Serdica J. Computing **13** (2019), No 1–2, 81–96

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

SMS SENTIMENT CLASSIFICATION BASED ON LEXICAL FEATURES, EMOTICONS AND INFORMAL ABBREVIATIONS*

Branislava Šandrih

Abstract. In this paper we investigate the influence of emoticons, informal speech, lexical and other linguistic features on the sentiment contained in SMS messages. Using the dataset of $\sim$ 6,000 samples, we trained a linear SVM classifier able to determine positive, negative and neutral sentiments. The dataset mostly contains messages in Serbian, but also in English and German. The classifier had an average accuracy score of 92.3% in a 5-fold Cross Validation setting, and F1-score of 92.1%, 74.0% and 93.3% in favor of positive, negative and neutral class, respectively.

1. Introduction. Exchange of short messages is one of the most popular communication styles of the present times. People communicate in this form using Facebook, Twitter, Instagram, WhatsApp, SMS, etc. Many researchers have focused their work on analysing datasets obtained from Twitter or Facebook.

ACM Computing Classification System (1998): J.5, H.3.5, I.7.5.

Key words: computer application in arts and humanities, web-based services, document analysis.

*Initial ideas and directions for this work were announced in the book of abstracts of the International Quantitative Linguistics Conference (QUALICO 2018).

Yet, not so many papers have been dedicated to analysis of SMS messages. This is probably due to personal nature of these messages, which makes obtaining a dataset with the size comparable to the size of datasets retrieved from micro blogging services a hard task. This is a paradox in some sense, since this is one of the oldest and most used forms of digital communication.

Any analysis of SMS messages has to cope with consequences of some specific circumstances. Firstly, each message is restricted to the length of 160 characters. Therefore, SMS messages often do not contain enough information for the analysis of their meaning. This is followed by the presence of many spelling errors and typos. Nowadays people mostly use post paid contracts with mobile network operators and therefore can concatenate and send many messages instead of one, but they still tend to write very short messages. One of the potential reasons of this shortness is a design of modern keyboards and the in-proportion of the size of a key and a finger tip. Another reason could also be a consequence of an old habit. Finally, SMS messages lost their popularity in favor of applications for instant messaging.

Another specificity can be seen when people text using a language with diacritics. For example, Serbian uses two alphabets (Latin and Cyrillic), the former with five additional letters with diacritics ('č', 'ć', 'đ', 'š', 'ž'). Mobile network operators restrict messages that contain diacritics to a length of 70, regardless of the number of diacritics contained within a message. This also applies to messages in Cyrillic. As a consequence, people usually omit the diacritics. Since electronic language tools contain words and their lemmas written in their correct forms, i. e., with diacritics, these tools cannot be applied to the SMS datasets, without some previous step of restoring diacritics.

Beside all these restrictions, authors of short messages have a need to express their mood, voice tone, facial expressions and much more of what oral communication contains. In the written communication, the only available tools are characters. The authors in [19] explore a dataset of Twitter messages and analyse types of transformations that occur in these texts. They notice that people tend to write messages in a way that people who read them can experience the whole emotional state of the author. For example, they use uppercase letters in the case of "shouting"; they excessively use emoticons in order to express their mood and attitude; other often used transformations are common abbreviations and shortened form of words.

In this research, a dataset of SMS messages from one person's smart phone is used. This dataset was initially described and analysed in [25]. Most of the senders were in their early twenties and they used the "popular" texting language.

This means the use of short forms of words, abbreviations and emoticons. Therefore, this work relies on the informality of the dataset and tries to discover the influence of modern language patterns on sentiments contained in messages. It should be noted that analysis performed on these texts slightly differs from the standard definition of Sentiment Analysis (SA). It can rather be considered a *mood analysis* since it tries to distinguish in what *tone* should a reader experience a message. Similar service is offered by commercial systems Twilio[1] and Nexmo.[2]

The dataset mostly contains messages in Serbian (more than 96%), and also in English and German. There are no external language resources used, but the method relies on, among others, predefined set of features that are modern short forms of common words in Serbian and English. Each message in the dataset was first manually annotated as having positive, negative or neutral sentiment. Next, a set of text features was extracted and SVM classifier was built using this set of features.

