IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

https://doi.org/10.15623/ijret.2018.0707020 Received: 09-05-2018, Accepted: 20-06-2018, Published: 11-07-2018

SENTIMENT ANALYSIS, A SUPPORT VECTOR MACHINE MODEL BASED ON SOCIAL NETWORK DATA

Alessandro Mario Muscolino[1], Salvatore Pagano[2]

[1]Indipendent researcher, Università degli studi di Catania
[2]Indipendent researcher, Università degli studi di Catania

Abstract

In this paper, we analyze the efficacy linguistic features in social media review to detecting the sentiment of any message. To calculate the sentiment polarity, we use a machine-learning method that applies text-categorization techniques to word vector of text in the document. We estimate the advantage of not discarding expressions of the informal and creative language used in micro blogging. The new features in this paper are the pre-classified data source which increase the calculation accuracy.

Keywords: - *sentiment analysis, social, big data, machine learning*

----------***----------

1. INTRODUCTION

In the last few years a digital revolution is taking place which is having a radical impact on today's communication and society.With the emergence of the World Wide Web and the exponential growth of Web 2.0 sites and applications that make communication and collaboration between people easier, the way of behaving and relating in one's social life and more generally, in society has changed drastically.

People express their opinions, points of view, feelings on social networks, on blogs, through reviews and find it more convenient to ask and give advice and opinions online on all topics: politics, medicine and health, news, gossip, products and events art. This huge amount of information has offered the opportunity to develop theories and technologies for automatic processing of natural language on a quantity of data never seen before.

The developed system derives from a supervised learning process based on a Support Vector Machine (SVM) classifier. The algorithm developed extracts, for each review, a vector of syntactic features, semantics and polarity that, appropriately administered to an SVM classifier, make it possible to construct a general model for sentiment analysis.

2. AUTOMATIC LEARNING APPROACH

The approaches based on machine learning rely on artificial intelligence algorithms to solve the problems of sentiment analysis. Generally, these algorithms take a set of examples (training sets) as input and output a general model for the classification[1].

From any text, the extraction of the features is the process of extrapolation of its characteristics and salient properties [2]. These properties must represent the basic characteristics of the text but their extraction isn't immediate: at the same time, they should discriminate and describe the original text as much as possible and reduce the large size of the source data and avoid redundancy[3]. The most used features in the literature are[4]:

Words, Identification of unigrams, bigrammes, n-grams of words in the document;

Parts of speech, adjectives, adverbs, nouns, verbs usually recognized through POS-tagging. The parts of the discourse are used to partially disambiguate the sense of the terms and for identifying adjectives and adverbs which are usually excellent indicators of the semantic orientation;

Syntax, recognition of syntactic combinations, usually obtained through parsers and syntactic structures in dependencies.[5] Some studies have shown that algorithms with syntactic features and algorithms with n-gram-based features provide similar results[6];

Opinion word, recognition of words that, by themselves, express a clear opinion;

Denial, negations that usually reverse opinions expressed;

Once the features have been extracted it is necessary to calculate their "weight" within the document [7]. One approach is based on presence: "0" if the feature does not appear and "1" if the feature appears in the document. Other approaches, usually used for word features, are those based on term frequency and inverse document frequency [5][8]. In general, in the information retrieval and in the classification of texts, it is preferable to weigh these features using the term frequency in order to obtain better results [9]. It has been shown that in sentiment analysis it is preferable to assign a value to the features based on presence / absence rather than on the term frequency.

3. THE SVM CLASSIFIER

The system realizes a classification based on a Support Vector Machines algorithm.

https://doi.org/10.15623/ijret.2018.0707020 Received: 09-05-2018, Accepted: 20-06-2018, Published: 11-07-2018

The planning and development work of the system has provided for the following activities [10]:

1) Definition of the characteristics to be extracted from the reviews.
2) Implementation of an automatic rating classification algorithm.
   - Pre-processing of review;
   - syntactic analysis;
   - construction of a tree of syntactic addictions;
   - lexical labeling and polarity calculation of each word of the review;
   - extraction of lexical, morphosyntactic and dependent features fromlexicon;
   - training;
   - test.

