Is Sentiment Analysis Domain-Dependent

Elena Brunova and Yuliya Bidulya

Tyumen State University, Tyumen, Russian Federation

Abstract: The purpose of this research carried out within applied linguistics is to consider the dependency of the sentiment lexicon and other sentiment analysis tools on the domain under study. For the experiment, we used the REGEX algorithm including the sentiment lexicon and formal grammar rules applied with the certain priorities. These rules and the corresponding syntactic models are similar to regular expressions which detect certain text elements, simplify each sentence and present the text as a formal model. The reviews in Russian from three domains (Bank Service Quality, Hotel Service Quality and Sightseeing) are analyzed; F1 measure is used as the efficiency criterion. The experiment does not reveal the domain-dependency of the algorithm applied. It is determined that the system generally detects positive reviews better than negative ones. When negative opinions are expressed, there is a tendency to use non-standard vocabulary and syntax.

Key words: Applied Linguistics • Natural Language Processing • Algorithm • Sentiment Analysis • Regular Expressions • User-Generated Content

INTRODUCTION

Sentiment analysis is the method for extracting opinions and emotions from texts in natural languages and their further processing. It is a classification method: the corpus is generally divided into two classes (positive and negative opinions) [1-2], though sometimes the third class (neutral) [3] and the forth class (mixed) [4] are added.

The first works on sentiment analysis were published in late 1990s- early 2000s [1; 5-6] and since then much has been done in this field. Sentiment lexicons have been built; algorithms have been developed [7-10]. All these successful studies were focused on the English language and it seemed attractive to apply their results to other natural languages, translating the lexicons and modifying the tools for syntax analysis. As far as the sentiment analysis of the Russian texts is concerned, the publications only start to appear [3; 11-13], most of them being literature reviews.

Sentiment analysis depends on the natural language under study due to essential differences in morphology and syntax. For example, English is a language with a fixed word order within a sentence, while in Russian the grammar relations are mainly expressed on the morphological level. Thus, the tools for the English syntax analysis are not very helpful for the Russian one.

Moreover, the attempts to build a universal sentiment lexicon, the principal and the most time-consuming sentiment analysis tool, sometimes fail even within the same language, but different domains [14]. Does sentiment analysis depend on the domain as it depends on the language.

The purpose of this research is to consider the dependency of the sentiment lexicon and other sentiment analysis tools on the domain under study.

Data and Methods of the Research: For the experiment, 80 reviews on the Bank Service Quality from [15], Hotel Service Quality and Sightseeing from [16] in Russian, were randomly selected. Six documents were removed from the corpus as they were the results of machine translation. Thus, the experimental corpus included 74 documents from three domains (32, 22 and 20 reviews from the Bank Service Quality, Hotel Service Quality and Sightseeing domains, respectively).

The sentiment analysis was carried out using the REGEX algorithm which consists of the sentiment lexicon and formal grammar rules applied with certain priorities. These rules and the corresponding syntactic models are similar to regular expressions which detect certain text elements, simplify each sentence and present the text as a formal model [17].
The lexicon consists of five classes, including two primary classes (positive and negative lexicons) and three secondary classes (increments, polarity modifiers and polarity anti-modifiers). The fragments of the sentiment lexicon are presented below.

**Positive Lexicon:** Bezopasnyj (safe), besplatnyj (free), vezhlivyj (polite), kompepentnyj (competent), chotkij (clear), effektivnyj (efficient).

**Negative Lexicon:** Agressivnyj (aggressive), bezvykhodnyj (hopeless), grubyj (rude), dosadnyj (annoying), obidnyj (offensive), trudnyj (difficult) …

**Increments:** Ochen’ (very), sovershenno (absolutely), nikogda (never) …

**Polarity Modifiers:** Ne (no), net (not), bez (without) …

**Polarity Anti-modifiers:** Tak (so), takoj (such) …

The experiment was carried out in the SENTIMENTO system implemented as an Apache-based Internet application [17]. The system has an opportunity to add words to the sentiment lexicon database, both by importing Excel files and by adding single words.

After the system performs the sentiment analysis, the user can confirm or reject the system conclusion.

For this purpose, the *Your conclusion* message and two buttons (*Positive* and *Negative*) are displayed. After the user presses one of these buttons, the system checks if the system conclusion matches the user’s one. Actually, both the system and the user answer the same question whether the document belongs to a certain class, e.g. whether this document is positive. To compare the system conclusion with the human one, the two-by-two contingency table is used, see Table 1.