2. Related Work. Much research has been done related to Sentiment Analysis in short texts based on emoticons and slang abbreviations. In [27], an experiment was conducted in order to determine the effects of three common emoticons on message interpretations. The results indicated that the contributions of emoticons were outweighed by verbal content. In [22], SA was performed on training data labelled with emoticons, i. e., using an approach independent of domain, topic and time. Classification was done on the data consisting of a large collection of blog posts which include an indication of the writer's mood, author's mood classification was performed in [12]. The study in [5] examined the influence of social context on the use of emoticons in Internet communication. Participants in the short chat were asked various questions and had to respond with a text, emoticon or a combination. The results showed that more emoticons were used in socio-emotional than in task-oriented social contexts. In [9], the task of automatic emotion analysis and generation in texts was explored. The authors classified texts by classes of emotions. At the end, they discussed the possibility of generating texts that express specific emotions. In [16] the task of recognising personal emotional state or a sentiment conveyed through text was addressed. The authors developed an Affect Analysis Model designed to handle the informal messages written in an abbreviated or expressive manner.

Many authors performed different methods for SA on Twitter data. Regardless of the dataset, the task of SA in the context of social media relies on

[1] Twilio, `www.twilio.com`

[2] Nexmo, `www.nexmo.com`

predefined sets of emoticons. The authors of [8] applied Machine Learning algorithms for classifying the sentiment of Twitter messages using distant supervision on training data consisting of Twitter messages with emoticons. They proved that standard Machine Learning algorithms have higher accuracy when trained on data with emoticons. In [20], the authors created a database of emoticons, gathering emoticons from numerous dictionaries of face marks and online jargon. They decomposed each emoticons into "mouth" and "eyes" elements and then analysed patterns of these semantic areas of emoticons. In [21], the importance of emoticons in Natural Language Processing was discussed. A supervised SA framework was proposed, which was based on data from Twitter, by utilising 50 Twitter tags and 15 smileys as sentiment labels [4]. The authors also explored dependencies and overlap between different sentiment types represented by smileys and Twitter hash tags. A SA on Twitter was performed in [15], where the gold standard was obtained by automatically annotating tweets based on their hash tags. The problems of spam, misspellings, slang and abbreviations, entity specificity in the context of the topic searched and pragmatics embedded in text were addressed in a multi-stage system.

In [10], an overview of scholarly research in the field of electronic communication was made, in order to investigate applications of emoticons in some facets of computer-mediated communication. The authors of [18] questioned whether predefined pictographic characters, a. k. a. "emojis", will come to replace earlier orthographic methods of para-linguistic communication, a. k. a. "emoticons". The focus in [26] was on using emoticon-rich texts on the Web in language-neutral SA. For that purpose, a Desktop application was implemented and tested.

Emoticon analysis was not the only approach that gave good results. The authors of [11] chose a supervised statistical Text Analysis approach, leveraging a variety of semantic and sentiment features in order to detect sentiments of short informal textual messages. They also use three general-purpose sentiment lexica that automatically capture many peculiarities of the social media language, containing common intentional and unintentional misspellings.

The authors of [2] chose a lexicon-based approach. They examined the similarity between Twitter feeds and SMS messages found on smart phones. They investigated common characteristics of both formats for the purpose of SA. Spoken utterance transcripts were analysed in [17]. The authors tried to estimate speaker's attitude towards the dialogue by exploiting this information. They applied SA tools to conversational data in order to extract sentiments that may be mapped onto speakers' experience of the dialogue as a whole, instead of performing standard analysis of participants' sentiments.

The approach proposed in this work is similar to the one in [17], in the sense that the sentiment of a message was assumed based on the author's texting style. There are no lexica of adjectives with sentiments used or created as, for example, in the work described in [14] and [13]. As in [11], we propose a method that relies on previously compiled sets of common abbreviations used in modern texting. Based on an assumption that authors of short messages tend to express their mood by the specific usage of characters (including grouping them into emoticons), different types of features are selected, classified and extracted. As in [1], we developed a Web service and a Web application, not for the text document classification itself, but for the extraction of the mentioned features.