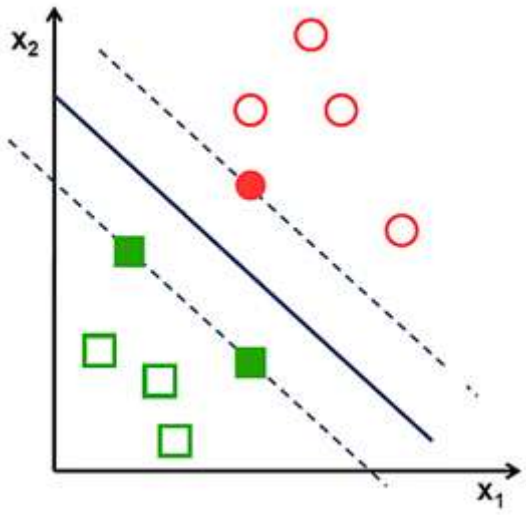

Fig 1: Linear SVM with support vectors and maximum distance between two classes.

To allow machine learning and the construction of a general model for classification, each review must be transformed into a vector of meaningful features to be supplied to the SVM algorithm as classes [11][12]:

- lexical features;
- morphosyntactic features;
- features dependent on the lexicon.

Lexical Features

Average length: number of blocks of at most 5 words in the review [AVG_LENGTH]
N-grammes of words: presence or absence of a sequence of consecutive words in the review [NG_TOKEN]
N-grammes of terms: presence or absence of a sequence of consecutive terms in the review [NG_LEMMA]
Punctuation: presence or absence of a "?" or a "!" at the end of the review [FINISH _PUNCT]
Emoticon: presence or absence of one or more positive emoticons [S_EMO_P] or negative [S_EMO_N] in the review [13]

Morphosyntactic Features

N-grams of parts of the speech: presence or absence of a sequence of consecutive parts of the speech in the review, corresponding to the main syntactic categories (noun, verb, adjective, adverb) [NG_POS]

Distribution of parts of the speech: percentage of distribution of names [P_DISTR_NOUN], adjectives [P_DISTR_ADJ], adverbs [P_DISTR_ADV] and numbers [P_DISTR_NUM] within the review.

Feature Dependent on the Lexicon

N-grams of opinion words: for each n-gram of consecutive lemma present in the review, the polarity of each of them is checked by extracting it, where present, from the lexicon. [NG_S].

Intensifiers: presence or absence of words that increase or decrease the polarity score of other words in the review. The information is obtained, where present, from the lexicon. [HAS_INTENS].

Negators: presence or absence of words that reverse the polarity score of one or more words in the review. The information is obtained, where present, from the lexicon [HAS_N].

Polarity modifiers: for every lemma of the tweet the presence of intensifiers or negators occurs, through the tree of syntactic dependencies, which change the value of polarity. In this case, the value of the feature was obtained by concatenating the intensifier / negator and the polarity of the lemma [14]. [S_WITH_MOD].

PMI: for each unigram, bigram, trigram and quadrigram of words in the review, a score is obtained, obtained by summing the polarities of the single terms present in the n-gram. For each n-gram we consider only the minimum value and the approximate maximum value to the nearest whole [PMI].

Polarity distribution: this feature calculates the percentage of positive words [S_DISTR_P] and negative [S_DISTR_N] in the review. Each value is rounded to the nearest multiple of 5.

More frequent polarity: this feature indicates the most frequent polarity between the lemmas present in the review, positive [S_MAJ_P] or negative [S_MAJ_N].

More frequent polarity in sections: divided the review into three equal parts, this feature indicates the most frequent polarity in each of the three parts using the polarity of the lemmas present in the lexicon. [S_POS_PRES].

Score by lexicon: is the score obtained by summing the individual a priori polarities indicated in the lexicon of all the words present in the review. [SCORE].

https://doi.org/10.15623/ijret.2018.0707020 Received: 09-05-2018, Accepted: 20-06-2018, Published: 11-07-2018

4. RESULTS

At the end of the classification process, for each class it's possible to group the texts as follows:
TP (true positive, true positive) number of documents/texts correctly inserted in the class;
FP (false positive, false positive) number of documents/texts incorrectly entered in the class;
FN (false negative, false negative) number of documents/texts incorrectly not inserted under class c;
TN (true negative, true negatives) number of documents/texts correctly not included in class c.

