A SEMINAR REPORT ON

Sentiment Analysis with Ensemble Classifiers

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

MASTER OF ENGINEERING (Computer Engineering)

BY

Kaushik S. Hande Exam No. 6989

Under the guidance of
Prof. A. G. Phakatkar

DEPARTMENT OF COMPUTER ENGINEERING

Pune Institute of Computer Technology
Dhankawadi, Pune
Maharashtra 411043

DEPARTMENT OF COMPUTER ENGINEERING

Pune Institute of Computer Technology
Dhankawadi, Pune
Maharashtra 411043

CERTIFICATE

This is to certify that the Seminar report entitled
"Sentiment Analysis with Ensemble Classifiers"

Submitted by
Kaushik S. Hande Exam No. 6989

is a work carried out by him under the supervision of Prof. A. G. Phakatkar and it is submitted towards the partial fulfillment of the requirement of Savitribai Phule Pune University, Pune for the award of the degree of Master of Engineering (Computer Engineering)

Prof. A. G. Phakatkar
Internal Guide
PICT, Pune

Dr. R. B. Ingle
Head
Department of Computer Engineering
PICT, Pune

Place:
Date:

ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Dr. S. S. Sonawane and Head of Department Dr. R. B. Ingle for their support.
I also sincerely convey my gratitude to my guide Prof. A. G. Phakatkar, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement from beginning till end to make this seminar a grand success.
I am also hugely indebted to my friends for all their help and support.
Above all I would like to thank my parents for their wonderful support and blessings, without which I would not have been able to accomplish my goal.

Contents

List of Figures

List of Tables

ABSTRACT

With the ever increasing social networking and online marketing sites, the reviews and blogs obtained from those, act as an important source for further analysis and improved decision making. These reviews are mostly unstructured by nature and thus, need processing like classification or clustering to provide a meaningful information for future uses carrying prediction and classification analysis. Reviews are classified as either positive or negative concerning a query term. This approach is useful for consumers who can use sentiment analysis to search for products, for companies that aim at monitoring the public sentiment of their brands, and for many other applications.We considered three different machine learning algorithms such as Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM)for classification of human sentiments. Classifier ensembles formed by Naive Bayes, SVM, and Maximum Entropy improves classification accuracy.

1 INTRODUCTION

Sentiment is an attitude, thought, or judgement prompted by feeling. Sentiment analysis is also known as opinion mining, it involves studing of people's sentiments towards certain entities. Internet is a resourceful place with respect to sentiment information. From a perspective of a user, people are able to express their views through various social media, such as forums, micro-blogs, or online social networking sites.

With the advent of Web 2.0 techniques, users started prefering to share their opinions on the Web. These user-generated and sentiment-rich data are valuable to many applications like credibility analysis of news sites on the Web, recommendation system, business and government intelligence etc. At the same time, it brings urgent need for detecting overall sentiment inclinations of documents generated by users, which can be treated as a classification problem. Sentiment analysis includes several subtasks which have seen a great deal of attention in recent years:

1. To detect whether a given document is subjective or objective.
2. To Identify whether given subjective document express a positive opinion or a negative opinion.
3. To determine the sentiment strength of a document, such as strongly negative, weakly negative, neutral, weakly positive and strongly positive.

In this work we are focusing on second subtask.

Besides individuals on social media marketers also need to monitor all media for information related to their brands — whether it's for public relations activities, fraud violations, or competitive intelligence. Thus, aside from individuals, sentiment analysis is also the need of companies which are anxious to understand how their products and services are perceived by the public.

The movie reviews are mostly in the text format and unstructured in nature. Thus, the stop words and other unwanted information are removed from the reviews for further analysis. These reviews goes through a process of vectorization in which, the text data are converted into matrix of numbers. These matrices are then given input to different machine learning classifiers for classification of the reviews.

Many researchers have focused on the use of traditional classifiers, like Naive Bayes, Maximum Entropy, and Support Vector Machines to solve such problems. In this work, we show that the use of ensembles of multiple base classifiers can improve the accuracy of review sentiment classification.

2 MOTIVATION

In traditional document classification tasks, the input to the machine learning algorithm is a free text, from which a bag-of-words representation is constructed — the individual tokens are extracted, counted, and stored as vectors.

