Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019)
IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8

Sentiment Analysis of Yelp Reviews by Machine Learning

Hemalatha S*[1], Ramathmika[2]
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education, Manipal, Karnataka, India-576104.
*[1]hema.shama@manipal.edu (Corresponding author), [2]ramathmikavs1999@gmail.com

***Abstract*— Sentiment analysis is a process of analyzing a piece of text written by a writer to identify and classify the opinions buried in that text and to determine whether the views of the writer about the topic is positive, negative, or neutral. Yelp is a review forum which provides reviews on local businesses. Users from anywhere in the world can post reviews and rate any business in this social networking site. In this paper, the textual yelp reviews of businesses are analyzed to assign a probability for the review as having positive or negative sentiment. The data considered for the sentiment analysis are the reviews on restaurants about food, service, price and ambience. Machine learning algorithms in the nltk library of python can prove to be very useful in any such research on Natural Language Processing and the library has been used extensively in this work. Each algorithm used has been analyzed and has been compared on the basis of their efficiency (confidence).**

Keywords*— *Sentiment; machine learning; yelp reviews; analysis; restaurant reviews

I. INTRODUCTION

The three fields of the computer science discipline, namely Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) are useful in developing many applications. The relationship between these fields is as shown in the Fig. 1.

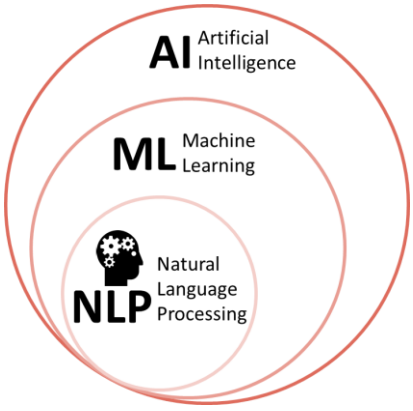

Fig. 1. Relationship between AI, ML, and NLP

AI teaches system to do intelligent things. **ML** teaches the system to do intelligent things that can learn from experience. **NLP** teaches the system to do intelligent things that can learn from experience and understand human language. NLP is with in the field of AI and goes hand in hand with ML. It is concerned with the preprocessing of data entered by a human in a worldly language which is not directly consumable by the computer. In Machine learning, without explicit programming, the computers are provided with the ability to learn automatically. The computer programs are developed such that the system improves by experience without human intervention [1]. The work presented in this paper involves the processing of the textual reviews written in English and deciding whether the review has a positive or negative sentiment. This is a simple application of Machine Learning, but similar projects on a large scale have numerous applications in the automation of computer responses.

Psychology is the scientific study of the human mind. What a particular human mind had in mind when he/she acted in a certain way (writing review in this case) was only a job for other well-read humans to decipher, but today the point of 'future' has been reached, where a computer can be expected to do a better job than human beings. With further advancements in the field of psychoanalysis of the human mind, many simple jobs can be automated, and humans can concentrate on the more important things. For instance, a call centre has a number of people taking calls of people reviewing a certain product of a company. With the development of learning algorithms, the task of analyzing these reviews can be given solely to the computer which will save many man-hours. Numerous such examples suggest the importance of such a project and the applications like this make this world a better place to live in.

Through machine learning, the computers will learn to do intelligent jobs such as predicting for the future or decision making. Machine learning algorithms are mainly categorized into four types: supervised learning, unsupervised learning, semi supervised learning and reinforcement learning [1].

A. Supervised learning

Labeled training data consisting of a set of example input and target output pairs is provided to the computer. The aim is to make the computer to learn a general rule that relates inputs to outputs. It is the task of a computer to learn a function that

maps an input to an output based on example input-output pairs. The algorithms in supervised learning analyzes the training data to generalize a function, that can be used for mapping new input examples.

