Business Reviews Classification Using Sentiment Analysis

Conference Paper · September 2015

DOI: 10.1109/SYNASC.2015.46

1 author:

Andreea Salinca
University of Bucharest
9 PUBLICATIONS **32** CITATIONS

SEE PROFILE

Business reviews classification using sentiment analysis

Tool/Experimental Paper

Andreea Salinca
Faculty of Mathematics and Computer Science
University of Bucharest
Bucharest, Romania
andreea.salinca@fmi.unibuc.ro

***Abstract*— the research area of sentiment analysis, opinion mining, sentiment mining and sentiment extraction has gained popularity in the last years. Online reviews are becoming very important criteria in measuring the quality of a business. This paper presents a sentiment analysis approach to business reviews classification using a large reviews dataset provided by Yelp: Yelp Challenge dataset. In this work, we propose several approaches for automatic sentiment classification, using two feature extraction methods and four machine learning models. It is illustrated a comparative study on the effectiveness of the ensemble methods for reviews sentiment classification.**

Keywords*— *sentiment analysis; opinion mining; classification; text reviews

I. INTRODUCTION

Sentiment analysis has become an important research area for understanding people's opinion on a matter by analyzing a large amount of information. Millions of people express their thoughts about various services or products using social networking sites, blogs or popular reviews sites. The active feedback of the people is valuable not only for companies to analyze their customers' satisfaction and the monitoring of competitors, but is also very useful for consumers who want to research a product or a service prior to making a purchase.

Yelp users' reviews express opinions and sentiments about businesses and service providers among a given rating, scaled from 1 to 5, which is used as a general metric review. Yelp challenge dataset contains information about 61 000 local businesses, 1.6 million reviews and 366 000 users in 10 cities across 4 countries on the globe. Yelp challenge dataset is much larger than previous released Yelp academic dataset, which contains 15 585 business and 335 022 users' reviews [1].

This paper presents the results of several machine learning algorithms for classifying Yelp reviews using sentiment analysis and natural language processing techniques. The huge number of user-generated *reviews* and ratings for restaurants, *businesses* and service providers are classified as either positive or negative with respect to the star ratings. We propose a method to automatically classify users' sentiments (positive or negative) using only the business text review. This is very useful because it allows users feedback to be expressed without manual intervention. By studding only the rating, it is very hard to analyze why the user has rated the business as 1 or 5 stars. However, the text review contains a more quantitative value for analyzing more than the rating itself. This paper also presents the preprocessing steps needed in order to achieve high accuracy in the classification task.

There is no previous research on classifying sentiment of text reviews using the latest reviews from Yelp challenge dataset. Determining the underlying sentiment of business review is a complex task taking into account several factors such as the connotation of a word depending on the context, language used, sentiment ambiguity when using words that don't express a particular sentiment or when using sarcasm. We show that a sentiment analysis algorithm built on top of machine learning algorithms such as Naïve Bayes and Linear Support Vector Classification (SVC) has accuracy above 90% using 1.6 million business reviews. Furthermore, in this paper we present the results of our experiments and ideas on how to further improve the obtained results.

II. RELATED WORK

Previous work regarding sentiment analysis classification using machine learning techniques in determining if the overall sentiment of a review is positive or negative used movie reviews as data. The authors use a unigram model and Naïve Bayes, maximum entropy classification, and support vector machines to perform the sentiment classification and achieve 80% accuracy. They concluded that their results outperform the method based on human tagged features [2].

Hu et al. perform the sentiment classification of a document at a sentence level instead of the whole document and extract features on which opinions have been expressed, identifying opinion words by proposing a technique that uses WordNet lexical database. For each feature, the related opinion sentence is included into positive or negative categories and computes a total count. The features are ranked according to the frequency of the appearance in the reviews. The authors provide a feature-based summary of reviews of products sold online [3].

Previous work using sentiment analysis and Yelp dataset reviews focused on predicting star rating using the text alone. The authors experiment different machine learning algorithms such as Naïve Bayes, Perceptron, and Multiclass SVM on a sample of 100 000 user reviews from Yelp dataset. They use

Bing Liu Opinion Lexicon for feature selection and some preprocessing techniques such as removing stop words or stemming (i.e. reducing the words to their root form). The best result (precision and recall) is obtained using Naïve Bayes and feature selection with stop words removed and stemming [4].

In [5] in order to capture word sentiments the authors used a supervised learning algorithm based on similarities between words which takes into account the rating of previous reviews for capturing the representation of words vectors. On a dataset of about 20 000 unique reviews from Yelp dataset, the accuracy reported is about 70%. Other work on opinion mining, and in particular in review mining using Yelp dataset, focuses on predicting a business' rating based on its only reviews' text for reducing the bias of the users. The authors create a bag of words representation of the top frequent words in all text reviews or top frequent adjectives after Part-of Speech combined with Linear Regression, Support Vector Regression and Decision Tree Regression. Using 35 645 reviews from Yelp Academic dataset, the Root Mean Square Error (RMSE) is around 0.6 [6].

