Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

Development of two data bases with comments in Bulgarian language and application of supervised learning approaches on them for comparative sentiment analysis. A brief overview

Daniela Petrova[1], *Violeta* Bozhikova[1]

1 – Technical University of Varna, Department of Software and Internet Techologies, 9010, 1 Studentska Street, Varna, Bulgaria

Corresponding author contact: daniela.petrova@tu-varna.bg

Abstract. The purpose of the current paper is to make an overview of the work done so far by the authors and make a summary of the results and reflections on the performed sentiment analysis on user comments in two different fields in Bulgarian language. As a starting point for the authors' work is the development of two databases with users' reviews and their preprocessing to become usable source of information for different types of analysis projects. As a result of the preprocessing is a revised Bulgarian language-driven algorithm for data preprocessing for Bulgarian language. The second part of the project is implemented into two steps: sentiment analysis using the supervised learning approaches developed for the two databases and a comparative sentiment analysis of the two databases, following their additional examination.

Keywords: Automatic Sentiment Analysis, opinion mining, supervised learning approaches, Bulgarian language

1 Introduction

In the present-day world, people are ever-more prone to seek, when choosing a product or service, for the prevailing opinion of the others expressed clearly in different comments, blogs and other sources of information. In the marketing sector it is essentially important for the companies to follow the comments and opinions people hold about their products or services. This leads to the growing need of automated analysis and opinion mining into the expressed user comments (Sadegh, 2012). Nowadays, there are many applications for sentiment analysis regarding the English language – different algorithms, applications, databases, while for the Bulgarian language one can find only a few.

We are aware of the work done by Preslav Nakov and Borislav Kapukaranov, which is fine-grained sentiment analysis for movie reviews in Bulgarian and is a pioneering work in this field on the subject of Bulgarian comments, as well as the creation of a dataset with Bulgarian movie reviews. (Kapukaranov, 2015)

While movie reviews are popular and widely available source for sentiment analysis, sentiment extraction on other type of customer feedback, for all we know, hasn't been done in Bulgarian so far. Thus, our contribution up to now could be generalized as follows:

- Created have been two databases with Bulgarian reviews, available for further researches;
- Prepared is a stop word list in Bulgarian;
- Proposed are refined steps for preprocessing Bulgarian reviews;
- Presented is a sentiment analysis on comments, other than movie reviews in Bulgarian.

2 Related work

Extensive research has been done into sentiment analysis of texts and reviews in English, but due to the specifics of the Bulgarian language they are not fully applicable for Bulgarian reviews of any kind.

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

As mentioned above, the first work on sentiment analysis of Bulgarian reviews is done by Preslav Nakov and Borislav Kapukaranov. There is another research done by Tsvetana Dimitrova and Valentina Stefanova with respect to the Bulgarian language. Their paper presents an attempt of semantic classification of adjectives in the Bulgarian WordNet (Dimitrova, 2018), which is also a step ahead in sentiment analysis of Bulgarian texts.

Another work is Ivelina Stoyanova's research into the automatic detection and tagging of phrases in Bulgarian language. According to her, multi-word expressions are a major part of the lexical system of the language and in the same time they are difficult to detect by an automatic system. Dealing with the problems connected with multiword expressions will enhance significantly the results of different natural language processing applications, including sentiment analysis. (Stoyanova, 2012)

3 Development of databases

The authors, having taken their first steps to sentiment analysis in Bulgarian language, made it known that there are no ready- to- use databases with reviews in Bulgarian. That fact drove the need to develop one, and later a second database with user comments in Bulgarian. An additional hurdle was the lack of sites with sufficient comments to compose a large enough database. Selected, thus, as the only appropriate and rich enough databases, were booking.com and grabo.bg.

