International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:3, 2014

Enhance the Power of Sentiment Analysis

Yu Zhang, Pedro Desouza

Abstract**—**Since big data has become substantially more accessible and manageable due to the development of powerful tools for dealing with unstructured data, people are eager to mine information from social media resources that could not be handled in the past. Sentiment analysis, as a novel branch of text mining, has in the last decade become increasingly important in marketing analysis, customer risk prediction and other fields. Scientists and researchers have undertaken significant work in creating and improving their sentiment models. In this paper, we present a concept of selecting appropriate classifiers based on the features and qualities of data sources by comparing the performances of five classifiers with three popular social media data sources: Twitter, Amazon Customer Reviews, and Movie Reviews. We introduced a couple of innovative models that outperform traditional sentiment classifiers for these data sources, and provide insights on how to further improve the predictive power of sentiment analysis. The modeling and testing work was done in R and Greenplum in-database analytic tools.

Keywords**—**Sentiment Analysis, Social Media, Twitter, Amazon, Data Mining, Machine Learning, Text Mining.

I. Introduction

SENTIMENT analysis, or opinion mining, is a relatively new branch of text mining and natural language processing (NLP). Generally speaking, sentiment analysis refers to data mining technologies for detecting subjective information, specifically the positive, negative, or neutral opinions toward a certain topic in the text. Because sentiment analysis can be applied to detect favorable and unfavorable opinions toward specific topics, brands, organizations, or products within large amount of text data, it is attracting ever more attention in marketing analysis, customer risk management, brand and reputation management, financial services, and social study, etc. [1].

A large data source for sentiment analysis is public social media, as it has become extremely popular amongst Internet users across the globe. Everyday millions of people share their lives and express their opinions on various topics on Twitter, Facebook, Amazon, Google+, etc. Both corporations and consumers can benefit significantly from mining information in these data sources. However, a big challenge is that the social media data are mostly unstructured online text, which cannot be downloaded or processed easily using traditional tools. Another obstacle is that it can be difficult to find relevant information from large datasets. For example, on Twitter users send over 400 million tweets per day; on Facebook there are over 2.5 billion "likes" and comments per day; on Amazon there are over 150 million reviews about the products they sell. If one wants to find out the sentiment scores about a certain product, say an iPhone 5S, he or she needs to extract data related only to this product from literally millions of online records.

Yu Zhang, PhD, is with EMC Corporation, Hopkinton, MA 01748 USA (phone: 1-347-574-3802; e-mail: kennethzhangny @gmail.com).
Pedro Desouza, PhD, is with EMC Corporation, Hopkinton, MA 01748 USA (phone: 1-214-704-6745; e-mail: pedro.desouza@emc.com).

Sentiment analysis has been studied and applied by scientists and researchers in the past ten years, and Twitter has been used as data source a number of times because its popularity [2]-[4]. The performances of different sentiment models have also been compared by some researchers [5]-[7]. However, there has not been any comparison between the efficiency of different sentiment models working with different social media data sources. It is intuitive that a good sentiment model for Amazon Reviews may not work so well with Twitter, as Tweets are normally shorter, with more frequent typos and grammar errors, than Amazon reviews.

There are a number of tools that can be used to conduct sentiment analysis, such as SAS Text Miner & Enterprise Miner, IBM SPSS, R, and online sentiment analysis tools like Meltwater, Google Analytics and Facebook Insights, etc. In this practice we used R to develop the sentiment analysis module because of its open-source feather and flexibility in text mining. The same module has also been created in Greenplum database for dealing with big data via two in-database analytic tools GPText and MADlib.

In this paper, we introduced several data mining methods for sentiment analysis, and we implemented these methods with three representative social media sources: Twitter (Tweets), Amazon Reviews, and Movie Reviews. Tweets are very short (up to 140 characters) and often contain misspellings or typos; Amazon Reviews are longer than Tweets, and usually are of better quality; Movie Reviews are even longer than Amazon Reviews, and of higher quality. The details of the three data sources will be mentioned in the next section.

Our approach here includes a fundamental task in sentiment analysis, which is classifying the polarity of a given text at the document – whether the expressed opinion in a document, a sentence, or an entity feature is positive, negative or neutral.

In addition to tradition classification methods, we developed a new sentiment analysis algorithm in order to enhance the predictive power and accuracy. All the methods were applied to each of the three data sources, and the misclassification rates have been used for performance comparison.