B. Semi-supervised learning

In this case, a training data set with some of the target outputs missing is given to the computer. The computer is made to learn by a little of labelled data with plenty of unlabeled data. Here, the use of entire labelled data and intelligent use of unlabeled data improves the model performance

C. Reinforcement learning

In reinforcement learning, the computer does decision-making and takes actions in a dynamic environment and the reward (or penalty) is given as the feedback for its actions. After several trials, the best policy will be learnt.

D. Unsupervised learning

Here unlabeled data is provided to the learning algorithm that infers a function describing the structure of unlabeled data. Unlabeled data is the one that is not classified or categorized.

To perform sentiment analysis of restaurant reviews obtained from Yelp dataset, the reviews are classified as having positive sentiment or negative sentiment using the classification algorithms such as Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Linear SVC (Support Vector Clustering).

E. Naive Bayes Algorithm

It is a classification technique which depends on Bayes theorem. It assumes that the presence or absence of a particular feature of a class is independent of other features in the class. It is the simplest algorithm to classify large data set.

F. Multinomial Naive Bayes Algorithm

Here each of the feature has multinomial distribution forming a feature vector that represent the frequency of occurrence of that feature in a particular instance. It is a specific instance of Naive Bayes classifier.

G. Bernoulli Naive Bayes Algorithm

Here the features are independent binary variables representing the presence or absence rather than probabilities as in Multinomial Naive Bayes.

H. Logistic Regression Algorithm

It is a mathematical model that estimates the probability of occurrence of a feature and it works with binary data. For example, if food quality is a feature, then probability of being good quality has the value 1 and that of bad quality has the value zero.

I. Linear SVC (Support Vector Clustering) Algorithm

It returns a hyperplane that categorizes the provided data set into good and bad.

II. Literature Survey

The research on NLP started long back in the 1960s with development of some NLP systems such as SHRDLU and ELIZA. SHRDLU works with restricted vocabularies and ELIZA is simulation of a psychotherapist which provides human-like interaction. With the invention of "conceptual ontologies" in the 1970s, many NLP systems were built which were based on complex sets of hand-written rules. Later in the 1980s, when computational power increased, ML algorithms were introduced for NLP. Some of the systems developed during this time used machine learning algorithms such as decision trees and produced systems of hard if-then rules similar to existing hand-written rules. The hidden Markov models used part-of-speech tagging in NLP and some statistical models which make probabilistic decisions were developed. Real valued weights were attached to the features making up the input data [2]. The cache language models upon which many speech recognition systems now rely are examples of such statistical models.

Sentiment Analysis is a major tool for a machine to understand human psychology. This technology is being studied extensively in order to implement in the fields where humans were needed to detect sentiment or feeling. It is of major importance in assistant chatbots and combined with speech recognition technology, it can also be used to substitute humans in call centres.

In the recent years some researches have been found in the literature. One among them is the work of Jiangtao Qiu et al., [3] which predicts the ratings of Yelp reviews that are nonrated, from their sentiments. The sentiments are analyzed at the aspects level and the ratings depend on the sentiment of aspects and the number of positive and negative aspects of the review. Another proposal for sentiment analysis from aspects, given by Ayoub Bagheri et al., [4] detects the aspects by employing certain heuristic rules to find the effect of opinion words in detecting the aspects. A new metric is proposed to score aspects based on aspect frequency. Alvaro Ortigosa et al., [5] proposed a sentiment analysis method for facebook messages using lexical based and machine learning approach and explored its application for e-learning. Sentiment analysis at the sentence level is proposed by Orestes Appel et al., [6] which uses a hybrid method consisting of NLP techniques, lexicon and fuzzy sets to estimate the semantic polarity. Aminu Muhammad et al., [7] proposed a sentiment analysis model for social media genres by integrating local context and global context. The relation of terms with their neighbours is local context and text genre is the global context. Sequence modelling proposed by Tao Chen and Ruifeng Xu et al., [8] for the distributed user and product representation learning improves the document-level sentiment classification performance. Marco Rossetti et al., [9] extended the topical method with an application focus on tourism domain. The user and item models are derived and analyzed for predictions and recommendations in the tourism domain. Anuja P Jain and Padma Dandannavar [10] applied machine learning algorithms for sentiment analysis of Twitter data. Xueying Zhang and Xianghan Zheng [11] compared the Chinese sentiments by different machine learning algorithms. Yan Zhu et.al, [12] proposed a multilayer classification which gives better results

