УДК 004.4.85
Н.І.Яворська, І.В.Миколюк, А.М.Стефанів, В.М.Бревус
ТНТУ ім. І.Пулюя, Україна

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ ТА ВИЗНАЧЕНННЯ ТОНАЛЬНОСТІ ТЕКСТУ

N.I. Yavorska, I.V. Mykoliuk, A.M. Stefaniv, V.M. Brevus

DATA MINING AND TEXT SENTIMENT ANALYSIS

The number of news articles published on various websites in general and news websites in particular had a dramatic increase over the last years. At the time of writing, text mining is generating a frenzy of debate in the scholarly publishing world. There is the usual misunderstanding, over-enthusiasm and unrealistic expectations that are associated with technology hype.

Josiah Stamp said: “The individual source of the statistics may easily be the weakest link.” Nowhere is this more true than in the new field of text mining, given the wide variety of textual information. By some estimates, 80 percent of the information available occurs as free-form text which, prior to the development of text mining, needed to be read in its entirety in order for information to be obtained from it. It has been applied to spam filters, fraud detection, sentiment analysis and identification of trends.

Text mining can be defined as the analysis of semi-structured or unstructured text data. The goal is to turn text information into numbers so that data mining algorithms can be applied. It arose from the related fields of data mining, artificial intelligence, statistics, databases, library science, and linguistics.

There are seven specialties within text mining that have different objectives. These can be decided by answers to the questions shown in the decision tree in Figure 1.

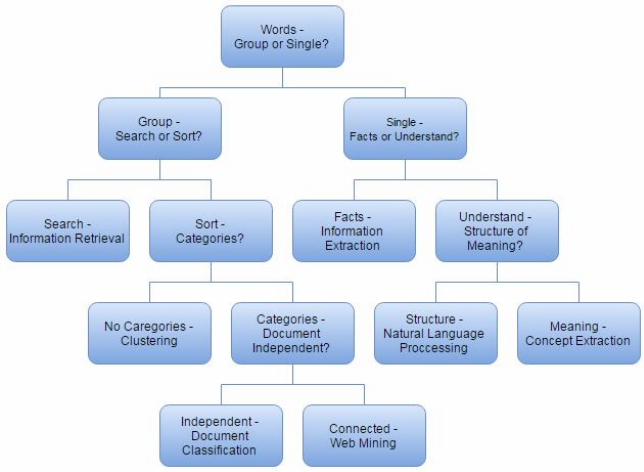

Fig. 1: Text Mining Decision Tree

As input data used comments from the Internet, which was divided into two parts: positive and negative according content.

The calculations are carried out based on the frequency distribution. The frequency of a particular observation is the number of times the observation occurs in the data. The distribution of a variable is the pattern of frequencies of the observation. Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. Frequency distributions are portrayed as histograms or polygons.

The results of calculations shown in the graph below.

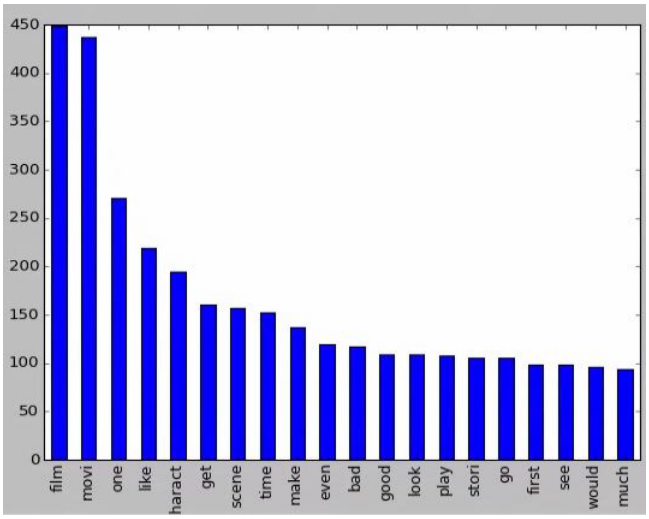

Fig. 2. Graph of words and number of occurrences in negative comments

After analyzing the negative comments were defined words that are used most often and allow you to identify similar comments. Figure 2 shows a histogram of the frequency of occurrence of words in the negative comments.

Similarly, in Figure 3 shows a histogram of the frequency of occurrence of words in the positive comments.

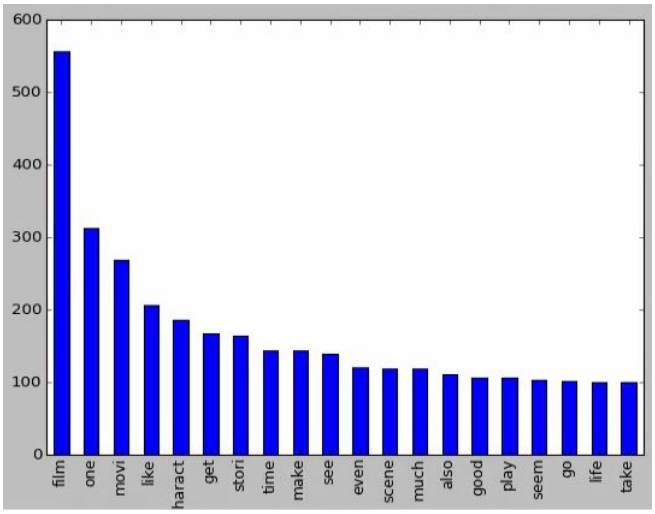

Fig. 3. Graph of words and number of occurrences in positive comments

It had conducted text classification sentiment analysis of internet comments to films. The frequency of occurrence of words in the text was determined and on the basis of an analysis of words belonging to a particular type comments like "positive" or "negative" and were found 5% most popular and most used words on "positive" and "negative" comments.