II. Methodology

A typical work flow of our sentiment analysis can be described as

ISSN: 2517-9942
Vol:8, No:3, 2014

Fig. 1 The work flow of the sentiment analysis module

Step 1. Set up input parameters & extract raw text data (online reviews, blogs, Tweets, or other documents).
Step 2. Process raw data:
. Clean up text
. Remove stop words
. Stemming
. Translate text into corpus matrices
Step 3. Conduct a few classifiers to calculate the polarity of the formatted data.
Step 4. Evaluate the accuracy and efficiency of each algorithm.
Step 5. Produce outputs (sentiment scores, spreadsheets, graphs).

In the data extraction step, we wrote a program in R that can extract Tweets automatically through the Twitter API. For comparison, we also downloaded movie reviews from Pang & Lee's website [8], and Amazon customer reviews from Amazon.com manually.

In the data processing step, feature information was extracted from the raw text. Features could be single words (unigrams), two-word phrases (bigrams), sentences, etc. Previous research [6] shows that best performance is often approached when using the unigram feature. Hence in this paper we illustrate our models and comparison based on the unigram feature.

In the modeling and classification step, we conducted five classifiers in R: SScore, Maximum Entropy, SVM, Weighted Maximum Entropy, and Weighted SVM. A thousand mixed documents have been used to build the training set.

In the evaluation step, each classifier was evaluated with different data sources, and the one that performs the best was used to produce output. The accuracy of a sentiment analysis is about how well it agrees with human judgments. According to research human raters usually agree only about 79% [9]. In other words, a model is considered nearly as good as humans if the accuracy is over 70%. When processing good quality documents, our methods often outperform the 79% human produced baseline; when processing poor quality text, the accuracy is lower but close to 70%.

In the output step, the sentiment scores together with other information were saved into Excel spreadsheets, and the sentiment distributions are visualized in a number of figures.

The rest of this section includes a detailed description of the data preparation and an introduction of each classifier.

A. Data Preparation

The data preparation includes both data extraction and data processing. In the data extraction step, a thousand Tweets (500 positive, 500 negative) containing emoticons were downloaded automatically through our R code by calling the Twitter API. The sentiment of each Tweet was determined by emoticons as in Pak et al.'s paper [4]:

Flag positive if the Tweet has at least one of the happy emoticons:

":)", ":-)", "=)", ":D", ":P", "(:", "(-:", "(=", etc.

Flag negative if the Tweet has at least one of the sad emoticons:

":-(", ":(", "=(", ")-:", "):", ")=", etc.

About 230 Amazon reviews were downloaded manually from Amazon.com, and the sentiment of each review was determined by the ranking stars (5 stars = positive; 1 star = negative); and about 400 movie reviews (200 positive and 200 negative) were randomly picked from Pang & Lee's website [8].

Fig. 2 shows examples of each document type.

Fig. 2 Examples of Movie Reviews, Tweets and Amazon Reviews. The positive and negative words are marked in green and red respectively

It is clear that movie reviews often consist of a couple of hundred words of relatively good quality; Amazon reviews are shorter but also of reasonable quality; Tweets are much shorter (one or two sentences only) and often contain typos, misspellings, and little or no grammatical features. It is therefore intuitive to consider using different classifiers for

International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:3, 2014

different data sources in order to maximize the accuracy or minimize the misclassification rate.

Before we can apply any data mining method for sentiment analysis, we need to transform the text files into digital format. At first all the text files are saved together as a Corpus, an object defined in R for storing a collection of text documents.

A series of transformations can then be implemented to all elements of the corpus in order to clean the text. This process includes converting texts into plain text documents, eliminating extra whitespace, converting all letters into lower case, removing numbers, punctuations and user-defined stop words, and stemming. Here are examples for these steps:

- Change all letters into lower case:
  "LOVE" will become "love"
- Remove numbers and punctuations:
  0-9, $#%@! ... will be removed
- Remove stop words (irrelevant to sentiment):
  you, me, we, us, in, on, to, here, there… will be removed
- Stemming (reduce words to their stems) (Fig. 3)

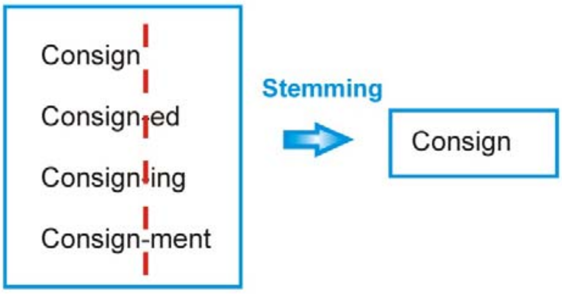

Fig. 3 Stemming

After these steps the corpus is considered "clean".

