Automatic Term Extraction for Sentiment Classification of Dynamically Updated Text Collections into Three Classes

Yuliya Rubtsova

The A.P. Ershov Institute of Informatics Systems (IIS),
Siberian Branch of the Russian Academy of Sciences
yu.rubtsova@gmail.com

Abstract. This paper presents an automatic term extraction approach for building a vocabulary that is constantly updated. A prepared dictionary is used for sentiment classification into three classes (positive, neutral, negative). In addition, the results of sentiment classification are described and the accuracy of methods based on various weighting schemes is compared. The paper also demonstrates the computational complexity of generating representations for N dynamic documents depending on the weighting scheme used.

Keywords: Corpus linguistics, sentiment analysis, information extraction, text classification and categorization, social networks data analysis.

1 Introduction

We live in a constantly changing world. Peoples' style and way of life, behaviors and speech are all changing. Natural language is constantly transforming and developing together with conversational speech: new words are included in active vocabulary, while old ones cease to be used. New words are born every day, and about half of them are slang. Slang responds to changes in all spheres more quickly than other types of language and is so important to modern society that last year 40 neologisms, some of which are slang words, were added to the Oxford English Dictionary. Slang is actively used in colloquial speech and written communication on social networking sites, as well as to express an emotional attitude towards a particular issue. Users of social networks are among the first to start using new terms in everyday language. Among about 1000 new words included in the Oxford English Dictionary near 40 were terms that came from social networks, such as "srsly", "me time" and "selfie". Accordingly, it is necessary to consider slang when developing sentiment classifiers, in particular when creating vocabularies of emotional language. Moreover, since active vocabulary is regularly updated with new terms, vocabularies of emotional language should also be updated regularly, and the weights of the terms in these vocabularies must be recalculated.

This paper presents an approach to extracting terms and assigning them weights in order to build a vocabulary of emotional language that is constantly updated.

P. Klinov and D. Mouromtsev (Eds.): KESW 2014, CCIS 468, pp. 140–149, 2014.

There will be a comparison of methods based on various weighting schemes and the computational complexity of recalculating the weights of terms in the vocabulary depending on the methods used will be demonstrated. All experiments to classify texts into three sentiment classes (positive, neutral, negative) were performed on two collections:

- Collection of short posts from microblogs [1];
- News collection.

2 Overview of Term Weighting Schemes

There are different approaches to the extraction of evaluative words from texts and the determination of their weight in the collection. In [2], the authors use a thesaurus to expand a vocabulary of evaluative words that had been collected manually. In corpus linguistics, methods of extracting terms based on measuring the relevance of a term to a collection are widely used, for example, the well-known methods based on the TF-IDF weighting scheme [3]. In [4], the authors show that variants of the classic TF-IDF scheme adapted to sentiment analysis task provide significant increases in accuracy in comparison to binary unigram weights. They tested their approach on a wide selection of data sets and demonstrated that classification accuracy enhanced.

The functioning of most existing methods of automatic and semi-automatic word extraction from texts are based on the assumption that all the data are known in advance, accessible and static. For example, to use a method based on the TF-IDF scheme [3], it is necessary to know the frequency each term occurs in the document, which means that the data set should not be changed during calculation. This greatly complicates computation is required for data calculation in real time. For example, when adding a new text to the collection, it is necessary to recalculate the weights for all terms in the collection. The computational complexity of recalculating all the weights in the collection is $O(N^2)$.

The Term Frequency – Inverse Corpus Frequency (TF-ICF) measure has been proposed [5, 6] in order to solve the problem of searching for terms and calculating their weights in real time. Information on the usage frequency of a term in other documents of the collection is not required in order to calculate TF-ICF, so the computational complexity is linear. The results of methods based on TF-ICF and TF-IDF have been compared [3] in order to evaluate the effectiveness of a method based on the TF-ICF weighting scheme for the task of extracting evaluative terms for a vocabulary of emotional language.

The formula for calculating the TF-IDF measure is as follows:

$$tfidf = tf \times log \frac{T}{T(t_i)} \quad (1)$$

Where tf is the frequency with which the term occurs in the collection (of positive or negative tweets), T is total number of texts in the positive and negative collections, and $T(t_i)$ is the number of texts in the positive and negative collections containing the term.