3. Annotation of Samples. A dataset of 6,298 SMS messages was collected in XML format. Each message contains information about sender's number, date, message body and some other metadata. Following is an example of a single SMS message from the dataset.

```
<sms address="+38164305****" date="1424530897293" type="1"
contact_name="Gri***" readable_date="21.02.2015 4:01:37 PM"
body="..."  />
```

Messages were previously manually labeled as either neutral (i. e., carries no sentiment information, 3,272 samples), positive (2,719) or negative (180 samples). Messages that contain less then 10 characters (including blanks, total of 127) were discarded, since the human annotator was not able to indicate the mood contained in these messages, due to the lack of information. Some examples of such messages are: `4a`, `716`, `z*`.[3] Therefore, the final dataset contains 6,171 samples (messages). Some sample messages along with their annotations are given in Table 1 and their English translations are in Table 2.

The annotation of these messages was not an easy task in many cases. It was important for it to be performed adequately, since the outcome of the later classification is tightly connected to the way in which messages were manually categorised. In Table 3 some of the messages that contain ambiguous sentiment are given, along with their possible categorisations. Their English translations are in Table 4.

4. Feature Selection. We selected features and divided them into three main categories. Different authors suggest different names and organise

[3]The first two messages are names of classrooms in a building, and the third is the continuation of a previous message, where an author made a typo and wanted to make a correction.

Table 1. Example of messages and their annotations

Body	Label
ae! :D cemo na fb da skupljamo ekipu? u koju cemo? :)))) Nisam mislila na zadatak, zadatak je interesantan. :)	POS
Eeeeeeeeeeeeeeee bre, djubre prehlada. :/ Brande moj, cu li ti...	NEG
Kredit je dopunjen sa 200,00 din i vazi do 24.11.2016. Poštovani, vozilo 15 je na adresi. Vaš GOLUB TAXI	NEU

Table 2. English translation of the messages from Table 1

Body
Great! :D Are we gathering people on Facebook? Where shall we go? :)))) I didn't mean about the task, the task is interesting. :)
Noooooooooooo, stupid cold. :/ Have you heard about the news...
Your account has been reloaded with 200,00 RSD and it expires in 24.11.2016. To whom it may concern, the taxi 15 has arrived. Your GOLUB TAXI

these features differently. For example, the group of features that [6] and [23] call "linguistic" features, [3] term "stylistic" features. In this work, the naming is most similar to the one suggested by [3]. We explain these categories in what follows.

Lexical Features. These features (70 in all), as suggested by [3], can be observed on the basis of the characters and of the words.

- **Character-based.** This group of features (63 in all) includes counts of each punctuation character,[4] lowercase and uppercase alphabetic characters, digits, diacritics, umlauts, etc. Apart from the absolute counts, ratios of all these numbers to the total number of characters in the message were added as additional features.

- **Word-based.** Word-based lexical features used especially for this task are (7 in all): average length of tokens,[5] average sentence length, ratio of short words (up to three letters) to the total number of tokens, number of distinct words, ratio of the number of distinct words to the total number of words, number of words that occur more than once in a sentence, and ratio of the number of words that occur more than once to the total number of words.

[4] In [3], the authors group punctuation features as a part of the syntactic ones.

[5] A token is considered to be a string of characters between two spaces, or between a space and a punctuation mark.

Table 3. Messages with confusing or multiple sentiments

Body	Label
Ozb, kako, gde? :) Ajd vazi, posalji link. Nisam znao :/ Spic braso	POS
Svasta :/ Zao mi je sto si se namucila :) Hvala ti...samo, ne znam koje drugo postoji :/ :)	NEG

Table 4. English translation of the messages from Table 3

Body
Really, how, where? :) OK, send me the link. I did not know :/ Top bro'
Nonsense :/ I am sorry that you put so much effort :) Thanks...but, I do not know which other is there :/ :)

Syntactic Features. This group, as suggested in [3], contains emoticons and abbreviations.