Blair-Goldensohn et al. built a system that can automatically summarize opinions from a set of reviews for a local service such a restaurant or hotel and aggregate the review sentiment per aspect (such as food, service, décor, value). They implement a custom built lexicon based on WordNet and use a classifier at the sentence level [7].

III. Sentiment Analysis Clasification

We use the Yelp Challenge Dataset [1] which consists of 1.6 million reviews (to around 7 GB of data) from 366 000 users from 61 000 local businesses which are stored in JSON format. The dataset contains information about business name, location, category, users, reviews, dates, stars. However, we have focused only on business raw text reviews and their star ratings for sentiment classification.

In order to use supervised learning and train a classifier, we usually require a hand labeled training data, but taking into account the large range of businesses and the large number of reviews, it would be very difficult to manually annotate the data to train a sentiment classifier for reviews.

We assume that the star rating is an accurate measure for the sentiment opinion of the review. The average rating of the Yelp reviews was about 3.7. The star rating of a business review is an integer from 1 to 5. We decided to eliminate all star ratings and their corresponding review which are equal to 3, and to keep all ratings above 4 which were considered as "positive" sentiments, and also to keep all ratings below 3 which were considered as "negative" sentiments. We obtain a dataset containing 1 346 545 reviews.

After splitting the filtered Yelp challenge dataset in 80% for training and 20% for testing, we use preprocessing techniques in order to extract a set of features. This data was used to train several classifiers. The classification experiments were conducted using 3-fold cross validation for evaluating the accuracy. Our approach is to use different machine learning classifiers and build feature extractors. The machine learning classifiers used are Naïve Bayes, Linear Support Vector Classification (SVC), Logistic Regression and Stochastic Gradient Descent (SGD) Classifier. For feature extraction we use different pre-processing techniques. Lastly, we measure the accuracy using the testing dataset.

A. Preprocessing

In the preprocessing phase all the punctuations and all the spaces from the review text are removed. We convert all capital letters to lowercase in order to reduce redundancy in feature selection task.

For feature extraction we implement two approaches, one approach is building a custom dictionary from the training dataset and the other approach is performing lexical analysis of text reviews. We also considered using an existing opinion lexicon, but the list of features to be extracted would have a high bias. Some features specific to Yelp dataset would not be taken into consideration and become irrelevant with respect to the sentiment mining. Even though the features set would contain only words scored for polarity (using a compiled list of negative and positive opinion words) some important features that are not found in the dataset would have been excluded.

B. Feature extraction

After preprocessing steps, for the first approach in features selection algorithm we loop over the entire training dataset to tokenize each word. We implement the following steps:

- Negations are grammatical constructions that change the meaning of a sentence to the opposite. An approach for handling negations is to append a different character such as "!" to every word between negation and the following punctuation [8]. We chose to modify only the word immediately before and after the negation term, rather than all the words until the end of the sentence. This step is done in order to express more information of the sentence context.
- Remove stop words (common words in English, but with no sentiment information) from the text reviews using Porter corpus of stop words from Natural Language Toolkit (NLTK).
- Remove words that have less than three letters.
- Apply stemming (reducing a word to its root form or base form) using Porter algorithm which is implemented in NLTK. This step is done for removing repetitive features and it is also useful because it reduces the size of the vocabulary by about 30 percent.
- Further, in the last step of the first approach for building the feature extraction algorithm, to form the feature vectors we used a bag of words representation for text reviews of all business using unigrams. We build a custom dictionary for mapping the frequency of occurrence of the words in the training set – each token occurrence is treated as a feature and the vector of all the token frequencies represents a multivariate sample. We remove the words with frequency lower than one for reducing the sparsity of the training matrix.

For the second approach of the feature selection algorithm we also choose to apply the preprocessing steps previously defined. Further, we remove stop words and words with less

than three letters and apply the stemming step as described above. Next we implement the following steps:

- We loop over the entire training dataset to tokenize each sentence and assign a POS (part-of-speech tag).
- Due to the fact that each word can have multiple senses, we apply Lesk's word sense disambiguation (WSD) algorithm built on top of WordNet which is implemented in NLTK. The algorithm returns a synonym set (Synset) that has the highest number of overlapping words between the context sentence and different definitions from each WordNet synonym set.
- After finding the closest meaning in the context of the word using WSD, we use a sentiment opinion lexicon SentiWordNet [9] for computing the positive and negatives scores of each word in each sentence of the text review. SentiWordNet assigns to each WordNet Synset three sentiment scores: positivity, negativity and objectivity (for each Synset the sum of the three scores is 1, and each score value is ranging from 0 to 1). For instance, for the word "slow", there are 3 senses of the verb "slow", 3 senses of the adjective "slow", and 2 senses of the adverb "slow" in WordNet. In SentiWordNet according to its POS for one sense of the adjective "slow", the negative score is 1, but for another sense the negative score is 0.
- For each tokenized sentence of the review we use all scores of the words in a sentence and compute the total positive and total negative score of the text review. We use total positive and negative scores as predictors.

