Sentiment Analysis of User-Generated Online Content

Bobicev V.
Catedra Informatica Aplicată
Universitatea Tehnică a Moldovei
Oraşul Chişinău, Republica MOLDOVA
victoria.bobicev@gmail.com

Sokolova M.
University of Ottawa,
Institute for Big Data Analytics,
Ottawa, Canada
sokolova@uottawa.ca

***Abstract* — This paper presents several experiments in the domain of automate text sentiment analysis. Comparison between machine learning (ML) and rule-based algorithms demonstrated that well-tuned rule-based methods obtain better results than general ML methods and it is necessary to use various types of features for obtaining satisfactory accuracy using ML algorithms.**

***Key words* — Natural Language Pricessing, Text analysis, Sentiment Analysis, Machine Learning Algorithms, Rule – based methods, Semantic Lexicons.**

I. Introduction

Sentiment analysis has become a major research topic for Computational Linguistics recently. This field of research is very new and there is no well established terminology. In various works it is referred as: Opinion Extraction, Opinion Mining, Sentiment Mining, Subjectivity Analysis. In [14] is provided a short review of terms and expressions used by different authors. They claim that phrases "sentiment analysis" and "opinion mining" were used parallels and denote the same field of study which itself can be considered a sub-area of subjectivity analysis. A system that summarized opinions from the reviews for a local service such as a restaurant or hotel described in [3] was called Sentiment Summarizer. One of the popular tasks is sentiment classification of movie reviews [15] where document classification methods were applied for detection of review' positive or negative polarity. Speaking more general most of such kind of systems extract opinions about certain topic and detect the sentiments of these opinions [9]. Such kinds of systems were proven extremely useful in politics for summarizing opinions of the voters [11]. Automate opinion mining and sentiment analysis were widely used in sociological analyses [12].

Such a wide range of applications obviously presuppose different problem definitions, various levels and techniques of analysis. The simplest task is the definition of positive, negative or neutral opinion or attitude [6]. In many cases two polarities of sentiments are definitely not enough. In SemEval 2007 Affective Text Task [18] six emotional labels (anger, fear, joy, sadness, surprise, disgust) and their intensity were used for classification task. In [1] opinions are divided in four top-level categories: reporting, advice, judgment and sentiment.

Another approach was used for news annotation in SemEval 2007 Affective Text Task [18]. 1000 newspaper titles were manually annotated with one of six emotional labels (anger, fear, joy, sadness, surprise, disgust) and their intensity. Annotators was guided mostly by personal feelings and as it was pointed in [2] sentences like "Scientists proved that men's perspiration raises women's hormone levels" or "100 killed in bomb attack" had been marked as negative. It was absolutely normal that facts worded in text raised emotions in readers. Keeping in mind that the annotated sentences were newspaper titles their purpose was to attract attention and evoke emotions. They mentioned that most newspapers want to give an impression of objectivity and avoid to express attitudes to given topic directly, however always managing to convey their opinion about the topic to the readers by highlighting some facts while possibly omitting others for example or by the choice of words.

[2] discussed the importance of clear separation of the good and bad news content from the good and bad sentiment expressed by the author. The paper described experiments on annotation of quotes (reported speech) from newspaper articles. After the first annotation experiment the authors noticed that people react to both facts and attitude on facts presented in text. It was pointed that "in the case of newspapers, it is mandatory to distinguish between three different "components": the *author*, the *reader* and the *text* itself". These three components of discourse were described in details in the theory of discourse [7]. In order to increase the comparatively low inter-annotator agreement of the first experiment authors elaborated detailed annotation guidelines, asked annotators to re-annotate the same set of quotations and obtained much better agreement of annotations.