- **Emoticons.** It is expected that emoticons have the highest influence on the impression about the mood of a message sender. Therefore, 102 emoticons in the form of regular expressions in Python are listed. Absolute counts of each emoticon occurrence per message were added as a single feature. Each emoticon was also assigned to one of the arbitrarily formed nine groups: smiley, happy, sad, surprised, kiss, wink, tongue, skeptic, miscellaneous. This information is also included in the set of features, i. e., nine additional features were added as aggregated counts of each emoticon type (e. g., total number of smiley emoticons, total count of all happy emoticons in a message etc.) The full list of emoticons is available on-line.[6]

- **Abbreviations.** Despite messages being mostly in Serbian, they often contain abbreviations commonly used in texting worldwide (short word forms, slang words, etc.). Therefore, a list of 135 common slang abbreviations in Serbian and 297 in English was compiled. Absolute count of each abbreviation occurrence is then used as a single feature per short message instance. Some of them are: *pozz* (orig. pozdrav, eng. greeting), *pls* (orig. please), *tnx* (orig. thanks), *k* (orig. OK), *msm* (orig. mislim, eng. I think), *vrv* (orig. verovatno, eng. probably), *stv* (orig. stvarno, eng. really), *mzd* (orig. možda, eng. perhaps), *nmp* (orig. nemam pojma, eng. I have no idea) etc. The full

[6] Full list of emoticons: `https://github.com/Branislava/sms_fingerprint/blob/master/features_extraction/emoji.py`

list of abbreviations is available on-line.[7]

Stylistic Features. The smallest group of features (total of 6) was selected after careful human analysis of the dataset, as it seemed that these features could help with differentiating formal and informal tone in messages. These features observe spaces after punctuation, whether a sentence starts with a word in lowercase, wrong punctuation such as `..` and `??`, words starting with "ne" (a likely typo, as negations in Serbian should be separated from a verb) and whether vowels repeat as in *oooook*.

For this particular dataset, we extracted the full list of features using the Web service described in the latter text. The dataset is available as a CSV file containing values for 621 features.[8]

5. Web service. We developed a Web service and a corresponding Web interface, since these features are often used for many tasks, especially in tasks of Sentiment Classification [5, 16, 26] and Authorship Identification [24]. The code was written in Python, and RESTful request dispatching was implemented using Flask micro-framework.[9] Most of the features are represented with corresponding regular expressions.

Web service for feature extraction can be used by sending POST requests to URIs listed in Table 5.[10] Body of a request should be a JSON string containing text to be classified as a value of key named *data*, and when features exist for Serbian and English, then the JSON object should also contain *lang_list* key.

Table 5. Feature Extraction using Web API

URI	Lang	Description
`/char_based_features`	no	Char-based lexical features
`/word_based_features`	no	Word-based lexical features
`/emoticon_features`	no	Emoticon syntactic features
`/abbreviation_features`	yes	Slang abbreviation syntactic features
`/stylistic_features`	no	Stylistic features
`/functionword_features`	yes	Counts of function words

For example, in order to extract emoticon features, the corresponding Unix *curl* command would be:

[7] Full list of abbreviations: `https://github.com/Branislava/sms_fingerprint/blob/master/features_extraction/language_resources.py`

[8] Extracted features for each message from the dataset: `https://github.com/Branislava/sms_sentiment/blob/master/dataset/sms.csv`

[9] `http://flask.pocoo.org/`

[10] The service is hosted at `http://147.91.183.8:12348/`

```
curl
-d '{"data": "This is a message :-)", "lang_list": ["sr", "en"]}'
-H "Content-Type: application/json"
-X POST http://147.91.183.8:12348/emoticon_features
```

The Web interface is available at `http://features.jerteh.rs/` and is also displayed in Figure 1.

