UDC 004.62

Methods and Techniques of Assessment of the Value Orientations of Social Media Users

K. O. Kiforchuk[1]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Institute of Physics and Technology*

Abstract

The purpose of the work is to investigate and review the general concept of sentiment analysis and the use of such approach in assessment of the value orientations of social media users. The article analyzes main components and steps of sentiment analysis and different methods and techniques which are applied at each stage. The paper also researches machine learning approach in sentiment analysis: different algorithms were reviewed in details. Authors use methods of analysis for research of technologies and means of sentiment analysis, its functions, opportunities and advantages of use; comparison methods for researching individual techniques and methods. The article states that there are many options for further improvements in sentiment analysis and authors propose an original approach for determining emotional component in text.

Keywords: sentiment analysis, social media analysis, machine learning

Introduction

Despite the rationalism of the modern world and the development of technology, a person is still an emotional being. Emotions rule the world: most human decisions and actions cannot do without the influence of an emotional factor. Emotions influence the mood of peoples, the relations between cultures and the behaviour of nations. Neither political leaders, nor historians, nor ordinary interested people can afford to ignore them.

In the modern world, social networks have become one of the most popular way of self-representation. Many people open up and trust on social networks much more readily than in real life.

Often, what you write on social networks does not go beyond the Internet. But there are exceptions when the expression of emotions in the virtual world affects our reality. One of the most illustrative examples can be the expression of discontent on social networks, which develops into protests, strikes or even civil wars. In addition, social media is also a good tool for rallying people who have a common emotional reaction to a certain incident [1].

Because of its importance, the emotion's role in social media has been the subject of considerable research and media attention. But it can only be definitely said whether the emotion is positive or negative: reliable determination of a specific emotion in social media content remains an unresolved issue. While debate continues about the emotional impacts of browsing social media in the course of day-to-day life, researchers have focused only on a limited set of emotions, rather than investigating the range of human emotion [2].

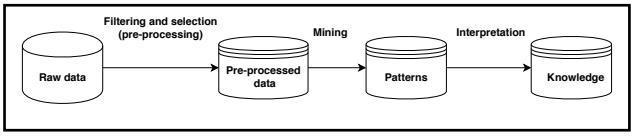

Fig. 1. Sentiment analysis main stages

1. Sentiment analysis

Most often, the study of the emotional tonality of a text in social media is done using a technique called «sentiment analysis» in order to infer emotion from the text, whether positive, negative or something more specific like sadness. Sentiment analysis is a field of study that analyses people's opinions, evaluations, attitudes, and emotions generally from written language. The texts from reviews are processed to get an accurate description of how the writer feels regarding the subject. Sentiment analysis is one of the most active research areas in natural language processing, web/social network mining, and text/multimedia data mining. The growing importance of sentiment analysis coincides with the popularity of social network platforms, such as Facebook, Twitter, and Flickr.

As shown in Figure 1 the analysis process involves several stages: data collection, data pre-processing and mining – that is what actually called «sentiment analysis».

There are several approaches to obtaining data from social media. For example, used existing open data sets obtained earlier by other researchers were used in [3]. The advantages of this approach are the absence of time losses and the ability to collect large data amounts. However, there is no guarantee of data quality. By contrast, there is a second way of gathering data – you can get information on your own and therefore be confident in the quality and veracity of the collected

data. This can be done with or without the Application Programming Interface (API) of the social platform being researched. Software for gathering social media information that do not use API provides a very flexible and high performance contents retrieval. Its gathering activities are not limited to the standard Web protocols and technologies, but also operate with other type of sources like remote databases, File Transfer Protocol (FTP), Network News Transfer Protocol (NNTP), mailboxes, file systems and other proprietary source protocols [4]. The main disadvantage of this approach is a lot of development time. In contrast to the previous method, Twitter API were used in [5] for information acquisition thereby reducing the time of the research but imposing restrictions on the ability to collect data.

As for pre-processing stage, there are many ways to filter out unnecessary or incomplete information. According to [5] the following techniques may be used: Unigrams, Unigrams except Stop Words, Bigrams, Bigrams except Stop Words, Most Informative Unigrams and Bigrams. While in [6] the following techniques were used: Word parsing and tokenization, Stop Words and Stemming. The completely another way was introduced in [7] – an emoticon dictionary and an acronym dictionary were firstly used for pre-processing messages from social media.