II. Methods of Analysis: Machine-Learning Methods

Supervised and semi-supervised machine learning techniques are the most widely used methods for subjectivity analysis. Opinionfinder system [21] used Naive Bayes classifier that distinguished between subjective and objective sentences using a variety of lexical and contextual features. It is interesting that this classifier was trained using subjective and objective sentences, which were obtained from a large corpus of unannotated data by rule-based classifiers [16]. In [17] five distinct classifier algorithms were coupled using simple voting scheme in order to achieve reliable accuracy. [18] compare several knowledge-based and corpus-based approaches aiming to find the best one. The authors used WordNet-Affect [20] in knowledge-based method and Naïve Bayes and LSA classifiers in corpus-based approaches. In general, sentiment lexicons are the most used lexical resource in opinion classification tasks. There are several well known lexical affective resources such as SentiWordNet [19], WordNet-Affect mentioned above, MicroWNOp [5]. Many researchers used lists of affective words or collocations created ad-hoc [8]. In [10] is described use of mechanical Turk for semantic lexicon creation.

III. DATA

For the experiments, we used a data set of Twitter messages introduced in [4].

Our goal was to assign every extracted tweet with a sentiment label. To ensure better quality of the sentiment labels, we asked that each tweet will be labeled independently by three annotators. The data annotation was a practical work for a graduate course "Semantic Interpretation of Text" which pre-requisites include Computational Linguistics and Natural Language Processing courses. Ten annotators were selected through a rigorous process. Those annotators manually assigned tweets with positive, negative and neutral labels [4]. In 53% of cases the three raters selected the same label, in 34% cases two of three labeled the tweet with the same sentiment, and 13% of tweets were uncertain, i.e. the three raters chose different sentiments. It should be mentioned that tweets with PHI were more difficult for annotators: only 48% of these tweets were labelled with the same category and 15% of tweets with PHI were uncertain. It may be attributed to irony and humor that some authors used in description of their health problems (e.g., Boy I sure had fun at the dentist today Psyche, Have developed an allergy to fried okra and Arbys chicken Joy). As such, humor and irony are difficult for sentiment classification. In Twitter, this difficulty is exemplified by shortness of messages. For instance, sentiment labeling of Headache good night was problematic for the annotators. Table 3 shows a few annotated tweets.

We applied Fleiss Kappa to evaluate inter-rater agreement. Fleiss Kappa is used to assess annotator agreement on two categories:

$$\kappa = \frac{\bar{P} - \bar{P}_{sent}}{1 - \bar{P}_{sent}} \tag{1}$$

where an average agreement per sentiment category:

$$\bar{P}_{sent} = \sum_{j=1}^{N} p_j^2 \tag{2}$$

and an average agreement per tweet:

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{3}$$

and

$$p_i = \frac{1}{n(n-1)}\left(\sum_{j=1}^{3} n_{ij}^2 - n\right) \tag{4}$$

$$p_j = \frac{1}{N \cdot n}\sum_{i=1}^{N} n_{ij} \tag{5}$$

Where

p_i evaluates raters' agreement on the i_{th} tweet;

p_j shows the ratio of all tweets assigned into the j_{th} sentiment category.

n_{ij} is how many raters assigned the i_{th} tweet into the j_{th} sentiment category;

n is the number of raters;

N is total number of tweets.

On our data, the annotators achieved the average agreement equal to $0 .56$ that indicated *moderate agreement* [13].

For machine learning experiments, we used tweets where at least two raters agreed on the sentiment label. The tweets for which all three raters selected different labels were considered ambiguous and discarded from future use.

IV. THE EXPERIMENTS

Finally, 1169 annotated tweets were used in the experiments. The main goal of our experiment was to compare rule-based (SentiStrength) and statistical (WEKA) approaches (also comparing with inter-annotator agreement).

SentiStrength[1] is a sentiment analysis (opinion mining) program. The heart of SentiStrength is a lexicon of 2310 sentiment words. The lexicon is used in a simple way. When SentiStrength reads a text, it splits it into words and separates out emoticons and punctuation. Each word is then checked against the lexicon for matching any of the sentiment terms. The overall classification for this text is the maximum positive and negative strength of each sentiment, which is 3 and -4. In addition to the lexicon, SentiStrength includes a list of emoticons together with human-assigned sentiment scores.

Weka[2] is a collection of machine learning (ML) algorithms for data mining tasks. It includes such algorithms as Naïve Bayes, several modifications; several modifications of decision trees and decision lists, support vector machine algorithm, lazy learning algorithms such as KNN and IBk and several others.