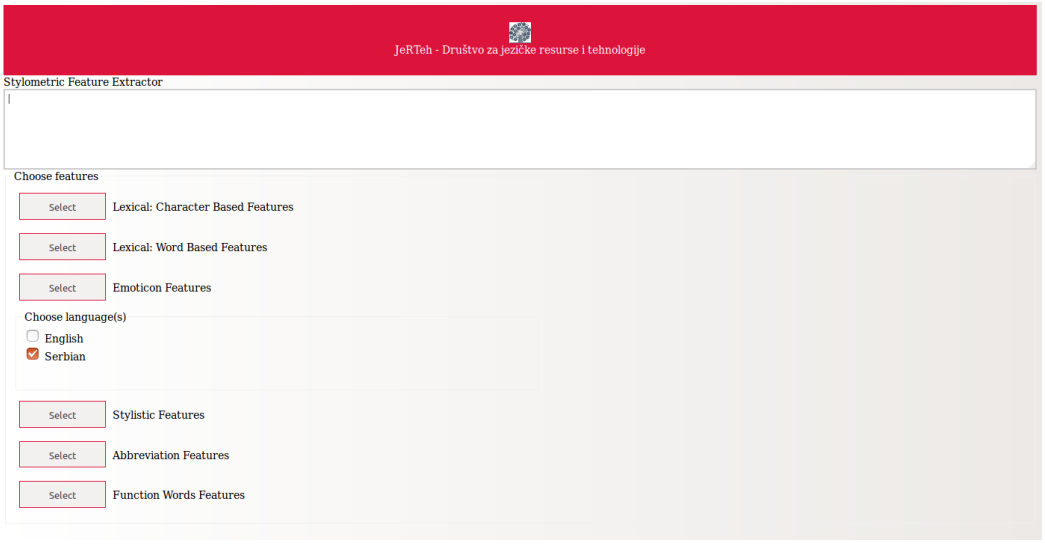

Fig. 1. Web interface for feature extraction

After entering text into the text area, user can select a group of features by clicking on a corresponding *Select* button. As a result a window pops up, with text in JSON having feature names as keys and their counts as values, as shown in Figure 2.

The code was written in Python, and RESTful request dispatching was implemented using the Flask micro-framework.[11]

6. Results and Discussion. Linear Support Vector Machine classifier is selected for this task, with default parameter $C = 1$. All values were normalised first, i. e., they are mapped to the interval $[0, 1]$.

We tried the 5-fold and the 10-fold Cross Validation (CV) settings, and the results were consistent. As evaluation metrics, we used accuracy (Acc), recall (R), precision (P) and F-score (F). The results of these metrics per each fold of a 5-fold

[11] `http://flask.pocoo.org/`

Fig. 2. Resulting JSON of lexical feature counts

CV, in favor of positive (pos), negative (neg) and neutral class (neg) are displayed in Table 6.

Table 6. Evaluation results in a 5-fold CV setting

Fold	**Acc**	**R** (pos)	**P** (pos)	**F** (pos)	**R** (neg)	**P** (neg)	**F** (neg)	**R** (neu)	**P** (neu)	**F** (neu)
1	.930	.901	.956	.928	.667	.963	.788	.968	.909	.938
2	.918	.887	.948	.916	.622	.920	.742	.961	.896	.928
3	.936	.912	.966	.938	.742	.852	.793	.969	.914	.941
4	.926	.883	.960	.920	.611	.880	.721	.975	.907	.940
5	.907	.878	.934	.905	.486	1.0	.655	.953	.885	.918
Avg	**.923**	.892	.953	**.921**	.626	.923	**.740**	.965	.902	**.933**

For easier analysis of the model performance, a confusion matrix obtained in a randomly selected iteration is shown in Figure 3.

It can be seen that the model suffers from dataset imbalance. Seventeen messages with negative sentiments were classified as neutral. Also, a noticeable number of errors occurred when positive samples are classified as neutral (70). In Table 7 we list some of the messages that were misclassified and then we try to explain why this happened.

In the case of message (1), the content means something positive (English translation would be "Good work"), but since there is no punctuation or emoticon, this message was classified as neutral. It is similar with message (2)—it translates as "How is the little baby", but this is very problematic case, because it is just a