C. Learning Methods

We adapt and apply four learning models: Multinomial Naïve Bayes, Support vector machines: Linear Support Vector Classification, Logistic regression and Stochastic Gradient Descent Classifier for sentiment analysis classification using Yelp challenge dataset. All algorithms are implemented in the scikit-learn machine learning library written in Python [10].

Naive Bayes is a traditional approach for text classification. In Multinomial Naïve Bayes algorithm, we represent a business text review using a feature vector extracted using the proposed feature extraction algorithms. The length of the feature vector is equal to the number of words in the extracted dictionary. We use a Boolean feature vector form representation, assuming that the occurrence of each word is more important than the frequency. We use a binary weighting function and assign 1 if the word is present in the dictionary and 0 otherwise. Laplace smoothing is used for avoiding over fitting.

We build the feature extraction vector using TF-IDF (term frequency-inverse document frequency) transform method to reflect how important a word is to a business text review and apply the following classifiers: Linear Support Vector Classification (SVC), Stochastic Gradient Descent Classifier and Logistic regression classifier. In our "bag of words" approach, each word is weighted using its frequency in the dictionary created with the proposed feature extraction algorithms described in previous section. We assume that the word frequency is related to the indicator of the sentiment. For instance if the word "great" is repeated in a text review: "great service and great view! The food was great also!", then we assume that the sentiment review would register as positive.

IV. Evaluation and Results

For evaluating the performance of the proposed feature extraction algorithms and several classifiers we use accuracy - which represents the percentage of test samples that are classified correctly from all test samples.

We build a framework system for exploring different approaches in feature extraction methods in combination with different classifiers as described in section III. In our first experiments, we used a sample of 10 000 business reviews dataset extracted from the filtered Yelp Challenge Dataset. We also split this dataset in 80% for training and 20% for testing.

As shown in Table I, the best accuracy score on test dataset was achieved by the system (92.6%) using the first approach presented in section III – B, when preprocessing steps are applied before the feature selection phase (removing punctuations) and removing stop words, applying stemming, handling negations and using unigrams in forming the feature vectors in bag of words representation. The result is obtained using Stochastic Gradient Descent (SGD) Classifier with the following parameters: the squared Euclidean norm L2 for penalty added to the loss function and a value of 0.0001 of the constant for multiplying the regularization term.

TABLE I. Classifier Accuracy

Features	Naïve Bayes	Linear SVC	Logistic Regression	SGD
Unigrams + Stop words + without removing punctuations + Without handling Negations	0.881	0.917	0.893	0.914
Unigrams + Stop words + without removing punctuations + Handling Negation	0.883	0.919	0.898	0.911
Unigrams + Stop words + removing punctuations + Handling Negations	0.897	0.923	0.900	0.926
Unigrams + POS + WSD	0.787	0.786	0.787	0.786

Using the same feature selection algorithm and applying Linear SVC we obtain a closed score (92.3%), but when applying Naïve Bayes for classifying positive and negative sentiments, the system performed slightly worse (89.7%). We obtain a similar accuracy score (90%) using Logistic Regression classifier (MaxEnt) with L2 norm for regularization and a tolerance of 0.0001 for the stopping criteria.

Without applying the step of handling negations in the feature selection algorithm the system's accuracy, regardless of the classifier used, decreases with 1%. Our belief that negations handling is an important step in sentiment analysis is confirmed. We show that without removing punctuations from the text review in the preprocessing phase, but applying the handling negations, the accuracy slightly decreases besides. We evaluate (Table II) the effectiveness of the learning algorithms using the first approach of the feature extraction algorithm based on precision and recall, the first for positive

sentiment identification, the other for negative sentiment identification and f1 measure - the harmonic mean of precision and recall (1.0 is the best value). Precision represents the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. Recall is as the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The results are obtained on the test dataset.