(i) Most of the authors apart from Pang et al., and Matsumoto et al., have used unigram approach to classify the reviews. This approach provides comparatively better result, but fail in some cases. The comment "The item is not good," when analyzed using unigram approach, provides the polarity of sentence as neutral with the presence of one positive polarity word 'good' and one negative polarity word 'not'. But when the statement is analyzed using bigram approach, it gives the polarity of sentence as negative due to the presence of words 'not good', which is correct. Therefore, when a higher level of n-gram is considered, the result is expected to be better. Thus, analyzing the research outcome of several authors, this study makes an attempt to extend the sentiment classification cation using unigram, bigram, trigram, and their combinations for classification of movie reviews.

(ii) Also a number of authors have used Part-of-Speech (POS) tags for classification purpose. But it is observed that the POS tag for a word is not fixed and it changes as per the context of their use. For example, the word 'book' can have the POS 'noun' when used as reading material where as in case of "ticket booking" the POS is verb. Thus, in order to avoid confusion, instead of using POS as a parameter for classification, the word as a whole may be considered for classification.

(iii) Most of the machine learning algorithms work on the data represented as matrix of numbers. But the sentiment data are always in text format. Therefore, it needs to be converted to number matrix. Different authors have considered TF or TF-IDF to convert the text into matrix on numbers. But in this paper, in order to convert the text data into matrix of numbers, the combination of TF-IDF and CountVectorizer have been applied. The rows of the matrix of numbers represents a particular text file where as its column represent each word / feature present in that respective file.

3 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques proposed in various refferences:

No.	Reference	Techniques	Desciption
1	Tweet sentiment analysis with classifier ensembles	Naive Bayes Maximum Entropy Support Vector machine,	Classify the dataset using different machine learning algorithms and n-gram model with ensemble classifiers
2	Classification of sentiment reviews using n-gram machine learning approach	Naive Bayes Maximum Entropy Support Vector machine,	Classify the dataset using different machine learning algorithms and n-gram model
3	Thumbs up?sentiment classification using machine learning techniques	Naive Bayes Maximum Entropy Support Vector machine,	Classify the dataset using different machine learning algorithms and n-gram model
4	Automatic opinion polarity classification of movie.	Naive Bayes (NB) and Markov Model (MM)	Accessed overall opinion polarity(OvOp)concept using machine learning algorithms
5	The sentimental factor: improving review classification via human-provided information.	Naive Bayes	Linearly combinable paired feature are used to predict the sentiment
6	Sentiment analysis using support vector machines with diverse information sources	Support Vector Machine (SVM)	Values assigned to selected words then combined to form a model for classification
7	Mining the peanut gallery: opinion extraction and semantic classification of product reviews	SVM, Machine learning using Rainbow, Naive Bayes	Information retrieval techniques used for feature retrieval and result of various metrics are tested
8	Sentiment classification using word sub-sequences and dependency sub-trees	Support Vector Machine (SVM)	Syntactic relationship among words used as a basis of document level sentiment analysis
9	Chinese comments sentiment classification based on word2vec and svm perf	Support Vector Machine (SVM)	Use word2vec to capture similar features then classify reviews using SVM

Table 1: **LITERATURE SURVEY**

4 PROBLEM DEFINITION AND OBJECTIVE

4.1 Problem Statement

To increase accuracy and efficiency of classification of reviews using ensemble classifiers.

4.2 Proposed Objective

To improve the accuracy of classification of reviews by using

1. Bag-of-words representation of text.
2. Unigram, bigram, trigram and combination of these.
3. Removing stopwords, Numeric and special character from text as they do not play any significant role.
4. Using combination of three different classifiers for text review classification.

5 PROPOSED SOLUTION

The reviews of IMDb dataset is processed to remove the stop words and unwanted information from dataset. The textual data is then transformed to a matrix of number using vectorization techniques. Further, training of the dataset is carried out using machine learning algorithm.