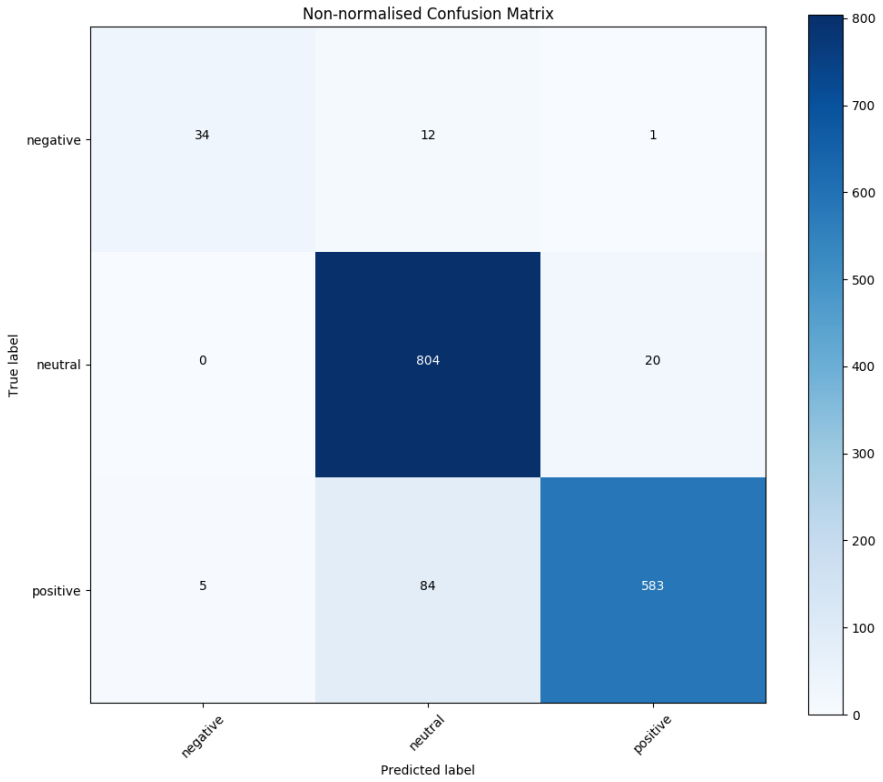

Fig. 3. The confusion matrix

question and it can be considered neutral. In the case of message (3), it was not annotated well by a human, since this message has a negative content (what can be concluded after the sad emoticon).

In the case of message (4), its sentiment is ambiguous and it is not surprising that the classifier got confused, since this message could be classified as both, positive and negative. Message (5) contains complaints, but there is also an exclamation mark. Most of the messages that contained exclamation marks were annotated as positive, and this sample was misclassified most probably due to this reason. It is similar with message (8), which is a simple statement, but the exclamation mark added makes it sound positive.

We can compare messages (6), (7) and (9). Message (9) contains ... and it was manually annotated as neutral. Yet, messages (6) and (7), having the similar structure, were manually annotated as negative. So the occurrence of ... is probably common for both the negative and the neutral class.

We also wanted to examine how our manually-handcrafted features influ-

Table 7. Miss-classified messages

#	**Message**	**Predicted label**	**True label**
1	Spic Braso	NEU	POS
2	Kako je bebica?	NEU	
3	tu sam :-(	NEG	
4	Hvala ti...samo, ne znam koje drugo :/ :)	POS	NEG
5	Jao.kako mi je tesko!	POS	
6	Pa bezveze, al sta da se radi...	NEU	
7	Ccc... Kako je bilo na kraju?	NEU	
8	Cao, Nikola je!	POS	NEU
9	Vec kasnim, ljuti se...	NEG	

ence the outcome of the classification. For this purpose, we selected a Gradient Boost ensemble classifier [7], because of the nature of its internal feature-selection algorithm. Gradient boosting is a sequential technique that combines a set of weak learners, usually decision trees, and delivers improved prediction accuracy in an iterative fashion. Trees are added one at a time and a gradient descent procedure is used to minimise the loss when adding new trees. After calculating error or loss, the outcomes predicted correctly are given a lower weight and the ones misclassified are weighted higher, until best instance weights are found. Twenty most influential features (according to this classifier, on the whole dataset) are displayed in the Figure 4.

The most influential lexical character-based features were the ratios of the characters ':', '!' and '?' to the total number of non-space characters. By analysing the dataset, we speculate that the presence of exclamation mark is evident in many messages that contain positive mood, while enumerations (followed by ':') and questions are usually contained in neutral messages.

Presence of emoticons from the sad, smiley or skeptic groups also influenced this classifier, which was one of our main assumptions. These groups of emoticons should be representatives for negative and positive sentiments.

A word-based lexical feature that contains average tokens length was also among the top five influential features. Also based on the observation, informative (neutral) and negative messages are usually shorter and tokens contain no character repetitions, which mimic the sender's utterance.

It can be seen that mood prediction based on a short message is a hard task even for humans. Due to privacy reasons, original, unprocessed dataset is not available online. Extracted features in a format of CSV file that can be read as a data frame into R or Python program, along with the code can be downloaded from `https://github.com/Branislava/sms_sentiment/`.