TABLE II. PRECISION, RECALL AND F-MEASURE OF CLASSIFIERS

Classifier	Positive			Negative		
	Precision	F1 Meas.	Recall	Precision	F1 Meas.	Recall
Naïve Bayes	0.874	0.714	0.604	0.900	0.937	0.976
Linear SVC	0.895	0.800	0.723	0.928	0.952	0.977
Logistic Regression	0.952	0.705	0.559	0.892	0.939	0.992
SGD	0.911	0.806	0.723	0.928	0.954	0.980

We found that the second approach of feature extraction, using POS (part-of-speech tag), applying WSD (word sense disambiguation) and computing a positive and a negative score using the sentiment opinion lexicon SentiWordNet was not useful. The accuracy of the system for all classifiers decreased with 14% when compared to the first approach results. We also found that the performance decreased during cross-validation, and thus these features were not included in the next experiments. Although this approach doesn't require training compared with the first approach, we believe that the meaning of the whole text review - which can be large - is too broad depending on the context. Sentiment ambiguity can occur, for instance a sentence of a text review containing a positive or negative word doesn't necessarily express any sentiment (e.g.: "you recommend a good restaurant close to this company").

Next, for evaluating the accuracy of the proposed methods we used the test business reviews dataset extracted from the filtered Yelp Challenge Dataset which contains about 1.34 million text reviews. We apply the preprocessing steps, remove stop words, apply stemming, handling negations and use unigrams for feature vectors representation as described in previous section (after splitting the large Yelp Challenge dataset in 80% for training and 20% for testing). The classification experiments were conducted using 3-fold cross validation for evaluating the systems accuracy.

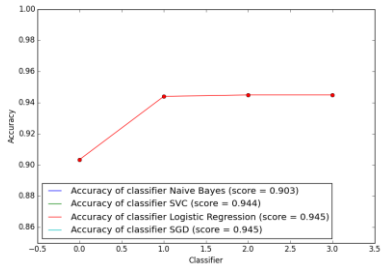

Fig. 1. Classification accuracy of bussiness reviews

A comparison of accuracy on the test dataset using different learning algorithms is shown in Fig 1. The best score of 94.4 % was obtained using the first approach on feature extraction algorithm as described in section III combined with Linear SVC on the test dataset. The same accuracy was obtained when using Logistic regression for classifying the positive and negative reviews. The system's accuracy slightly decreases by 5% when using the Naïve Bayes classifier.

V. CONCLUSIONS AND FUTURE WORK

This paper explores the usage of several feature extraction methods and classifiers for classifying business text reviews using a large dataset: Yelp challenge dataset containing more than 1.6 million reviews [1]. Our best classifiers Linear SVC and SGD have obtained an accuracy of 94.4% using the first approach proposed in the feature extraction algorithm. In terms of performance, Naïve Bayes and Logistic Regression classifiers tend to have slightly worst results.

Sentiments expression, feelings, emotions and opinions are a difficult task in humans' analysis. Wilson et al. found 82% agreement by two persons in the assignment of phrase-level sentiment polarity to automatically identify subjective expression [11]. Sentiment analysis methods proved to perform well for classifying sentiments of Yelp business reviews by taking into account the star rating given by the users. We think that the systems' accuracy could still be improved by exploring the usage of bigrams or trigrams, word chunks, or even part-of-speech (POS) as features in order to distinguish between the same word features that are used as different POS.

REFERENCES

[1] [Online]. Available: https://www.yelp.com/dataset_challenge/dataset

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, vol. 10, pp. 79-86, 2002

[3] H. Minqing and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004

[4] X. Yun, X. Wu and Q. Wang, "Sentiment Analysis of Yelp's Ratings Based on Text Reviews", pp. 1-5, 2014. [Online]. Available: http://cs229.stanford.edu/projects2014.html

[5] J. Jong, "Predicting Rating with Sentiment Analysis," pp. 1-5, 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/Jong-PredictingRatingwithSentimentAnalysis.pdf

[6] F. Mingming and M. Khademi, "Predicting a business star in Yelp from its reviews text alone", arXiv preprint arXiv:1401.0864, 2014

[7] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, "Building a Sentiment Summarizer for Local Service Reviews", *Proc. of the WWW Workshop on NLP Challenges in the Information Explosion Era*, NLPIX, 2008

[8] D. Sanjiv, M. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web", Management Science, pp. 1375-1388, 2007.

[9] A. Esuli and F. Sebastiani, "SentiWordNet:A Publicly Available Lexical Resource for Opinion Mining", *Proceedings from International Conf. on Language Resources and Evaluation (LR)*, 2006, pp. 417-422

[10] "scikit-learn" [Online]. Available: http://scikit-learn.org/

[11] T. Wilson, J. Wiebe and P. Hoffmann "Recognizing contextual polarity in phrase-levelsentiment analysis", *Proc. of the conference on Human Language Technology and Empirical Methods in NLP*, 2005