Sentiment Analysis of Social Network Posts in Slovak

Rastislav KRCHŇAVÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
Rastislav.krchnavy@gmail.com

Sentiment analysis is part of Natural Language Processing problem. In sentiment analysis text is classified into categories, which describes several states of sentiment. Our data for training and validation are Fecebook posts from pages managed by Seesame (Slovak PR agency collaborating with us on this work). The texts from dataset are written in Slovak and many of them have grammar mistakes, missing diacritics and contain a lot of emoticons and emojis. In sentiment analysis the main goal is to determine whether the text is positive or negative, eventually neutral.

There has not been any existing solution developed for Slovak language yet, but for other languages some research has been done. In Czech language, which is similar to Slovak, a machine learning solution achieved 72 % in three class classification [1].

In sentiment analysis there are two main approaches – machine learning and lexicon based approach. In our work we will compare different approaches and we will also try to build a classifier which combines both of them for achieving the best accuracy. In each approach we can analyse text on sentence, document and aspect level.

First step of our solution is pre-processing the text. In the phase of pre-processing the original raw texts are transformed into set of features which comes to training and testing the classifier. The most important steps of pre-processing are extracting the emoticons and emojis, lemmatizing the words, segmentation and deleting stop words[1].

Extraction of emoticons and emojis means, that these figures in text are replaced by their name, because in further processing there are allowed only alphanumerical characters. In phase of lemmatizing, all words are converted to their lemas, which contains information about semantic meaning of words but not about their forms. This step is very important for Slovak language. By removing stop words, we can choose

* Supervisor: Marián Šimko, Institute of Informatics, Information Systems and Software Engineering

[1] Words with no semantic value

Spring 2016 PeWe Workshop, April 2, 2016, pp. 65–66.

between using a list of stop words or removing words based on their grammatical categories.

After pre-processing we will build the classifier. We use a NLTK[2] library, where the machine learning classifiers been implemented yet. In our research we will compare these algorithms:

- Naïve Bayes classifier – machine learning approach
- Maximal Entropy classifier – machine learning approach
- Lexicon based approach based on Slovak Sentiment Lexicon[3]
- Classifier based on combination of machine learning and lexicon based approach

Last step of our experiment is validation. Data in our dataset are divided into five categories (strongly negative, negative, neutral, positive, and strongly positive) and contains over 1200 posts[4]. The main metric we focus in our experiment is accuracy, but we also measure precision and recall for each class.

The best results are about 80 % accuracy in two classes Naïve Bayes classifier and about 81.5 % in combination of Naïve Bayes and lexicon bases classifier. With increasing number of classes Naïve Bayes shows the best preforming classifier with about 59 % in three classes, about 45 % in four classes and about 40 % in five classes.

The exact results will be published in May 2016, but now we can say that in two class classification the results are good, because human raters can agree in 79 % [2].

Extended version was published in Proc. of the 12th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2016), STU Bratislava.

References

[1] Koktan, M.: *Automatické rozpoznávání (analýza) sentimentu*. Diplomová práca, Západočeská univerzita v Plzni, Plzeň, 2012.

[2] Onegva, M.: How Companies Can Use Sentiment Analysis to Improve Their Business. *Mashable*. 2010.

[2] Natural Language Toolkit – http://www.nltk.org/

[3] Available on Web – https://github.com/okruhlica/SlovakSentimentLexicon

[4] At the time of writing this abstract (April 2016)