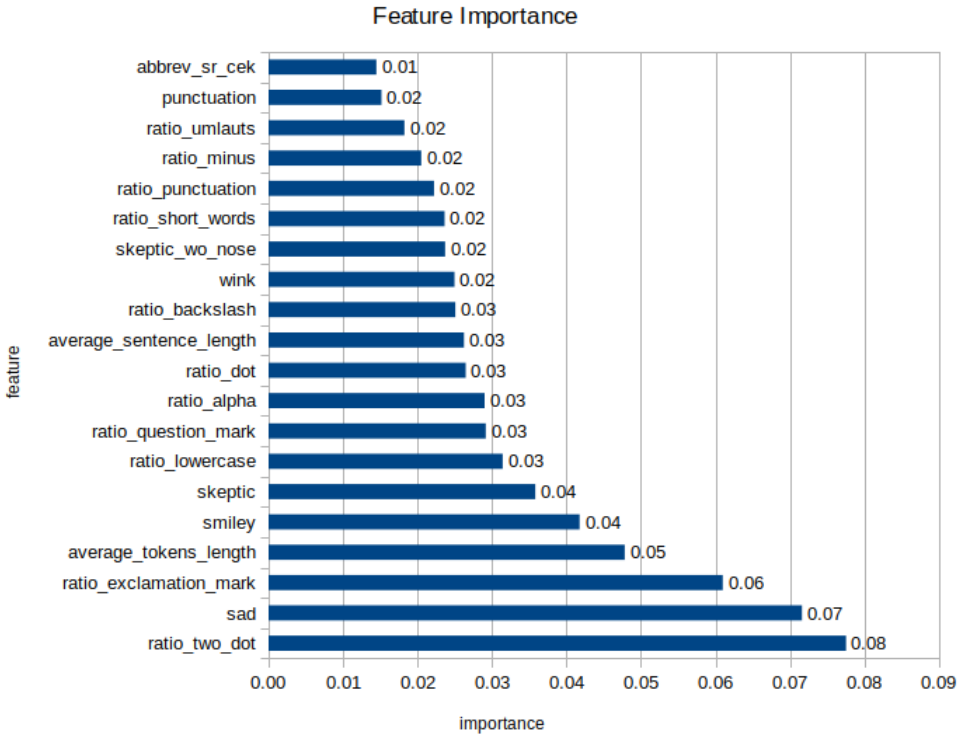

Fig. 4. Feature Importance

7. Conclusion and Future Work. After careful feature construction, it can be concluded that our model had satisfactory performance on this dataset. Many messages contain multiple sentiments and they are very hard both to annotate and to classify. One solution for this would be to perform sentence-based sentiment classification. Another approach would be to perform Emotion Recognition on these messages.[12]

The next experiment will be dedicated to evaluation of the same procedure on different datasets, differing in origin (SMS, Twitter, Facebook, etc.), size and language. The most important contribution of this paper is the non standard approach for specific use case of sentiment analysis. Instead of using predefined lexica, we suggest that for very short texts, when lexica cannot be applied, the distribution of characters themselves should also be taken into consideration.

Acknowledgements. This research was supported by the Serbian Ministry of Science under grant No 178006.

[12]Each message would then contain indicators of presence of certain moods, like anger, surprise, happiness, fear, disgust etc.

REFERENCES

[1] Aleksieva-Petrova A., E. Minkov, M. Petrov. A web application for text document classification based on k-nearest neighbor algorithm. *Serdica Journal of Computing*, **11** (2017), No 2, 183–198.

[2] Andriotis P., A. Takasu, T. Tryfonas. Smartphone message sentiment analysis. In: G. Peterson, S. Shenoi (eds). Advances in Digital Forensics X. *IFIP Advances in Information and Communication Technology*, **433** (2014), 253–265.

[3] Cristani M., G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, V. Murino. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In: Proceedings of the 20th ACM international conference on Multimedia, 2012, 1121–1124.

[4] Davidov D., O. Tsur, A. Rappoport. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics (Posters Volume), 2010, 241–249.

[5] Derks D., A. E. Bos, J. Von Grumbkow. Emoticons and Social Interaction on the Internet: the Importance of Social Context. *Computers in human behavior*, **23** (2007), No 1, 842–849.

[6] Ebert S. Artificial Neural Network Methods Applied to Sentiment Analysis. PhD thesis, Ludwig-Maximilians-Universität München, 2017.

[7] Friedman J. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of statistics*, **29** (2001), No 5, 1189–1232.