The next step is to transform the formatted corpus into a DocumentTermMatrix object whose rows represent documents and columns represent terms. Each entry, for instance Entry (*i, j*), represents either the frequency or the presence of Term *j* in Document *i*. Figs. 4 & 5 illustrate the term frequency matrix and term presence matrix, respectively.

	term1	term2	term3	term4	term5	term6	term7	term8	term9	term10	term11	term12	term13	term14
Doc1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc2	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Doc3	0	2	0	0	1	0	1	0	0	0	0	0	0	0
Doc4	0	0	1	0	0	0	0	0	0	3	5	0	2	0
Doc5	0	0	0	2	0	0	2	0	0	0	0	0	0	0
Doc6	0	0	3	0	0	0	0	0	4	0	0	0	0	0
Doc7	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Doc8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc9	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Doc10	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The number of times term 13 appears in document 4

Fig. 4 Term Frequency Matrix

	term1	term2	term3	term4	term5	term6	term7	term8	term9	term10	term11	term12	term13	term14
Doc1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc2	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Doc3	0	1	0	0	1	0	1	0	0	0	0	0	0	0
Doc4	0	0	1	0	0	0	0	0	0	1	1	0	1	0
Doc5	0	0	0	1	0	0	1	0	0	0	0	0	0	0
Doc6	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Doc7	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Doc8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc9	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Doc10	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1: presence
0: absence

Documents show that term-presenc matrix often outperforms term-frequency matrix

Fig. 5 Term Presence Matrix

As the data is in matrix form, all the classification models can be applied and compared. Documents show that classifiers often perform better with term-presence matrix than with term-frequency matrix [6], therefore we conducted our comparisons with the term-presence matrices in this practice.

B. SScore Classifier

A simple approach of sentiment analysis is to use a list of positive and negative sentimental words to evaluate reviews. For each review the number of positive and negative words is counted and the polarity of the review is determined by the higher count. Hu & Liu's opinion lexicon contains over 2,000 positive words and over 3,700 negative words [10] which has been used as a baseline by a group of researchers at Stanford University in their Twitter sentiment classification [2].

Another example is that NetBase Solutions Inc. [11] computes a Net Sentiment Score (NSS) for classification:

$$\text{NSS} = \left(\frac{\text{Positive Mentions-Negative Mentions}}{\text{Positive Mentions+Negative Mentions}}\right)\times 100 \qquad (1)$$

The NSS can be any number between -100 and 100, and the higher the number is the more positive the review is considered. We used this formula to construct the SScore classifier. A difference is that the NSS is localized to individual sentences, but we applied this formula to score documents. The two methodologies are very similar.

The SScore Classification is one of the most straight-forward methods for sentiment analysis, and it can be implemented easily. The accuracy may not be as high as most machine learning algorithms in many cases, though it can serve as a good baseline for performance evaluation and comparison of algorithms.

C. Maximum Entropy (MaxEnt)

Maximum entropy (MaxEnt) or log-linear classification has been proven effective in a number of natural language processing applications [12], [6], [13].

Let $(f_1, \ldots, f_m)$ be a predefined set of m features that can appear in a document d. The features can be unigrams, bigrams or n-grams. Let $n_i(d)$ be the number of times f_i occurs in document d. Each document can then be represented by the document vector

$$\vec{d} = (n_1(d), n_2(d), \ldots, n_m(d)) \qquad (2)$$

The classifier of Maximum Entropy method is defined as

$$P_{ME}(c\,|\,d) = \frac{1}{Z(d)}\exp\Big(\sum_{i=1}^{m}\lambda_{i,c}F_{i,c}(d,c)\Big) \qquad (3)$$

where $Z(d)$ is a normalization factor:

$$Z(d) = \sum_{c}\exp\Big(\sum_{i=1}^{m}\lambda_{i,c}F_{i,c}(d,c)\Big) \qquad (4)$$

and $F_{i,c}$ is a feature/class function for feature f_i and class c, defined as

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

The $\lambda_{i,c}$'s are feature-weight parameters that are set so as to maximize the entropy of the induced distribution subject to the constraints that the expected values of the feature/class functions with respect to the model are equal to their expected values with respect to the training data.