There is also a lot of diversity at the mining stage. Some researchers use machine learning for sentiment analysis, which requires training a machine learning algorithm on the specific corpus to be analyzed [8]. More typically in the academic literature, however, researchers use a «dictionary» method, which essentially involves counting up the number of words thought to signify a particular emotion and dividing by the total number of words in a given text to derive an estimate for that emotion. Dictionaries of these emotion words are not honed to the particularities of a corpus but rather are pre-specified by an expert or crowdsourced judging process, which means they generally underperform machine learning methods [8]. The most popular dictionary method is likely Linguistic Inquiry and Word Count (LIWC), which refers to itself as «the gold standard in computerized text analysis». Updated last in 2015, LIWC's dictionary contains 620 words thought to signify positive emotion and 744 words thought to signify negative emotion. LIWC's negative emotion word category comprises three specific word categories for anger, anxiety and sadness [9].

2. Machine learning approach in sentiment analysis

Using machine learning algorithms and techniques in sentiment analysis is the most scalable approach: with its use it is possible to process hundreds of thousands of social media messages.

There are a number of algorithms that are used to train machines to perform sentiment analysis, but all of them have similar processes. There are two main stages of machine learning algorithm: training and prediction. Figure 2 represents the process of model training.

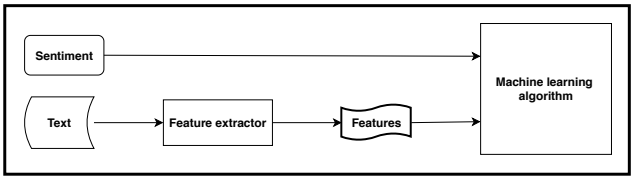

Fig. 2. ML model training process

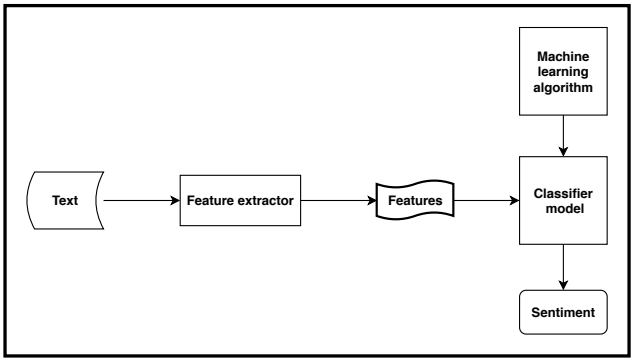

Fig. 3. Process of ML model prediction

As can be seen from Figure 3, prediction is an inverse process to the training: given a machine learning algorithm and its classifier model, appropriate sentiment of the text is returned.

Each of the machine learning algorithms has its own advantages and disadvantages. However, combination of different methods can provide unique and outstanding results. In the next subsections most used algorithms will be explained.

2.1. Naive Bayes

Naive Bayes is a group of probabilistic algorithms that, for sentiment analysis classification, assigns a probability that a given word or phrase should be considered positive or negative.

Naive Bayes machine learning algorithm is based on Bayes theorem, which is described by the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Essentially, this states «the probability of A given that B is true equals the probability of B given that A is true times the probability of A being true, divided by the probability of B being true».

Denote $D = \langle d_i \rangle, i = 1, 2, \dots n$, represents document, where d_i is corresponding to a letter, a word, or other attribute of text, and a set of $C = \{c_1, c_2, \dots c_k\}$ is predefined classes. Text classification is to assign a class label $c_j, j = 1, 2, \dots k$ from C to a document.