An agreement was reached that it was only the comments on hotels and guest houses in the biggest towns and resorts in Bulgaria, which could be taken from both sites that reflect the substantial amount of reviews to use for a database. The result was a database (later referred as Database 1) of 100 082 reviews in total, 72 078 of which positive and 28 004 – negative. Both sites have different ways of structuring their user comments. In booking.com every user is allowed to leave a positive and a negative review for every hotel. While in grabo.bg there is only one field that allows the users to enter their comments, which are then rated with stars from 1 to 5. In order to unify the reviews from both sites, the comments, rated with 4 or 5 stars were marked as positive, and those with 1, 2, or 3 stars – as negative. Apparently, the difference between the number of positive and the number of negative reviews would seem quite big. One of the reasons is that in booking.com the field "negative opinion" was frequently left empty or was filled with sentences like "There is no such thing", "We are satisfied with everything", "Everything was OK" and those comments had to be removed as they would be marked wrongfully as negative. (Petrova, 2021)

After the first sentiment analysis on the database using the most common supervised learning approaches and their inconclusive results, the author decided to create a second database, so that it could be used for comparison with the same algorithms but other data. As there wasn't found another source of data, grabo.bg was used again – this time with comments on different products and services – such as restaurants, theaters, beauty salons, shops and other entertainment. A database (Database 2) with 105 052 reviews was created, 90 384 of which positive and 15 062 – negative.

4 Preprocessing

The critical assessment of the two databases revealed that certain sentences are used as comments more than once, such as "We are very satisfied", "Everything was perfect" and so on. In order to have a database with unique reviews and consequently - accurate analysis, the first step of the preprocessing was removing the duplicate rows. This was done by a short script which could find and delete duplicates. It was also found that the comments were expressed in Bulgarian and non-Bulgarian language but written in Latin letters. So as not to lose any data, those comments, written in Latin letters were transformed in Bulgarian, and those written in English or other languages were removed. A logical further development of the databases, therefore, could be scripts that find and corrects misspelled words, which could lead to an ever better and more correct database. This would be important as the vectorizers used in the analysis count the words in the sentences and the number of times they are used in the review and when a word is misspelled it will be wrongfully counted as a different word. In addition, the data was cleared of multiple empty spaces and punctuation.

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

Based on the adjustments made to the databases, the authors propose the following steps in the pre-processing phase of Bulgarian language-driven data, which would be used in sentiment analyses (see Fig.1).

Fig. 1. Steps of preprocessing of data in Bulgarian language

The next two steps of the preprocessing are recommendable, despite language being used, but their application on Bulgarian texts is a serious obstacle. These steps are removing stop words and stemming or lemming. There are plenty of stop word lists in English and applications for stemming and lemming, but this is not the case for Bulgarian language. This led to the creation of an own stop word list, which were removed from the comments. The difference in the databases before and after removing the stop words could be seen in Fig. 2 and Fig. 3 with the most common words before and after stop words removal.

Fig. 2. Most common words before stop words removal

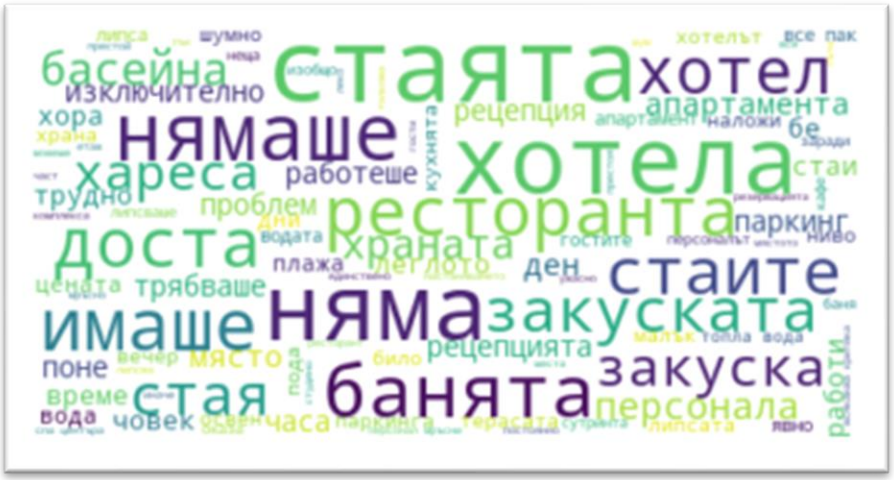

Fig. 3. Most common words after stop words removal

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

It could easily be seen that the first figure consists mainly of prepositions, pronouns and other parts of speech with little effect on sentiment analyses and opinion mining, which leads to the need of removing all stop words from the documents.