[8] Go A., R. Bhayani, L. Huang. Twitter Sentiment Classification using Distant Supervision. In: CS224N Project Report, Stanford, 2009, 1–12.

[9] Inkpen D., F. Keshtkar, D. Ghazi. Analysis and Generation of Emotion in Texts. In: KEPT 2009. International Conference on Knowledge Engineering Principles and Techniques, 2009, 3–13.

[10] Jibril T. A., M. H. Abdullah. Relevance of Emoticons in Computer-Mediated Communication Contexts: An Overview. *Asian Social Science*, **9** (2013), No 4, 201–207.

[11] Kiritchenko S., X. Zhu, S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, **50** (2014), 723–762.

[12] Mishne G. Experiments with Mood Classification in Blog Posts. In: Proceedings of Style 2005. ACM SIGIR 2005 workshop on stylistic analysis of text for information access. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.2693&rep=rep1&type=pdf`, 5 February 2020.

[13] Mladenovic M., C. Krstev, J. Mitrovic, R. Stankovic. Using lexical resources for irony and sarcasm classification. In: BCI '17. Proceedings of the 8th Balkan Conference in Informatics, New York, USA, 2017, Art. 13, 1–8.

[14] Mladenovic M., J. Mitrovic, C. Krstev, D. Vitas. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, **46** (2015), 599–620.

[15] Mukherjee S., A. Malu, A. R. Balamurali, P. Bhattacharyya. Twisent: a multistage system for analyzing sentiment in twitter. In: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, 2531–2534.

[16] Neviarouskaya A., H. Prendinger, M. Ishizuka. Compositionality Principle in Recognition of Fine-Grained Emotions from Text. In: Proceedings of the Third International ICWSM Conference, 2009, 278–281.

[17] Ojamaa B., P. K. Jokinen, K. Muischenk. Sentiment analysis on conversational texts. In: Proceedings of the 20^{th} Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania, No 109, 233–237.

[18] Pavalanathan U., J. Eisenstein. Emoticons vs. Emojis on Twitter: A Causal Inference Approach. `arXiv:1510.08480v1`, 28 October 2015.

[19] Petrovic M. M., N. Ljubešic, D. Fišer. Nestandardno zapisivanje srpskog jezika na tviteru: mnogo buke oko malo odstupanja? *Anali Filološkog fakulteta*, **29** (2017), No 2, 111–136. (in Serbian)

[20] Ptaszynski M., P. Dybala, R. Komuda, R. Rzepka, K. Araki. Development of Emoticon Database for Affect Analysis in Japanese. In: Proceedings of the 4^{th} International Symposium on Global COE Program of the knowledge Federation, 2010, 203–204.

[21] Ptaszynski M., R. Rzepka, K. Araki, Y. Momouchi. Research on Emoticons: Review of the Field and Proposal of Research Framework. In: The Seventeenth Annual Meeting of The Association for Natural Language Processing, 2011, 1159–1162.

[22] Read J. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In: Proceedings of the ACL student research workshop, 2005, 43–48.

[23] Repar A., S. Pollak. Good Examples for Terminology Databases in Translation Industry. In: eLex 2017. The 5^{th} biennial conference on electronic lexicography, Netherlands, 19-21 September 2017, 651–661.

[24] Šandrih B. Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting. In: 3^{rd} International Conference Computational Linguistics in Bulgaria (CLIB 2018), 203–210. `http://dcl.bas.bg/clib2018/wp-content/uploads/2018/07/CLIB_2018_Proceedings_v2_final.pdf`, 27 May 2020.

[25] Šandrih B., D. Vitas. Kvantitativni pregled jezika kratkih poruka. *Naučni sastanak slavista u Vukove dane*, **47** (2018), No 3, 155–165. (in Serbian)

[26] Škorić M. Classification of Terms on a Positive-negative Feelings Polarity Scale Based on Emoticons. *Infotheca: Journal for Digital Humanities*, **17** (2017), No 1, 67–91.

[27] Walther J. B., K. P. D'Addario. The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication. *Social science computer review*, **19** (2001), No 3, 324–347.

Branislava Šandrih
Faculty of Philology
University of Belgrade
3, Studentski trg
11000 Belgrade, Serbia
e-mail: `branislava.sandrih@fil.bg.ac.rs`

Received July 28, 2018
Final Accepted September 26, 2019