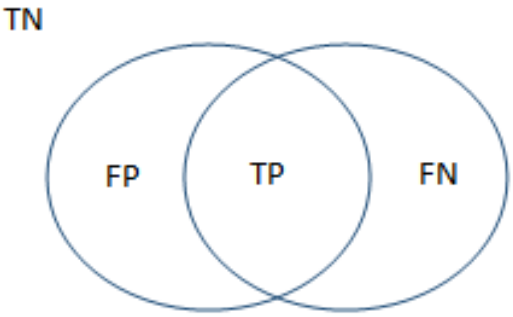

Fig 2: Diagram for the definition of evaluation metrics

These metrics are calculated as follows:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$score = \frac{2 * precision * recall}{precision + recall}$$

Each of them allows to measure different aspects of the classification [15].
Accuracy: measure the percentage of correctly classified documents / texts.
Precision: measure the correctness of the classifier in terms of percentage of documents / texts correctly labeled in a certain class compared to the total number of documents labeled in that class.
Recall: measures the completeness of the classifier in terms of the percentage of documents / texts correctly labeled in a class with respect to the total number of documents that should have been classed in that class.
Score: it is the geometric mean between precision and recall and allows to have a good measure of correctness and total completeness of the classifier [16].

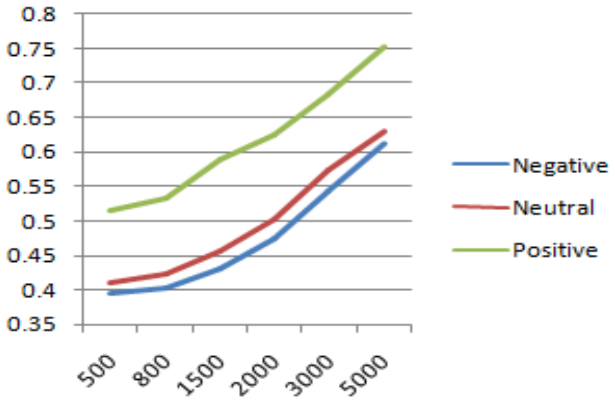

Chart -1: Trend of metrics in relation to the number of review used for training set

The chart shows the trend of the relative metrics in relation to the number of reviews used for training set.

5. CONCLUSION

Based on the usually subjective nature of classification problems, the evaluation of the text class performance was carried out experimentally, measuring its ability to make correct decisions during the grading process.

Our experiments on the sentiment analysis on social media review show that the characteristics of a part of the speech can be useful for the analysis of sentiment in the microblogging domain.

Despite the more than good results obtained, the classifier described in this document can be greatly improved by refining and refining the tools used, enriching them with other techniques and new approaches.

Further improvements of the results of the classification can be obtained by preceding the sentiment polarity classes with a subjectivity classification process in order to identify texts that present only facts from those that express sentiments and opinions. it is also possible to consider adding some irony detection techniques in order to recognize and classify the ironic tweets or those using metaphors more accurately.

A very interesting experiment is the realization of a mixed systemto carry out a polarity classification of texts by exploiting and combining the results of both the lexicon-based approach and the one based on stochastic / statistical automatic learning.

REFERENCES

[1] Russell, S., & Norvig, P. (2005). AI a modern approach. Learning, 2(3), 4.
[2] Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. IEEE Journal of Selected topics in applied earth observations and remote sensing, 7(6), 2094-2107.

https://doi.org/10.15623/ijret.2018.0707020 Received: 09-05-2018, Accepted: 20-06-2018, Published: 11-07-2018

[3] Gamon, M. (2004, August). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of the 20th international conference on Computational Linguistics (p. 841). Association for Computational Linguistics.

[4] Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 35(3), 399-433.

[5] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[6] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[7] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.

[8] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[10] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. Neural computation, 13(3), 637-649.

[11] Keerthi, S. S., & Gilbert, E. G. (2002). Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning, 46(1-3), 351-360.

[12] Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 12(5), 526-558.

[13] Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL student research workshop (pp. 43-48). Association for Computational Linguistics.

[14] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[15] Khan, A. Z., Atique, M., & Thakare, V. M. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89.

[16] Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508-524). Springer, Berlin, Heidelberg.