Adopted, for the stemming of the words, was the only working module that the authors could find, developed for Bulgarian language and Python, by Prelsav Nakov from the University of California, Berkley (Nakov, 2003). It removes the suffixes of the words and leaves the roots of the words. As it was sometimes very time consuming, this stage of the preprocessing was used as a test whether it increases or not the final accuracy of the sentiment analysis of the comments.

After applying all preprocessing steps mentioned above, the size of the two databases was reduced with around 10 000 reviews each.

Table 1. Databases after preprocessing

Comments	Database 1	Database 2
Positive	63 714	84 489
Negative	25 624	14 357
Total	**89 341**	**98 846**

5 Results

Applied on the so databases with user comments were applied different methods for sentiment analysis using supervised machine learning algorithms:

- Naïve Bayes classifier - despite its simplicity, the Naïve Bayes classifier could also be a preferred technique for classification of texts and could be extremely effective in different spheres. Throughout the course of operation, Naïve Bayes takes a random model for generating a document. Using the rule of Naïve Bayes this model is inverted to predict the foremost probable category of the new document; (Ye, 2009)
- Logistic regression - belongs to the family of classifiers mentioned as the exponential or log-linear classifiers. The foremost important difference between Naive Bayes and logistic regression is that logistic regression is a discriminative classifier while naive Bayes is a generative classifier; (Wankhade, 2017)
- Logistic regression with bi-grams;
- Support Vector Machines (SVM) it's a statistical classification method, first introduced by Vapnik (1995). This model may well be used for binary and multi categories classification. SVM seeks for the hyperplane, represented by vector w that divides the positive from the negative vectors with an optimal margin; (Ye, 2009)
- Support Vector Machines with bi-grams.

Chosen for all the methods was Term Frequency and Inverse Document Frequency (Tf-idf) vectorizer, which has proven to give higher results in the previous work of the author. Tf-idf weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the amount of times a word appears within the review as it calculates values for every word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a robust relationship with the document they seem in. (Ramos, 2000)

The results of all predictions are listed in Table 2. The highest accuracy in sentiment prediction using Database 1 give both Logistic regression with bi-grams and Naïve Bayes classification, while for Database 2 – the most accurate is SVM with bi-grams. It is clear that after using stemming in the preprocessing steps the final accuracy is higher than without stemming, although the difference is not big. (Author, 2021)

Table 2. Classification results

Methos	Database 1	Database 2

Technical University of Varna
Annual Journal
https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

	Without stemming	
Naïve Bayes	**0.8683**	0.8530
Logistic regression	0.8503	0.9355
Logistic regression using bi-grams	0.8575	0.9338
SVM	0.8239	0.9375
SVM using bi-grams	0.8413	**0.9451**
	After stemming	
Naïve Bayes	0.8639	0.8489
Logistic regression	0.8508	0.9387
Logistic regression using bi-grams	**0.8644**	0.9410
SVM	0.8326	0.9385
SVM using bi-grams	0.8475	**0.9467**

It can also be seen that the final accuracy is significantly higher when using Database 2. The reasons for such 10% difference could be the following:

- Database 2 has around 10% more negative comments;
- Difference in the bag of words or phrasing.

To find out whether the difference in the final accuracy is perceived as a difference in the count of negative and positive comments, the two databases were transformed to be equal: both to have 63 714 positive and 14 357 negative reviews. This led to a smaller database of 78 071 documents, but gave a clearer view of the reason for the different results. The same models and supervised learning algorithms were used on the transformed databases. The final accuracy was, once again, with 10% higher for Database 2. This suggests that the reason could be the variety in the phrasing and bag of words, used in the two databases. To verify the accuracy of this statement, a comparative analysis of the most commonly used positive and negative words in both databases was done. The results of this analysis are shown in Table 3. (Author, 2021)

Table 3. Comparative analysis of the most commonly used words

Words	Database 1	Database 2
„страхот“ (awesome)	8.4045	10.0331
„довол“ (satisfied)	7.9813	9.3768
„вкусн“ (delicious)	6.0880	7.8047
„прекрас“ (wonderful)	8.0667	7.2126
„не“ (no)	-9.4110	-14.8634
„разочаров“ (disappointed)	-3.9197	-6.0161
„ужас“ (terrible)	-5.2182	-5.6326
„не довол“ (not satisfied)	-2.9348	-5.9610
„зле“ (bad)	-3.2440	-3.5335

There is an evident distinction in the weights of the most common words in the two databases, again in favor of Database 2, which could be the reason for the difference in the final accuracy results.