than traditional methods. Guoshuai Zhao et. al, [13] used social users' rating behaviour about the user-services, in the social media to predict their ratings. Tri Doan and Jugal Kalita [14] performed sentiment analysis using a variant of online random forests with incremental learning. These are some of the important researches found in the literature about sentiment analysis. The methodology proposed in this paper is explained in the next section.

III. Proposed Method

This paper enlightens the implementation of Machine Learning algorithms which processes textual, statistical data provided by the Yelp dataset as a part of the Yelp Dataset Challenge. It is an application of supervised learning. The aim is to classify the business reviews into graded categories such that they can be roughly sorted in an order from bad to good or negative to positive.

There is a sequence of steps involved from taking raw data to predicting whether a review is of positive or negative sentiment and to what degree. Tokenizing, part of speech tagging, assigning probabilities of good or bad on the basis of occurrence to every adjective that occurs within the reviews are just some of the major steps to be followed [2]. The focus of this work is going to remain on a dataset provided by Yelp, which has reviews for businesses. The nltk, a natural language processing library in python is used to fetch the algorithms.

The following algorithms have been used:

1. Naive Bayes
2. Multinomial Naive Bayes
3. Bernoulli Naive Bayes
4. Logistic Regression
5. Linear SVC (Support Vector Clustering)

The classification in the proposed model is achieved by taking the mode (out of 5 results if three or more give positive sentiment then the output is positive) of the outputs of those algorithms. Natural language processing has become a step by step procedure of the 60+ years of research that has been put into it. Following are the steps of implementation:

1. Data processing is one of the most important aspects of machine learning. In order to ensure the quality of training data, over 16,000 reviews are classified to positive and negative sentiments and the machine is trained to higher levels of precision. This is done by sorting the data and converting it to a txt format file which is easy for Python program to read.
2. The second step to train a NLP machine is tokenization of data. Each review is word tokenized and passed on to the next step of classification.
3. One of the most important steps in the classification of these reviews is the Part of speech tagging where each token is tagged to its part of speech and only the adjectives are considered for the classification purpose. Since adjectives are strongly descriptive words, this step is of major importance to the training.
4. The occurrence of each of these adjectives is then mapped into a tuple (word, frequency) in order to assign the probability of each word occurring or not occurring in a review depending on its type.
5. The first 50000 most occurring adjectives are chosen and assigned probabilities of them making a review of positive or negative sentiment. This step will determine the confidence of the output of each review that is tested later. It is one of the most time consuming steps of the procedure.
6. Then the first 5000 feature sets are taken as the training set and the last 5000 as the testing set.
7. Applying the classification algorithms on these training and testing sets give their respective accuracies.
8. Every time a review is entered by the user, it is analyzed by the trained machine. The mode of the results of these five algorithms is taken and define confidence for the sentiment of that review.

IV. Exparimentation and Results

The experimentation is carried out in Python, in a machine with intel core i5 processor and 4GB RAM. The accuracy of each algorithm depends on the quality of the training and testing sets. The accuracy obtained for each of the algorithms and the accuracy of the proposed model is shown in Table 1.

Table 1. Accuracies in Different Algorithms

Algorithm	Accuracy
Naive Bayes	79.12
Multinomial Naive Bayes	78.92
Bernoulli Naive Bayes	73.22
Logistic Regression	78.88
Linear SVC (Support Vector Clustering)	75.32
Proposed	78.44

These numbers are good for a practice project but aren't acceptable for implementation in real world machines. This accuracy can be improved by providing manually sorted data as an input training set and providing more number of reviews for training. The capacity of the computer is also a major issue to be kept in mind. The maximum desired accuracy is 90%. Further, more advanced algorithms, better methods of tokenizing and better classification can give a smaller gap between the desired efficiency and actual results.