By applying Bayes theorem to sentiment analysis we get the following interpretation:

$$P(c_j|D) = \frac{P(D|c_j)P(c_j)}{P(D)} \tag{2}$$

Where $P(c_j)$ is prior information of the appearing probability of class c_j, $P(D)$ is the information from

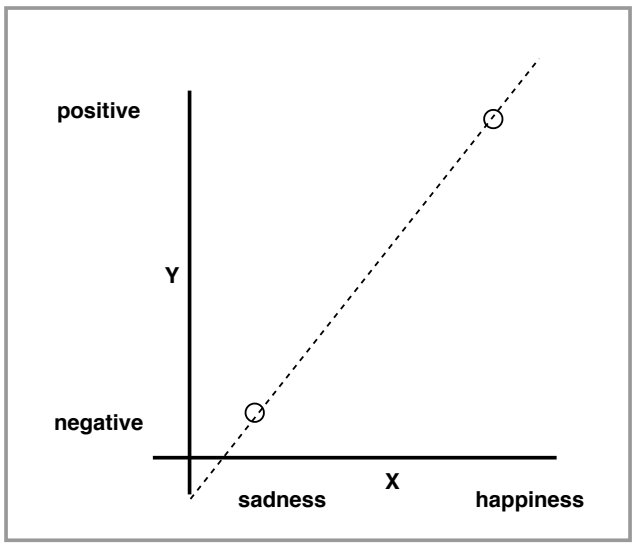

Fig. 4. Linear regression in sentiment analysis

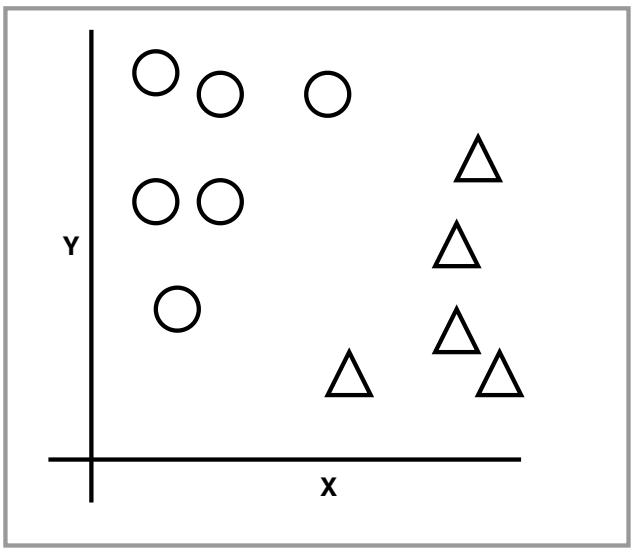

Fig. 5. Distribution of circle and triangle tags

observations, which is the knowledge from the text itself to be classified, and $P(D|c_j)$ is the distribution probability of document D in classes space. Bayes classifier is to integrate these information and compute separately the posteriori of document D falling into each class c_j, and assign the document to the class with the highest probability [10].

Basically, Naive Bayes calculates words against each other. With machine learning models trained for word polarity, it is possible to calculate the likelihood that a word, phrase, or text is positive or negative.

When techniques like lemmatization, stopword removal and term frequency-inverse document frequency (TF-IDF) are implemented, Naive Bayes becomes more and more predictively accurate.

2.2. Linear regression

Linear regression is a statistical algorithm used to predict a Y value, given X features. It attempts to find the linear relationship between X and Y. Variable Y is considered to be a dependent on an explanatory variable X.

In terms of sentiment analysis input variable X represents words and phrases while output variable Y shows its polarity: whether word or phrase has «positive» or «negative» meaning.

Linear regression fits a correlation between inputs and outputs as linear equation. Figure 4 shows how linear relationship can be used for determining sentiment polarity.

2.3. Support vector machines (SVM)

A support vector machine is another supervised machine learning model, similar to linear regression but more advanced. SVM uses algorithms to train and classify text within sentiment polarity model, taking it a step beyond X/Y prediction.