In Machine Learning (ML), texts are represented through their most essential characteristics (i.e., features). Such features can be found through statistical analysis of the data, often referred to as feature selection. In this study, tweets are short texts, with a high variety of lexical units and shortenings. Semantic feature construction in such texts can be challenging. Instead, we opted for a statistical approach to select features for ML experiments. The most frequent representation of a document in ML experiments is so called "bag of words" when all words in document are considered features. For more

[1] http://sentistrength.wlv.ac.uk/
[2] http://www.cs.waikato.ac.nz/ml/weka/

compact representation the sparse words are excluded from the list. The most frequent words such as prepositions, articles, conjunctions and pronouns are considered “noice” words which do not convey any important information about text and usually are removed from the list of features as well. We, however were not convinced that they do not help in classification, thus we decide to verify their influence. Initially, we used three benchmark representations:

- Bag of Words 3 This representation excludes words that appear less than three times in the data. Usually, such words have the highest proportion of noise (spelling errors, mistypes, uncommon slang).
- Bag of Words 5 bag of words occurred at least 5 times in the data set; this representation identifies words common in the data set, i.e. appearing in several data units.
- Bag of Words 10 bag of words occurred at least 10 times in the data set; this representation identifies the most sailent words in the data set.

These three sets were used in two modifications: with the most frequent words and without.

The last representation included features created on the base of SentiStrength lexicon. We used only words which appeared both in our corpus of tweets and in SentiStrength lexicon.

After experimenting with various ML algorithms we selected several with the best accuracy for our task: modifications of Naïve Bayes (NB) and Support Vector Machine (SVM). We used 10-fold cross-validation for the selection of the most accurate classifier. Table 1 presents F-scores for all used representations and all algorithms. F-scores are calculated as follows:

$$Fscore = \frac{2tp}{2tp + fn + fp} \tag{6}$$

where

tp- correctly recognized positive examples,
tn- correctly recognizednegative examples,
fp- negative examples recognized as positives,
fn- positive examples recognized as negatives.

TABEL I. RESULTS OF THE EXPERIMENTS WITH THE BASIC FEATURE SETS AND THE FOLLOWING ALGORITHMS: NB – NAIVE BAYES; DMNB DISCRIMINATIVE MULTINOMIAL NAIVE BAYES; NBMULTUNIMIAL – MULTINOMIAL NAIVE BAYES CLASSIFIER; SVM – SUPPORT VECTOR MACHINE CLASSIFIER.

algorithm features	**N of features**	**NB**	**DMNB text**	**NBMulti nomial**	**SVM**
bow10 without stop-words	139	0.546	0.562	0.555	0.568
bow10	261	0.544	0.564	0.564	0.552
bow5 without stop-words	347	0.580	0.582	0.577	0.582
bow5	504	0.568	**0.599**	<u>**0.6**</u>	0.567
bow3 without stop-words	708	0.557	0.590	0.576	0.562
bow3	884	0.576	**0.595**	0.589	0.545
SentiStrength	568	0.582	<u>**0.617**</u>	**0.603**	**0.616**

In general, lexicon-based features are better than frequent words (0.617 vs. 0.6) for all tweets.

- Among the features collected from the frequent words the bag of words occurred in texts at least 5 times gives the best result in all three cases (all, PHI and non-PHI tweets) although stop-words were helpful only in the case of non-PHI set.
- In general, we cannot say definitely that stop-words are not helpful in classification.
- It should be noted that SVM performed better on the sets without stop words in all cases.
- Among the lexicon-based features SentiStrength feature set gives the best result in all experiments.
- Among the utilized machine-learning algorithms the best was DMNB algorithm showing the best results in three of six experiments. In general, Bayesian classifier was the best, although in one case SVM outperformed it.

The next set of experiments included analysis of negations in text. Negation in text can change the sentiment from one polarity to completely opposite. For example, “I was so happy about it” and “I was not happy about it”. Thus, we analysed each tweet and added the feature which indicated absence or presence of negation words in text. If the negation appeared near the word that presented a feature it was connected to the negation and formed a new feature. The results are presented in Tab 2.