The Precision and Recall are calculated. Precision (P) is the fraction of retrieved documents that are relevant P = tp/(tp+fp). Recall (R) is the fraction of relevant documents that are retrieved R = tp/(tp+fn). Then F1 measure is calculated using the formula [9]:

\[
F1 = \frac{2P\times R}{P + R}
\]

**Experiment:** At the first stage of the experiment, the REGEX algorithm was tested within a single domain. The results are presented in Table 2.

Then we checked the domain-dependency of the REGEX algorithm. The algorithm was tested within a larger number of documents from the Bank Service Quality domain and from two more domains: the Hotel Service Quality and Sightseeing. The results are presented in Table 3.

<table>
<thead>
<tr>
<th>Table 1: Two-by-two contingency table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct (the user says yes)</td>
</tr>
<tr>
<td>Incorrect (the user says no)</td>
</tr>
<tr>
<td>Selected (the system says yes)</td>
</tr>
<tr>
<td>Not selected (the system says no)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Sentiment analysis efficiency within a single domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>Precision (POS)</td>
</tr>
<tr>
<td>Recall (POS)</td>
</tr>
<tr>
<td>F1 (POS)</td>
</tr>
<tr>
<td>Precision (NEG)</td>
</tr>
<tr>
<td>Recall (NEG)</td>
</tr>
<tr>
<td>F1 (NEG)</td>
</tr>
<tr>
<td>F1 (average)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: The sentiment analysis efficiency within three domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Precision (POS)</td>
</tr>
<tr>
<td>Recall (POS)</td>
</tr>
<tr>
<td>F1 (POS)</td>
</tr>
<tr>
<td>Precision (NEG)</td>
</tr>
<tr>
<td>Recall (NEG)</td>
</tr>
<tr>
<td>F1 (NEG)</td>
</tr>
<tr>
<td>F1 (average)</td>
</tr>
</tbody>
</table>
RESULTS AND DISCUSSION

As it can be seen from Tables 2 and 3, the Precision, Recall and F1 values for the Bank Service Quality domain slightly decreased after the testing corpus had been expanded. This occurred due to the fact that some sentiment words easily identified by a user had not been included in the sentiment lexicon and hence were not identified by the system. This can be easily corrected by adding such words to the sentiment lexicon.

The experiment demonstrated that the Precision, Recall and F1 values did not decrease when our sentiment lexicon and REGEX algorithm were applied to other domains. On the contrary, the average F1 values for the Hotel Service Quality and Sightseeing domains are even slightly higher than those for the Bank Service Quality domain (0.955, 0.95 and 0.875, respectively). It means that our algorithm does not depend on the domain, though it definitely depends on the language. The secondary classes of the lexicon (polarity modifiers, anti-modifiers and increments) are universal for any Russian text. However, if we try to apply their translations to other languages (e.g. English), the results will be not so accurate. For example, due to the non-occurrence of double negation in English, such words as never, nobody, or nothing, cannot be considered polarity increments, they are definitely polarity modifiers, since they do not increase the polarity, but change it into the opposite ones. As for the formal grammar rules, some of them should be also changed in case of English. For example, the word order within a sentence or its fragment is not significant for the interpretation of Russian sentences, but it is really critical for the English sentences.

Our special attention was paid to the system errors (fp and fn). In these cases the documents were analyzed manually to determine the problems of the REGEX algorithm, both the lexicon and formal grammar rules, which result in inaccurate sentiment analysis. They can be grouped as below:

Ambiguity:

Multiple Meaning: Some words may have a sentiment meaning along with a non-sentiment one, e.g. S menya dovol’no! (That's it for me!), but Komnata dovol’no udobnaya (The room is rather comfortable.) In the first example dovol’no is a sentiment (negative) word, in the second one, it is an increment.

Parametrical Words: Some words from sentiment lexicons appear domain-specific, e.g. dolgo (long) can be ranged into the positive lexicon when evaluating the battery operation (the Smartphone domain), but it can be ranged into the negative lexicon in evaluating the client’s time consuming (the Bank Service Quality domain). These are parametrical words, i.e. the words denoting the amount of some domain-specific parameter (battery life, the client’s time consuming, etc.) [18].

Collocations: The behavior of a single word may differ from the behavior of the same word within a collocation, e.g. tak (so) is a polarity anti-modifier, whereas tak sebe (not up too much) has a negative meaning and òàê tak derzhat’! (that's the way to go!) has a positive one.

Ambiguous È (And): The i (and) conjunction generally connects the coordinated parts of the sentence, e.g.:

Devushka predlozhila kartu Visa Gold, ne obyasniv mne usloviya obsluzhivaniya i ne pokazav tarify. (The girl offered me a Visa Gold card without explaining the service terms and showing the service rates.)