Unlike Naive Bayesian classifier, Maximum Entropy method makes no assumptions about the relationships between features; therefore it may potentially perform better than Naive Bayesian when conditional independence assumptions are not met.

D. SVM

Support vector machine (SVM) has been shown to be highly effective for text classification and generally outperforms Naive Bayesian and some other classifiers [6]; [14]. The basic idea of SVM in two-category case is to find a hyperplane that separates the document vectors in one class from those in the other and maximizes the margin at the same time. In our case, we define

$$\underline{c}_j \in \{-1, 1\} \qquad (6)$$

where -1 represents negative opinions and 1 represents positive. The hyperplane can then be defined as

$$\underline{w} = \sum_j \alpha_j \underline{c}_j \underline{d}_j, \ \ \alpha_j \geq 0 \qquad (7)$$

The $\underline{d}_j$'s are training vectors and also called support vectors. The α_j's are obtained by solving a dual optimization problem. Classification of test vectors consists of determining which side of hyperplane they fall in [15]; [6].

E. Weighted Maximum Entropy (WMaxEnt)

In order to further enhance the predictive power of Maximum Entropy classifier, we created a new model by adding a weight scalar on each term in the term-presence matrix so that the sentimental terms (both positive and negative terms) weigh more than others. The weighted term-presence matrix is defined as

$$M_w = M * diag(w_1, w_2, ..., w_N) \qquad (8)$$

where M represents the original term-presence matrix, M_w represents the weighted term-presence matrix, and w_i represents the weight for term i (column i in M).

We then conducted the Maximum Entropy method on weighted training and test sets, and we called this new method Weighted Maximum Entropy Classifier, or WMaxEnt.

F. Weighted SVM (WSVM)

Similar to the concept of Weighted Maximum Entropy classifier, we ran SVM with weighted training and test sets, and called it Weighted SVM Classifier, or WSVM.

The results show that the Weighted Maximum Entropy and Weighted SVM methods outperform other classifiers in most of our comparisons.

III. Results and Comparison

The results of our sentiment analysis module consist of sentimental scores (in this case positive, negative or neutral), summarized text information, and graphical results. Fig. 6 shows a few examples of the sentiment result for Tweets regarding the topic "HSBC" on November 15^{th}, 2012. SScore was used as the classifier, as it produced the lowest misclassification rate amongst all the classifiers.

ID	SentiClass	Author	Text
1	Positive	chris_greenf	Well done to #HSBC Norwich today for their Open for Business day- useful networking & advice. @StoreyJoanna
2	Neutral	LucasRouxdeLuze	#Indian merchandise #exports will expand more in the next eight years than those of any other country including #China- according to #HSBC
3	Neutral	AnsonBailey	Provident Financial Holdings- Inc. (PROV) Ex-Dividend Date Scheduled for ... http://t.co/raH96YF8 #hsbc
4	Neutral	NitWiterazzi	@suchitrak You are the lady who brought an entire empire down #HSBC with your sheer willpower.. The other person was Gandhiji.. Jai ho Luv u
5	Positive	CathrynHayes	Why Do Banks Like Franchising - & how easy is it to get finance? #HSBC http://t.co/YbauolsP #franchisetips
6	Negative	BarringtonBex	Just got asked by assistant in #hsbc hertford if my weekly wages were a tip.. i cant decide how i feel about this incident.
7	Negative	suchitrak	Senior vice president- business relationship manager at #HSBC- John Cruz discovered massive fraud and money laundering. http://t.co/9ygPRq4r
8	Negative	asher4manu	#hsbc representatives are soo slow. #lloydstsb #BestInTheWorld
9	Neutral	samrose6	When did banks stop being able to give u cash over the counter if their ATMs have stopped working #awkward #HSBC
10	Positive	MsMaduna	#HSBC is actively recruiting for fantastic female talent for a range of careers. 300 open vacancies http://t.co/p4fWI8FX @iROCK_UK
11	Negative	PHPology	@PHPology @HSBCUKBusiness maybe someone at #hsbc will reply back to me rather than ignore a #UX flaw. hmmmmmmm
12	Positive	CathrynHayes	'Growth Pioneers' take control of British business success http://t.co/6ushzL9n #HSBC
13	Positive	d_mcconnell	RT @wherewomenwork: #HSBC is actively recruiting right now for fantastic female talent for a range of brilliant careers. 300 open...