Let there be two tags: circle and triangle, with two data features: X and Y. The output of SVM classifier

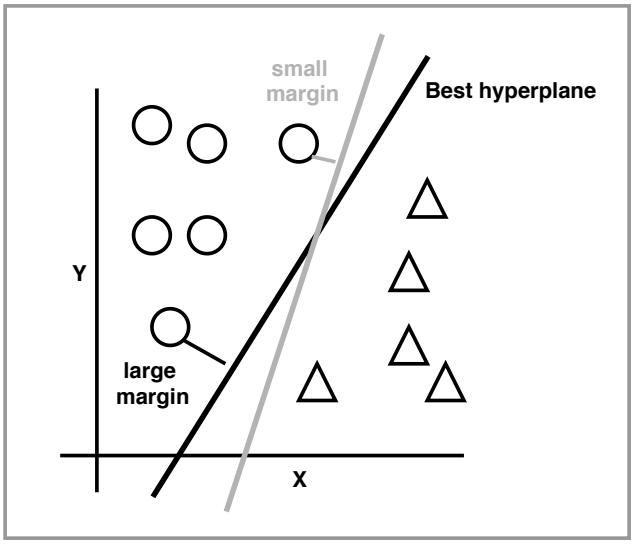

Fig. 6. SVM hyperplane

is an X/Y coordinate that is either circle or triangle. Let there be distribution of tags as shown in Figure 5

The SVM classifier assigns a hyperplane that best separates the tags. In two dimensions this is simply a line (like in linear regression). Anything on one side of the line is a circle and anything on the other side is a triangle. For sentiment analysis this would be positive and negative.

In order to maximize machine learning, the best hyperplane is the one with the largest distance between each tag. Figure 6 illustrates the difference between best hyperplane and any other.

Using SVM, the more complex the data, the more accurate the predictor will become.

Conclusions

There is a lot of diversity in sentiment analysis: this method of text mining can be done with the use of various natural language processing techniques.

The most competitive and perspective technique is machine learning. However, this study has many op-

tions as well: different algorithms can be used in sentiment analysis. The most common machine learning methods are the following: Naive Bayes, Linear Regression and SVM.

Each of these algorithms has its own pros and cons:

1) Naive Bayes
   - Pros: require less data, simple and fast;
   - Cons: can't work with categories that missing in training data set, works only with independent predictors;
2) Linear regression
   - Pros: easy to implement;
   - Cons: assumes linear relationship between dependent and independent variables, sensitive to outliners;
3) SVM
   - Pros: effective in high dimensional spaces, memory efficient;
   - Cons: bad performance with large sets and noisy data, directly provide probability estimates;

From the foregoing it is clear that each part of sentiment analysis has many options for research. Combinations of different approaches at each stage can give brand new results. Even existing methods have a margin for improvement.

One of the possible improvements in the quality of determining the emotional component of a text may be the use of associative chains for each emotion. With this approach, for each sentiment, a sequence of words is identified that is associated with a given emotion. This technique allows you to find hidden emotional subtexts. In addition, it becomes possible to identify several emotions in one text. Additional research is needed to compare associative chains approach with the existing techniques.

References

[1] J. T. Jost *et al.*, "How social media facilitates political protest: Information, motivation, and social networks," *Political Psychology*, vol. 39, pp. 85–118, 2018.

[2] G. T. Panger, *Emotion in Social Media*. PhD thesis, University of California, Berkeley, 2017.

[3] B. Naiknaware, B. Kushwaha, and S. Kawathekar, "Social media sentiment analysis using machine learning classifiers," *International Journal of Computer Science and Mobile Computing*, vol. 6, pp. 465–472, 6 2017.

[4] F. Neri, C. Aliprandi, F. Capeci, M. Montserrat, and T. By, "Sentiment analysis on social media," in *Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (Istanbul), IEEE/ACM, IEEE, 2012.

[5] S. Jayasanka, T. Madhushani, E. Marcus, A. Aberathne, and S. Premaratne, "Sentiment analysis for social media," in *Proceedings of Information Technology Research Symposium*, vol. 4, (Moratuwa), University of Moratuwa, 2013.

[6] R. Singh and R. Kaur, "Sentiment analysis on social media and online review," *International Journal of Computer Applications*, vol. 20, no. 121, pp. 44–48, 2015.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, pp. 30–38, 2011.

[8] A. Reagan, B. Tivnan, J. Williams, C. Danforth, and P. Dodds, "Benchmarking sentiment analysis methods for large-scale texts: a case for using continuum-scored words and word shift graphs." 2015.

[9] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," tech. rep., University of Texas at Austin, 9 2015.

[10] W. Zhang and F. Gao, "An improvement to naive bayes for text classification," *Procedia Engineering*, vol. 15, pp. 2160–2164, 2011.