The detection of this conjunction is very important for some formal grammar rules, in particular, for the rule concerning polarity modifiers, as in the example above. The ne (here- without) modifier changes the polarity of two positive words: obyasniv (explaining), pokazav (showing) which are connected by i (and).

However, sometimes i (and) does not connect the coordinated parts of the sentence, but behaves as an increment, e.g.:

Mogli by ostavit’ vkhod i besplatnym. (They could leave the entrance free.)

Irony and Sarcasm: Sometimes the author uses positive words whereas the meaning of the sentence is negative. If such words are between the quotation marks, we can detect them using one of our formal grammar rules, e.g. ‘privlekatel’nuye usloviya’ (‘attractive terms’). If the quotation marks are not used, our algorithm fails to detect real opinion, e.g.:

VOT ETO SERVIS (POS)! U nas mnogotysyachnyj gorod! I net vozmozhnosti vnesti nalichnuy’! Krasota (POS)! A esli lyudjam kredit nado platit’? (THAT’S QUITE THE SERVICE (POS)! Many thousands of people live in our town! And there’s no opportunity to top up your account using cash! Wonderful (POS)! And what if people have a loan?)
Thwarted Expectations: In some reviews the beginning is positive and the conclusion is negative, the turning point is usually marked with no (but), odnako (however), etc., e.g.:

Kogda otkryvaesh' nomer, zahodish', to hochetsya skazat' "O! Kruto (POS)!". No potom ponimaesh', chto vsjo ochen' ne udobno (NEG) i ne produmano (NEG). (When you open a room you’d like to say “Oh, cool (POS)!” But later you understand that everything is very uncomfortable (NEG) and ill-designed (NEG)).

Negative Lexicon Can Be Used to Describe the Domain Itself, but Not the Opinion: This problem usually occurs when horror books, or movies, are analyzed. We encountered it when analyzing the reviews on the hotel service, e.g.:

Na kovrah v koridore frazy iz ‘Prestuplenija i nakazaniija’, mrachnye tona (NEG), polumrak (NEG). Vse Vam daet ponjat', chto otel' sdelan 'po Dostoevskomu'. Dazhe to, chto v tsentral'nom korpulse potolki ne otdelany (NEG), a prosto pokrasheny v seryj (NEG) tsvet pozvoljaet Vam prosnut'sja i uvidet' etot mir glazami Raskol'nikova. (There are quotations from ‘Crime and Punishment’ on the carpets in the corridor, gloomy (NEG) colors, poor (NEG) light. All this implies that the hotel is designed ‘a la Dostoevsky’. Even the fact that in the central building, the ceilings are not finished (NEG), but simply painted grey (NEG), helps you to wake up and see this world with Raskolnikov’s eyes.)

Wrong Stemming: As stemming is very important for our algorithm, stemming errors may influence on the sentiment analysis results. In case of wrong stemming, the word is not detected by the system, or is ranged to the wrong class, e.g. lyub may mean not only lyubit’ (love, pos), but also lyuboj (any, neutral).

Wrong Spelling And/or Punctuation: In case of wrong spelling, the word is not detected by the system, or is ranged to the wrong class. In case of wrong punctuation, the formal grammar rules concerning punctuation mark within a sentence do not work correctly.

CONCLUSION

Our experiment demonstrated that the REGEX algorithm does not depend on the domain under study. When applied to other domains, its efficiency does not decrease. The system errors were analyzed. The lexicon was extended. The stemming errors were corrected. The spelling and punctuation check was added to the algorithm. As for the irony and sarcasm, it is often impossible to detect it with the sentiment analysis tools.

It can be seen from Table 3 that the efficiency for positive reviews is higher than for negative ones (0.92 and 0.87, respectively). This tendency remains when two more domains were added, see Table 4. In the Bank Service Quality and Hotel Service Quality domains (two domains out of three), the efficiency for positive reviews is slightly higher than for negative ones. It means that the system generally detects positive reviews better than negative ones. This can be explained by the fact that an author expressing positive opinions and emotions often uses standard, clichéd words and phrases. On the contrary, an author expressing negative emotions tends to use non-standard language means. The variability in case of negative emotions is just the feature L. Tolstoy described, ‘Happy families are all alike; every unhappy family is unhappy in its own way.’ [19]. When the natural language variability increases, the risk of the absence of a positive or negative word in the sentiment lexicon also increases.

However, the tendency of higher efficiency of identifying positive reviews is not observed in the Sightseeing domain, F1 (POS) =0.94; F1 (NEG) =0.96. When expressing a negative opinion about sights, an author generally is not so emotional, as in this case he or she is not involved into person-to-person conflicts. Hence, the variability of language means is also not so high.

REFERENCES

15. www.banki.ru
16. www.tripadvisor.ru