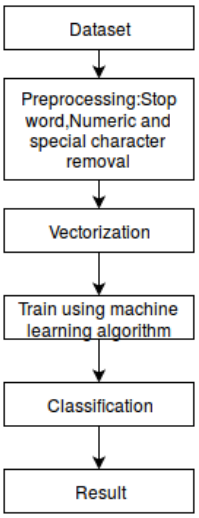

Figure 1: Diagrammatic view of the approach

5.1 Preprocessing

The text reviews sometimes consist of absurd data, which need to be removed, before considered for classification. The usually identified absurd data are:

1. Stop words: They do not play any role in determining the sentiment.
2. Numeric and special character: In the text reviews, it is often observed that there are different numeric (1,2,...5 etc.) and special characters which do not have any effect on the analysis. But they often create confusion during conversion of text file to numeric vector.

5.2 Bag of Words

After the preprocessing of text reviews, reviews are represented by a table in which the columns represent the terms (or existing words) in the reviews and the values represent their frequencies. Therefore, a collection of reviews after the preprocessing step addressed later can be represented as illustrated in Table 2, in which there are n reviews and m terms.Each review is represented as $review_i = (a_{i1}, a_{i2}, ..., a_{im})$, where a_{ij} is the frequency of term t_j in the $review_i$. This value can be calculated in various ways.

1. CountVectorizer: It converts the text reviews into a matrix of token counts. It implements both tokenization and occurrence counting. The output matrix obtained after this process is a sparse matrix.

2. Calculation of CountVectorizer Matrix: An example is considered to explain the steps of calculating elements of the matrix Garreta and Moncecchi (2013) which helps in improving the understandability. Suppose, three different documents containing following sentences are taken for analysis:

   (a) Sentence 1: "Movie is nice".

   (b) Sentence 2: "Movie is Awful".

   (c) Sentence 3: "Movie is fine".

   A matrix may be formed with different values for its elements size 4 * 6, as there exists 3 documents and 5 distinct features. In the matrix given in Table 3 , the elements are assigned with value of '1', if the feature is present or else in case of the absence of any feature, the element is assigned with value '0'.

5.3 Classifiers

Naive Bayes, Support vector machine and Maximum Entropy are used as classifiers for sentiment analysis .

- Naive Bayes (NB) method: This method is used for both classification as well as training purposes. This is a probabilistic classifier method based on Bayes' theorem. In this work, multinomial Naive Bayes classification technique is used. Multinomial model considers word frequency information in document for analysis, where a document is considered to be an ordered sequence of words obtained from vocabulary 'V'. The probability of a word event is independent of word context and it's position in the document.

- Support vector machine (SVM) method: This method analyzes data and defines decision boundaries by having hyper-planes. In binary classification problem, the hyper-plane separates the document vector in one class from other class, where the separation between hyper-planes is desired to be kept as large as possible.

- Maximum entropy (ME) method: In this method, the training data is used to set constraint on conditional distribution. Each constraint is used to express characteristics of training data. In Maximum Entropy (ME) if a word occurs frequently in a class, the weight of word-class pair becomes higher in comparison to other pairs. These highest frequency word-class pairs are considered for classification purpose.

The movie reviews of acl IMDb dataset is considered for analysis, using the machine learning algorithms discussed. Then different variation of the n-gram methods i.e., unigram, bigram, trigram, unigram + bigram, unigram + trigram, and unigram + bigram + trigram are applied to obtain the result.

5.4 Ensemble Classifiers

In practice, classifiers are built to classify unseen data, usually referred to as a target dataset. In a controlled experimental setting, a validation set represents the target set. Actually, in controlled experimental settings the target set is frequently referred to as either a test or a validation set. These two terms have been used

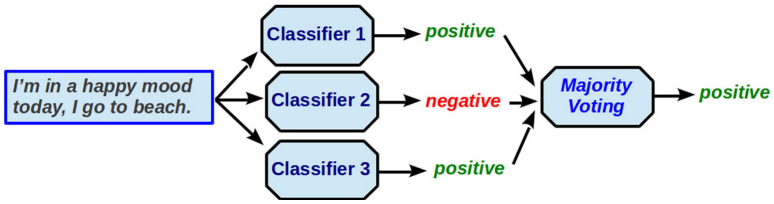

Figure 2: Diagrammatic view of the ensmble approach

interchangeably, sometimes causing confusion. In our study, we assume that the target/validation set has not been used at all in the process of building the classifier ensembles. Once the base classifiers have been trained, a classifier ensemble is formed by the average of the class probabilities obtained by each classifier or the majority voting.

