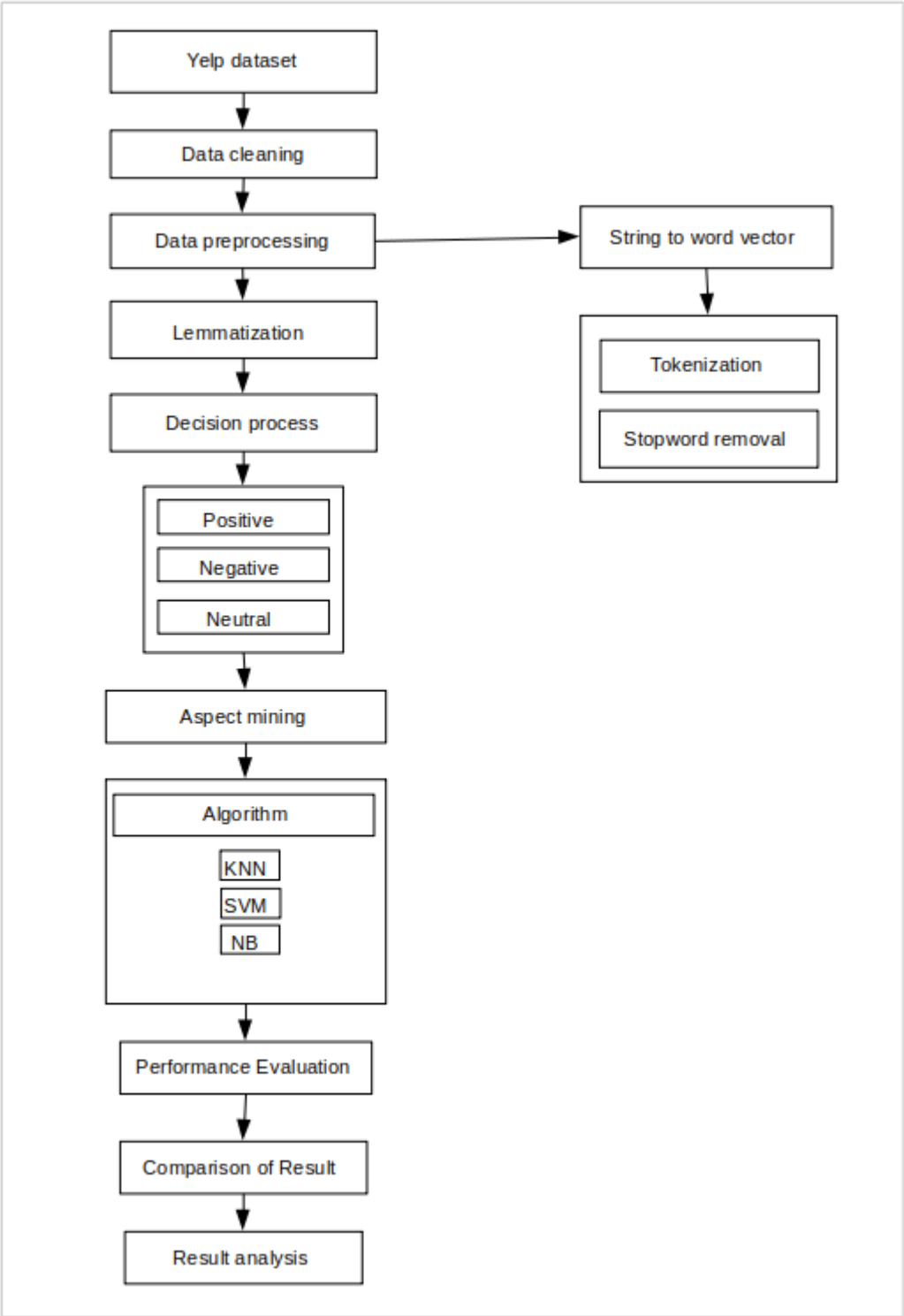
Support vector machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which creates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting on the training data. Overfitting is caused by incorporating the random errors or noise in the model. Therefore the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keeps the model from over-fitting. The complexity of the model has to be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. A SVM can make some errors to avoid over-fitting. It tries to minimize the

number of errors that will be made. Support vector machines classifiers are applied in many applications.

K-Nearest Neighbour:

This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according6 to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance.

WORK FLOW DIAGRAM



RESULTS

Algorithm	*Accuracy*
Multinominal Naive Bayes	65.26
Support Vector Machine	59.24
K- Nearest Neighbor Classifier	59.64

Table 1

CONCLUSION

In this project, we carried out an experiment to find the predictive performance of different classifiers. We select three popular classifiers considering their qualitative performance for the experiment. We choose yelp dataset. Here, the Multinominal Naive Bayes has good score compared to other models By using this model we can predict the positive, negative and average reviews. In order to compare the classification performance of for supervised learning algorithms, classifiers are applied on same data and results are compared on the basis of there accuracy and according to experimental results in table 1, it can be concluded that Naive bayes is the best as compared to SVM and K-Nearest Neighbour. After analysing the quantitative data generated from the computer simulations, Moreover their performance is closely competitive showing slight difference. So, more experiments on several other datasets need to be considered to draw a more general conclusion on the comparative performance of the classifiers.

REFERENCE

https://medium.com/@pmin91/aspect-based-opinion-mining-nlp-with-python-a53eb4752800

https://www.kaggle.com/omkarsabnis/sentiment-analysis-on-the-yelp-reviews-dataset

https://www.kaggle.com/manovirat/aspect-based-sentiment-analysis

https://devopedia.org/aspect-based-opinion-mining