Fig. 6 Sentiment results for Tweets regarding "HSBC"

By comparing the Score column and the Text column, one can see that most of the Tweets seem to be scored correctly, and the misclassification rate is lower than 30%.

The sentiment result can be summarized and visualized as in Fig. 7.

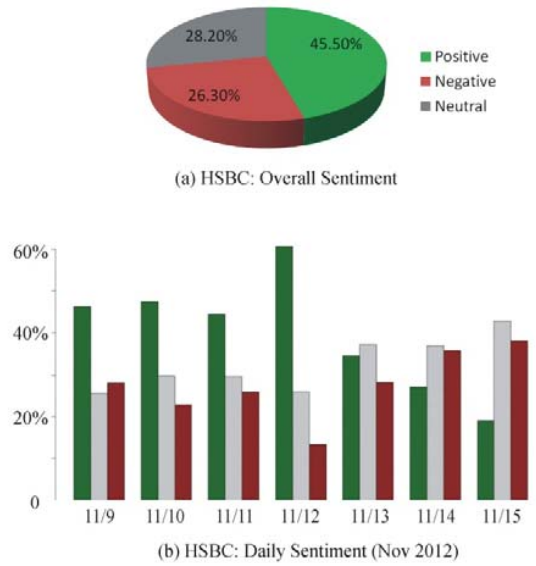

Fig. 7 The sentiment distribution for Tweets related to the topic "HSBC". (a): The total distribution of positive, negative and neutral based on seven days Tweets. (b): The daily distribution of Tweets between Nov 9th and 15th 2012

In Fig. 7 (a) we can see based on our sentiment analysis, in this seven-day period 45% Tweets expressed positive opinions about HSBC; 28% Tweets showed negative opinions; and around 26% of them were neutral. In Fig. 7 (b) we can see the daily sentiment distributions, which can be used for marketing analysis and customer risk analysis. For example, it may be useful to know why there are more positive Tweets on Nov 12th than other days, and why the number of negative Tweets increases in the last three days consecutively. Was this high positive number caused by the release of a new product or a promotion? Was the negative Tweets caused by competition from other banks? This kind of information will help the HSBC marketing department figuring out how to prevent the current clients from attrition and attract new clients.

The performances of all the five classifiers, specifically the misclassification rates, are compared in Fig. 8 for each of the three social media data sources (Twitter, Amazon Reviews and Movie Reviews).

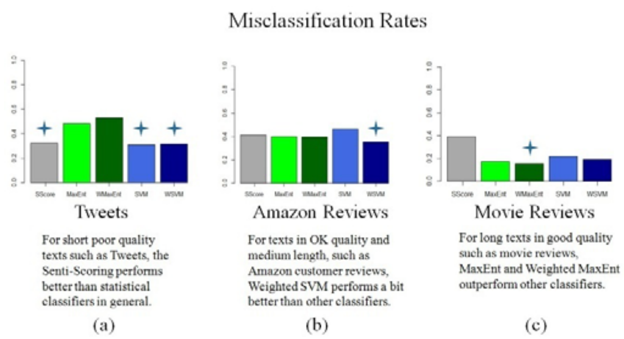

Fig. 8 Comparison of misclassification rates of the five classifiers for (a) Tweets, (b) Amazon Reviews, and (c) Movie Reviews

Fig. 8 shows that the best classifier(s) vary when dealing with different kinds of data sources. For Tweets, SScore, SVM, and WSVM outperform MaxEnt and WMaxEnt by a noticeable percentage. The lowest misclassification rate is around 33%. When working with Amazon Reviews, the WSVM classifier outperforms others and reaches about 28% misclassification rate. For Movie Reviews, the misclassification rates are generally lower than for the other two data sources except the SScore method, and WMaxEnt reaches 18% misclassification rate and is therefore considered the best classifier.