6 Conclusions

It should be noted, from all that has been mentioned above, that the results for the final accuracy depend not only on the different methods for sentiment analysis using supervised learning, but also on the bag of words and phrasing used in the database. With the two databases commensurable in their positive and negative comments and all other conditions being equal, the results and the final accuracy fail to be consistent.

The results of this research are two databases in Bulgarian, ready to be used in different projects as well as a refined algorithm for preprocessing of Bulgarian language-driven data. The applied methods for sentiment analysis for this project were only using supervised learning, but they could be combined with unsupervised learning in search for higher results in the future work of the authors. Another approach for further investigation on the two databases could be cross-domain sentiment classification –

https://doi.org/10.29114/ajtuv.vol6.iss2.261
Vol. 6 Issue 2 (2022)
Published: 2022-12-31
ISSN 2603-316X (Online)

which is the application of a sentiment classifier, designed to use labeled data on a particular domain to classify sentiment of reviews on a different domain, although it sometimes results in poor performance because the train domain-specific words might not appear in the test domain.

Acknowledgments

The scientific research, the results of which are presented in this publication, was carried out and funded as part of a project for financial aid for PhD students PD6/2021 in Technical University - Varna, specifically financed by the state budget.

References

Dimitrova, T., Stefanova, V. (2018). The semantic classification of adjectives in the Bulgarian Wordnet: Towards a multiclass approach. *Cognitive Studies | Etudes cognitives*, 2018(18). https://doi.org/10.11649/cs.1709

Hajmohammadi, M. S., Ibrahim, R., & Othman, Z. A. Opinion mining and sentiment analysis: a survey. INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY, 2(3c), 171-178. https://doi.org/10.24297/IJCT.V2I3C.2717

Kapukaranov, B., Nakov, P., Fine-grained sentiment analysis for movie reviews in Bulgarian, Proceedings of Recent Advances in Natural Language Processing, p. 266-274, Hisar, Bulgaria, Sep.7-9 2015. https://aclanthology.org/R15-1036.pdf

Nakov, P. (1998). BulStem: Design and Evaluation of Inflectional Stemmer for Bulgarian., Retrieved from https://www.researchgate.net/publication/250443777_Design_and_Evaluation_of_Inflectional_Stemmer_for_Bulgarian

Petrova, D. (2021) Automatic Sentiment Analysis on Hotel Reviews in Bulgarian – Basic Approaches and Results, IEMAICLOUD - London April 2021, https://doi.org/10.1007/978-3-030-92905-3_5

Petrova, D. (2021) Comparative assay on sentiment analysis on two databases in Bulgarian language, Interdisciplinary Conference on Mechanics, Computers and Electrics, Ankara, Turkey, 27-28 November 2021, ISBN: 978-625-409-707-2, to be published

Ramos, J.E. (2003). Using TF-IDF to determine word relevance in document queries. Tech. Rep., Departament of Computer science. Rutgers University Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf

Стоянова Ив.”Автоматично разпознаване и тагиране на съставни лексикални единици в българския език“, BAS, Sofia, April 2012 Retrieved from https://ibl.bas.bg/wp-content/uploads/2014/10/IStoyanova-avtoreferat.pdf

Wankhade, M., Chandra, A.,Rao, S., Dara, S.,Kaushik, Baij. (2017). A sentiment analysis of food review using logistic regression. International Conference on Machine Learning and Computational Intelligence-2017, 2456-3307, https://www.researchgate.net/publication/334654833

Ye Q., Zhang, Z. & R.Law. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, *Expert Systems with Applications 36*, 2009, p.6527-6535, https://doi.org/10.1016/j.eswa.2008.07.035