5.5 Mathematical Model

$S = \{s, e, I, O, fmain|\phi\}$

where,
s = start state
e = end state
I = Inputs to the system
I = {y,D}

D = { d1, d2, d3 ..... dn }=set of n different reviews
y ϵ { 0 , 1 } = The class label of the reviews
0 = Negative reviews
1 = Positive reviews
O = Output of the system
O = {D_o, y }
where

$D_o = d_1, d_2, d_3.....d_n$=set of n different test reviews
y ϵ 0 , 1 = The class label of the test reviews
$f_{main} = \{f_{vectorizer}, f_{tf-idf}, f_{classifier}\}$
$f_{vectorizer}$ = function for vectorising each review
f_{tf-idf} = function for tf-idf
$f_{classifier}$ = classifier for the prediction of class of review
O$\Leftarrow f_{vectorizer} \cup f_{classifier}$
O$\Leftarrow f_{tf-idf} \cup f_{classifier}$

6 CONCLUSION

It makes an attempt to classify movie reviews using different supervised machine learning algorithm. We also used CountVectorizer to improve the accuracy of classification. This algorithm is further applied using n-gram approach on IMDb dataset. It is observed that as the value of 'n' in n-gram increases the classification accuracy decreases i.e., for unigram and bigram, the result obtained using the algorithm is remarkably better; but when trigram classification are carried out, the value of accuracy decreases. Also ensemble approach increases the classification accuracy.

References

[1] Nadia F.F., Da Silva A, Eduardo R. Hruschka a , Estevam R. Hruschka Jr., "Tweet sentiment analysis with classifier ensembles", in Decision Support Systems Volume 66, Pages 170-179, 2014

[2] A Tripathy, A Agrawal, SK Rath , "Classification of sentiment reviews using n-gram machine learning approach", in Expert Systems with Applications Volume 57, Pages 117–126, 2016

[3] S. Das and M. Chen,"Yahoo! for Amazon: Extracting market sentiment from stock message boards," *Proceedings of the Asia Pacific Finance Association Annual Conference,* 2001.

[4] Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 79-86, 2002.

[5] B. Pang and L. Lee,"Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval,vol. 2, no. 1-2, pp. 1-135, 2008.*

[6] Beineke, P. , Hastie, T. ,Vaithyanathan, S., "The sentimental factor: improving review classification via human-provided information.," *In Proceedings of the 42nd annual meeting on association for computational linguistics (p. 263). Association for Computational Linguistics 2004.*

[7] Dave, K. , Lawrence, S. ,Pennock, D. M., "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *Proceedings of the 12th international conference on World Wide Web (pp. 519–528). ACM.*

[8] Matsumoto, S. , Takamura, H. ,Okumura, M., "Sentiment classification using word sub-sequences and dependency sub-trees," *n Advances in knowledge discovery and data mining (pp. 301–311). Berlin Heidelberg: Springer* 2005.

[9] Mullen, T. , & Collier, N. "Sentiment analysis using support vector machines with diverse information sources." *In EMNLP: 4 (pp. 412–418) .,* (2004).

[10] Niu, T. , Zhu, S. , Pang, L. , & El Saddik, A. . I ."Sentiment analysis on multi-view social data."*n Multimedia modeling (pp. 15–27). Springer,* 2016.

[11] Salvetti, F. , Lewis, S. , & Reichenbach, C. . "Automatic opinion polarity classification of movie."*Colorado research in linguistics, 17 , 2* 2004.

[12] Yuan Wang , Zhaohui Li, Jie Liu , Zhicheng He , Yalou Huang , and Dong Li "Word Vector Modeling for Sentiment Analysis of Product Reviews" *NLPCC 2014,*496, pp. 168-180, 2014.

[13] Zhang, D. , Xu, H. , Su, Z. , & Xu, Y. " Chinese comments sentiment classification based on word2vec and svm perf." *Expert Systems with Applications, 42 (4), 1857–1863* .2015

[14] Na, J.C., Sui, H., Khoo, C., Chan, S., and Zhou, Y."Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews "*Proceedings of the Eighth International ISKO Conference (pp. 49-54)* 2004.

[15] Rui Xia , Chengqing Zong , Shoushan Li ,"Ensemble of feature sets and classification algorithms for sentiment classification"*Information Sciences 181 1138–1152(2011)*

[16] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL),* 2002.