The primary objective of this work is to be able to teach the computer to read and understand a sentence typed by a human in a real world language (English) and to be able to classify this into positive or negative sentiment. The accuracy of this classification is judged by its confidence which is the number of algorithms that give correct prediction over a total number of algorithms used. To efficiently classify the reviews given to a business and judging the business by the basis of the

confidence is what this work focuses on. Further research into this technology and ever increasing knowledge of such machines owing to their massive processing power sees some major advancements in the field of sentiment analysis and it will facilitate various industries that have to directly interact with humans.

Fig. 2 is a screenshot of a fraction of the list of words that the machine checks within every new review it encounters. This is the preprocessed data. Only the relevant words are taken here.

pos : neg or neg : pos is a probability ratio of that particular word occurring is a positive or a negative sentiment respectively.

Satisfactory results have been obtained for simple English sentences as mode of the outputs of the algorithms are used for deciding sentiment. Thus the poor accuracy in each of the algorithms is compensated. Long and descriptive sentences are encouraged in order to get better and more accurate results. The machine may not be competent enough to detect sarcasm as the training set had genuine reviews. Some sample statements and their confidence measures are shown in Fig. 3 and Fig. 4 respectively.

```
Most Informative Features
                   worst = True              neg : pos    =     18.9 : 1.0
               overcooked = True              neg : pos    =     15.6 : 1.0
                 mediocre = True              neg : pos    =     15.3 : 1.0
              knowledgable = True              pos : neg    =     14.4 : 1.0
                disgusting = True              neg : pos    =     14.3 : 1.0
                    ripped = True              neg : pos    =     14.3 : 1.0
                    poorly = True              neg : pos    =     13.6 : 1.0
             disappointing = True              neg : pos    =     13.5 : 1.0
                    stated = True              neg : pos    =     12.5 : 1.0
                     awful = True              neg : pos    =     11.9 : 1.0
                      poor = True              neg : pos    =     11.3 : 1.0
                  horrible = True              neg : pos    =     11.0 : 1.0
                  terrible = True              neg : pos    =     10.7 : 1.0
                    savory = True              pos : neg    =     10.4 : 1.0
                     drunk = True              neg : pos    =      9.8 : 1.0
                  genuinely = True              pos : neg    =      9.7 : 1.0
                    filthy = True              neg : pos    =      9.6 : 1.0
                      rude = True              neg : pos    =      9.5 : 1.0
                   happier = True              pos : neg    =      9.0 : 1.0
                    quaint = True              pos : neg    =      9.0 : 1.0
                      scam = True              neg : pos    =      9.0 : 1.0
                  talented = True              pos : neg    =      8.6 : 1.0
                    photos = True              pos : neg    =      8.5 : 1.0
                   fitness = True              pos : neg    =      8.4 : 1.0
                    delish = True              pos : neg    =      8.4 : 1.0
                      lake = True              pos : neg    =      8.4 : 1.0
```


Fig 2. List of words and their probabilities of being a ‘pos’ (positive) or ‘neg’ (negative) sentiment.

```
("bad food!,italian dishes were not authentic"))
("good food!"))
("bad food!,asian dishes were not authentic"))
("it was bad"))
("cheap food and great quality"))
("The food was great! the ambiance was lively and the waiters were very friendly. well lit and great fragrance."))
("Very bad taste of the food. I don't like the smell of that place. terrible service and dark and cold atmosphere. very dull"))
("amazing broccoli better"))
("amazing broccoli"))
```

Fig 3. Sample input statements

```
('pos', 0.6)
('pos', 0.8)
('neg', 0.6)
('neg', 0.6)
('pos', 1.0)
('pos', 0.8)
('neg', 1.0)
('pos', 1.0)
('pos', 1.0)
```