TABEL II. RESULTS OF THE EXPERIMENTS WITH THE BASIC FEATURE SETS TAKING INTO CONSIDERATION NEGATIONS.

algorithm features	**N of features**	**NB**	**DMNB text**	**NBMulti nomial**	**SVM**
bow10 without stop-words	192	0.543	0.557	0.557	0.561
bow10	358	0.546	0.563	0.562	0.543
bow5 without stop-words	431	0.563	0.58	0.579	0.576
bow5	632	0.572	**0.603**	0.594	0.573
bow3 without stop-words	812	0.549	0.585	0.573	0.553
bow3	1030	0.569	0.6	0.582	0.546
SentiStrength	610	0.572	**0.614**	0.601	**0.619**

It is seen from the table that the the sets of features became larger and the results are better but only slightly. It indicates thst the number of negations is small and these words have not much influence on tweet sentiment polarity.

The last experiment used SentiStrenght for sentiment categorization. SentiStrenght is rule – based system and cannot be tested using 10-fold cross-validation as it does not need training. Thus, the result of SentiStrenght classification is presented in percents of correctly assigned labels.

SentiStrenght gives the possibility to train the initial lexicon with initial sentiment labels using the annotated data. Thus, we experimented with the initial SentiStrenght lexicon and with the same lexicon after the trainig. The next two experiments were made with the words common between SentiStrenght lexicon and our set of tweets. It also contained two stages:

without and with trainig. The results of these experiments are presented in the tab. 3.

TABEL III. SRESULTS OF THE EXPERIMENTS WITH THE BASIC FEATURE SETS TAKING INTO CONSIDERATION NEGATIONS.

	N of features	**all tweets 1169**
SentiStrength initial lexicon	691	61%
SentiStrength initial lexicon optimized with all		64%
SentiStrength common words	568	49%
SentiStrength common words optimized with all		58%

V. CONCLUSIONS

Comparison between ML and rule-based algorithms demnstrated that well-tuned rule-based methods obtain better results than general ML methods and it is necessary to use various types of features for obtaining satisfactory accurasy usng ML algoritms.

REFERENCES

[1] Nicholas Asher, Farah Benamara, Yvette Yannick Y. Mathieu. 2009. Appraisal of opinion expressions in discourse, Lingvisticae Investigationes, Vol. 32, No. 2., pp. 279-292.

[2] Balahur Alexandra, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen & Mijail Kabadjov (2009). Opinion Mining on Newspaper Quotations. Proceedings of 'IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology', pp. 523-526.

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

[4] Bobicev, V., M. Sokolova, Y. Jafer, D. Schramm. Learning Sentiments from Tweets with Personal Health Information, Proceedings of Canadian AI 2012, Springer, 2012.

[5] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming), chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

[6] Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. Processing, 1-6

[7] Barbara J. Grosz, Candace L. Sidner. Attention, intentions, and the structure of discourse. Journal Computational Linguistics, Vol. 12 Issue 3, 1986, pp. 175-204.

[8] Gregory Grefenstette, Yan Qu, James G. Shanahan, David A. Evans Coupling niche browsers and affect analysis for an opinion mining application In Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO , 2004)

[9] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.

[10] Saif M. Mohammad, Peter D. Turney Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, LA, California.

[11] Tony Mullen and Robert Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAICAAW), pages 159–162, 2006

[12] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010.

[13] Osman, D., J. Yearwood, P. Vamplew. Automated opinion detection: Implications of the level of agreement between human raters. Information Processing and Management, 46, 331–342, 2010

[14] Bo Pang and Lillian Lee Opinion mining and sentiment analysis Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135

[15] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002, pp.79-86.

[16] E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In EMNLP-2003.

[17] Sanjiv Ranjan Das, Mike Chen, (2007), "Yahoo for Amazon! Sentiment Extraction from Small Talk on the Web," Management Science, v53, 1375-1388.

[18] Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text, Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007).

[19] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2200-2204. ELRA.

[20] Strapparava, C., Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. 4th International Conference on Language Resources and Evaluation, 1083–1086

[21] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 347-354. Association for Computational Linguistics Morristown, NJ, USA, 2005.