Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 733 – 743

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew

Yaakov HaCohen-Kerner[a, 1*], Haim Badash[a]

[a]*Dept. of Computer Science, Jerusalem College of Technology, 9116001 Jerusalem, Israel*

Abstract

In this research, given a corpus containing blog posts written in Hebrew and two seed sentiment lists, we analyze the positive and negative sentences included in the corpus, and special groups of words that are associated with the positive and negative seed words. We discovered many new negative words (around half of the top 50 words) but only one positive word. Among the top words that are associated with the positive seed words, we discovered various first-person and third-person pronouns. Intensifiers were found for both the positive and negative seed words. Most of the corpus' sentences are neutral. For the rest, the rate of positive sentences is above 80%. The sentiment scores of the top words that are associated with the positive words are significantly higher than those of the top words that are associated with the negative words.

Our conclusions are as follows. Positive sentences more "refer to" the authors themselves (first-person pronouns and related words) and are also more general, e.g., more related to other people (third-person pronouns), while negative sentences are much more concentrated on negative things and therefore contain many new negative words. Israeli bloggers tend to use intensifiers in order to emphasize or even exaggerate their sentiment opinions (both positive and negative). These bloggers not only write much more positive sentences than negative sentences, but also write much longer positive sentences than negative sentences.

Peer-review under responsibility of KES International

Keywords: Blog corpus; Hebrew; Natural Language Processing; Negative words; Positive words; Seed lists; Sentiment

1. Introduction

The research presented in this paper was performed in the blog domain, which is one of the most popular domains in the Internet. A blog (a truncation of weblog) is a website consisting of informational posts composed by an individual author or a group of authors. The posts are appearing in reverse chronological order (the most recent post appearing first). Blogs typically enable other users to comment or respond to the blog post. Nowadays, there are

[1] Corresponding author. Tel.: +972-2-6751018; fax: +972-2-6751046.
E-mail address: kerner@jct.ac.il

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.257

hundreds of million public blogs in existence. Processing of blog posts presents challenges due to the large number of words present in the text set, their dependencies and the large number of training documents.

The selected application domain is personal blog posts written in Hebrew. We downloaded a corpus containing blog posts written in Hebrew. Given these blog posts, we are interested to answer the following research questions:

Q1(a). Is it possible to learn new positive words using a basic/extended list of positive words?

Q1(b). Is it possible to learn new negative words using a basic/extended list of negative words?

Q2. Can we discover special groups of words that are associated with the list of positive and negative words?

Q3. What is the distribution of the sentences (neutral, positive, and negative)?

Q4. What are the scores of the top words associated with the positive and negative words and what can we learn from these scores?

To answer these questions, we worked with two seed lists containing sentiment words in Hebrew. These lists were manually generated by us. Each one of these lists contains both positive and negative words. The first list is relatively a small list, containing only 45 words (22 positive and 23 negative). The second list, the largest list, contains 168 words (85 positive and 83 negative). Our motivation to perform experiments with two seed sentiment lists (basic and extended) is to check whether there is any different in the results obtained by these two lists. An example for a question is whether the use of the extended seed sentiment list can discover more positive and negative sentiment words than the use of the basic seed sentiment list.

We defined and activated the following algorithm. Given a blog corpus, we spilt it into sentences. For each sentence, we count the number of positive words (PW) and negative words (NW) included in the sentence according to a given seed sentiment list. Then, we give a sentiment value (+1, -1, 0) to the sentence at hand, according to the value of (PW-NW); i.e., +1 if PW-NW>0, -1 if PW-NW<0, and 0 otherwise. Moreover, for each specific word in the discussed sentence, which is not found in the sentiment list, we add the value of (PW-NW) to the sentiment score of the specific discussed word. After activating this process for all the sentences in the corpus, we have sentiment values for all the words in the corpus, which are not included in the sentiment list. We sorted these words according to their sentiment scores. The words with the highest positive scores are stored in the list of top words associated with positive words, and the words with the lowest negative scores are stored in the list of top words associated with negative words.

This paper is organized as follows: Section 2 supplies relevant background about the Hebrew language, sentiment lexicons, and their expansions, and sentiment blog lexicons and sentiment blog classification. Section 3 presents the two seed sentiment lists that our algorithm works with. Section 4 describes the examined corpus, the experimental results and their analysis. Section 5 presents a summary and proposals for research directions.

2. Relevant background

2.1. The Hebrew language

Hebrew is a Semitic language. It is written from right to left and it uses the Hebrew alphabet. Most Hebrew words are based on three (sometimes four) basic letters, which create the word's stem (root). Except for the word's stem, there are a few other components, which create the word's declensions, such as: belongings, conjugations, objects, prepositions, prefix letters, subjects, terminal letters, and verb types. Overview on these components can be seen at[1].

In Hebrew, it is impossible to find the declensions of a certain stem without an exact morphological analysis based on the components mentioned above.

The English language is richer in its vocabulary than Hebrew. The English language has about 40,000 stems, while Hebrew has only about 3,500 and the number of lexical entries in the English dictionary is 150,000 compared with only 35,000 in the Hebrew dictionary[2].

However, the Hebrew language is richer in its morphology forms. According to linguistic estimates, the Hebrew language has 70,000,000 valid (inflected) forms, while English has only 1,000,000[2]. For instance, the single Hebrew word וכשישתוהו is translated into the following sequence of six English words: "and when they will drink it". In comparison to the Hebrew verb, which undergoes a few changes the English verb stays the same.

In Hebrew, there are up to seven thousand declensions for only one stem, while in English there is only a few declensions. For example, the English word drink has only four declensions (drinks, drinking, drank, and drunk). The relevant Hebrew stem שתה ("drank") has thousands of declensions. Eight of them are presented below: (1) שתיתי