Fig. 4. Output after sentiment analysis is done on the data in Fig. 3

V. Limitations and Future Scope

Some limitations of the project include accuracy and less diversity of categorization. The training set can never be perfect as it is not an easy task to separate 20,000 reviews into good or bad just by reading them. The quality of the training data set and testing data set is a major aspect in determining the efficiency of any such algorithms. Moreover, the classification is binary and gives an overall result of the goodness or badness of the review. It does not detect sarcasm and does not deal with individual traits such as food, ambiance, cost, service among others. The future scope is to rate business based on different features, to use better and more dataset to train and to detect sarcasm

VI. Conclusion

Analysis of the results shows that we have successfully been able to get a satisfactory level of accuracy in the classification of Yelp reviews using supervised learning. By attaining higher accuracy for such experiments we can see a day in the near future where it will be normal for day to day machines to understand our sentiments by words straight out of our mouth. Natural language processing is just one small application of the mammoth that is machine learning. Machines have already started to replace humans in most fields of life. Low level employment in many companies can be substituted by machines and it can save them much money. Machine learning is one of the most important and interesting topics to study. Most things can be narrowed down to just a bunch of numbers and probabilistic occurrences. The power that mathematics has provided us, exponentiated by the computation ability of a computer is supreme and we can only benefit from it if we learn to tame it.

References

[1] A. Ethem, Introduction to Machine Learning, MIT Press. , 2010..

[2] D. Jurafsky and H. M. James, Speech and Language Processing, 2000.

[3] Q. Jiangtao, L. Chuanhui, L. Yinghong and L. Zhangxi, "Leveraging sentiment analysis at the aspects level to predict ratings of reviews," Information Sciences, vol. 451–452, p. 295–309, 2018.

[4] A. Bagheri, M. Saraee and F. de Jong, "An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews," Lecture Notes in Computer Science., vol. 7934, pp. 140-151, 2013.

[5] O. Alvaro, M. M. José and M. C. Rosa, "Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, vol. 31, pp. 527-541, 2014.

[6] A. Orestes, C. Francisco, C. Jenny and Hamido, "A hybrid approach to the sentiment analysis problem at the sentence level," Knowledge-Based Systems, vol. 108, pp. 110-124, 2016.

[7] M. Aminu, W. Nirmalie and L. Robert, "Contextual sentiment analysis for social media genres," Knowledge-Based Systems, vol. 108, pp. 92-101, 2016.

[8] C. Tao, X. Ruifeng, H. Yulan, X. Yunqing and W. Xuan, "Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis," IEEE Computational intelligence magazine, pp. 34-44, 2016.

[9] R. Marco, S. Fabio and Z. Markus, "Analyzing user reviews in tourism with topic models," Information Technology Tourism , vol. 16, p. 5–21, 2016.

[10] P. J. Anuja and D. Padma, "Application of Machine Learning Techniques to Sentiment Analysis," in IEEE conference, 2016.

[11] Z. Xueying and Z. Xianghan, "Comparison of Text Sentiment Analysis based on Machine Learning," in 15th International Symposium on Parallel and Distributed Computing, 2016.

[12] Z. Yan, M. Melody and M. Teng-Sheng, "Multi-Layer Text Classification with Voting for Consumer Reviews," in IEEE International Conference on Big Data (Big Data), 2016.

[13] Z. Guoshuai, Q. Xueming and X. Xing, "User-Service Rating Prediction by Exploring Social Users' Rating Behaviors," IEEE TRANSACTIONS ON MULTIMEDIA, vol. 18, no. 3, pp. 496-506, 2016.

[14] Tri Doan and Jugal Kalita, "Sentiment Analysis of Restaurant Reviews on Yelp with Incremental Learning", in IEEE International Conference on Machine Learning and Applications, 2016, pp 697-700