IV. Discussion and Future Work

Sentiment Analysis is a relatively new concept and technology in the world of data mining and machine learning. In the last fifteen years, scientists have made great achievements in improving the predictive power of sentiment analysis, even though they face various difficulties from data quality control, data extraction and data transformation. For example, Pang and Lee have, since 2002, done tremendous work in creating several machine learning models to solve sentiment analysis problems [6], [7], [16], [17]; Pak and Paroubek conducted sentiment analysis with Twitter API and Tweets [4]; Duric and Song did research on feature selection for sentiment analysis based on content and syntax models [18], etc. There have also been some comparisons between the performances of sentiment methods [5], [19]. However, there is insufficient research on how to select appropriate method(s) for text data from different resources and of differing quality.

This paper provides an innovative approach in this aspect: three popular social media data sources have been included for performance comparison. For each data source five sentiment classifiers were created, two of which are innovative models for enhancing the predictive power (Weighted MaxEnt and Weight SVM), and their performances were compared quantitatively. Our research shows that the selection of sentiment models should take data feature and data quality into account. A model that works well with Tweets may not be a good choice for analyzing customer reviews.

In additional to the models shown in this paper, we have conducted more research on improving the power of the sentiment analysis. For example, we found that the misclassification rate can be lowered by grouping some terms or features that often appear together, since by doing so the sparseness of the term-presence matrix can be decreased. Further, the usage of another statistic metric tf-idf, or term frequency-inverse document frequency, also shows slight improvement in accuracy, as it reflects how important a word is to a document in a collection of corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.

We will continue our research on improving the predictive power of sentiment analysis for various social media data sources. We will add the sentence-based sentiment scores into account [5], and see if it helps to improve performance. We will also create and test our models with other social media sources, such as Facebook Comments, Google+, Yelp, etc., and report insights in future publications.

As a novel and efficient analytic tool for both corporations and customers in various industries, sentiment analysis is in real needs, and these needs and technical challenges will keep us active for years to come.

ACKNOWLEDGMENT

This research is sponsored by EMC Corporation and EMC Consulting – Global Services.

REFERENCES

[1] IBM. *IBM Research - Tokyo / Text Mining*. Available from: http://www.trl.ibm.com/projects/textmining/takmi/sentiment_analysis_e.htm.
[2] Alec Go, R.B., Lei Huang, *Twitter Sentiment Classification Using Distant Surpervision*. 2009, Stanford University.
[3] twitrratr. Available from: http://twitrratr.com/.
[4] Alexander Pak, P.P., *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* IREC 2010, Seventh International Conference on Language Resources and Evaluation 2010.
[5] Rudy Prabowo, M.T., *Sentiment Analysis: A Combined Approach.* Journal of Informetrics, 2009. **3**(2): p. 143-157.
[6] Bo Pang, L.L., *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Proceedings of EMNLP, 2002: p. PP. 79-86.
[7] Bo Pang, L.L., *Opinion Mining and Sentiment Analysis.* Now Publishers Inc, 2008.
[8] Bo Pang, L.L., *Movie Review Data.* Cornell University.
[9] Ogneva, M., *How Companies Can Use Sentiment Analysis to Improve Their Business.* Mashable, 2010.
[10] Bing Liu, M.H., *A list of positive and negative opinion words or sentiment words for English.* UIC.
[11] Bricker, E., *Can Social Media Measure Customer Satisfaction?* 2011, NetBase Solutions Inc.
[12] Soo-min Kim, E.H., *Automatic Identification of pro and con Reasons in online reviews.* Proceedings of COLING/ACL, 2006.
[13] A Berger, S.D.P.a.V.D.P., *A Maximum entropy approach to natural language processing.* Computational Linguistics, 1996. 22(1).
[14] Joachims, T., *Text categorization with support vector machines: learning with many relevant features.* Proceedings of the European Conference on Machine Learning (ECML), 1998: p. 137-142.
[15] Burges, C.J.C., *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. 2: p. 121-167.
[16] Bo Pang, L.L., *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts.* Proceedings of the ACL, 2004: p. 271-278.
[17] Bo Pang, L.L., *Seeing stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales.* Proceedings of ACL, 2005.
[18] Song, A.D.a.F., *Feature Selection for Sentiment Analysis Based on Content and Syntax Models.* Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2011. ACL-HLT 2011: p. 96-103.
[19] Ishwinder Kaur, A.J.H., *A Comparison of LSA, WordNet and PMI-IR for Predicting User Click Behavior.* CHI, 2005.