DOI: https://doi.org/10.31649/2307-5392-2019-4-35-42

S. D. Shtovba, Dr. Sc. (Eng.), Professor; O. V. Shtovba, Associate Professor, PhD; O. V. Yahymovich; M. V. Petrychko

IMPACT OF THE SYNTACTIC DEPENDENCIES IN THE SENTENCES ON THE QUALITY OF THE IDENTIFICATION OF THE TOXIC COMMENTS IN THE SOCIAL NETWORKS

Social networks often become a medium for threats, insults and other components of cyberbullying. A huge number of people are involved in online social networks, therefore, there is a need for automation of the activities to protect users from anti-social behavior. One of the important tasks of such activity is the identification of the toxic comments that contain threats, insults, obscene etc. The bag of words statistics and bag of symbols statistics are typical features for the toxic comments identification. The effect of syntactic dependencies in sentences on the quality of identification of the social network toxic comments is studied in the article. Syntactic dependences are relationships with proper nouns, personal pronouns, possessive pronouns, etc. 20 syntactic features of sentences have been verified in the total. The article shows that 3 additional specific features significantly improve the quality of toxic comments identification. These three features are: the number of dependences with proper nouns in the singular, the number of dependences that contain bad words, and the number of dependences between personal pronouns and bad words. The experiments are based on data from kaggle- competition "Toxic Comment Classification Challenge". The original kaggle-task of categorizing the toxic comments was modified to the classification one with two alternatives: a neutral comment and a toxic comment. For our experiments, the original dataset with 159751 comments was reduced to 106590 comments due to problems with human-free extraction of the syntactic features. The toxic comment rate is 12.8% in the modified dataset. We use mean of the error rates for each types of misclassification as the metric of quality due to unbalanced dataset. A decision tree is used as a classifier. The decision trees were synthesized for two splitting rules: Gini index and entropy criterion.

Key words: *text mining, natural language processing, syntactic dependencies, toxic comments, social network, identification, machine learning, features selection.*

Introduction

Greater part of the users are involved in the online social networks. Some people address the social network occasionally, others – practically live in them. For certain users social network is a rest, but there are people who do not take decisions without the discussion in the network. According to the portal Statista, in October 2018 six most popular networks took the hurdle of one billion of active users. The most popular is the Facebook, the amount of its active users exceeded 2.2 billion persons. Social networks become the medium for threats, insults and other components of cyberbulling. Taking into account the fact that great number of people are involved in online social networks, there appears a need for automation of the activity to protect the users from antisocial impact. One of the important directions of such activity is the identification in the social networks the toxic comments, containing threats, insults, scorn to others, etc.

Various approaches are used for the automatic identification of the toxic comments, first of all methods of the text mining. The simplest variant is the naïve Bayes classifier on the whole lexicon of the comments from the training set [1]. The bag of words statistics and bag of symbols statistics is often used, i. e., frequencies of words and symbols without taking into account their order in the sentence and relationship between them are calculated. As a rule such features are taken into account: length of the comment, number of capital letters, number of exclamation marks, number of question marks, number of grammar mistakes, number of tokens with nonalphanumeric characters, number of abusive, aggressive and threatening words in the comment, etc. [2]. The more bad words the comment contains, the higher chances to classify it as the toxic one. There appear problems with the statistics of the bad words. The authors of the toxic comments intentionally distort bad words, for instance, instead of *shit*, they write *shiiit*, *sh1t*, *sh!t*, *shi**, *shyt*, *siht*, that is why, the scientists develop special technologies for identification of the hidden offensive words [3, 4] The word order is taken into account by a certain set of the stable word combinations, for instance, by n-grams [5], but this considerably increases the computational complexity of the models construction and does not always reveal the semantics of the comment.

The aim of the research is the study of the impact of the syntactic dependences of the words in the sentence on the quality of the toxic comments identification. Syntactic dependences are relationships with the proper

nouns, personal pronouns, possessive pronouns, etc. Unlike the n-grams method and naive Bayes approach, the model, based on the syntactic dependences is not restricted to the lexicon of the training set. In this model various proper names, personal pronouns, possessive pronouns are allocated in separate groups, i. e., the generalized features are used. If there is another specimen from this group in the test set, this will not influence the modeling. For the allocation of the syntactic dependences we will make use of the information technology from the previous research [6] of one of the coauthors. For the verification of the efficiency we will compare the results of the identification of toxic comments on two sets of features: on typical – on the base of the bag of words statistics and bag of symbols statistics and on the expanded, that additionally contains the statistics of the syntactic dependences. Experiments are carried out on the data of "Toxic Comment Classification Challenge".

Data sets

Data sets "Toxic Comment Classification Challenge" was suggested by the company Jigsaw for kaggle-competition [7]. The data set comprises 159751 text comments. They are written mainly in English [8]. For each comment the membership to six categories of toxicity is indicated: toxic ; severe toxic; obscene; threat; insult; identity hate. The comment may have multiple toxicity, i. e., belong to two, three, even six categories of the toxicity simultaneously (Fig. 1). The comment may be neutral, i. e., does not belong to any category of the toxicity. For instance, the comment *"Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be be banned."is neutral*. The comment *«Hi! I am back again! Last warning! Stop undoing my edits or die!»* is toxic and threatening, the comment *«Would you both shut up, you don't run wikipedia, especially a stupid kid.»* is toxic and insulting. 16225 comments are toxic, the rest – is neutral. The distribution of the comments according to multiplicity of the toxicity is shown in Fig. 2. It is seen that only the comments with high multiplicity of toxicity 5 and 6 seldom occur.

Additionally to typical features on the base of bag of words statistics and bag of symbols statistics we propose a number of specific features, which take into account syntactic dependences between the words in the comment. We have created the corresponding programming module for the parsing of English comments. Programming module is written in Java in the Eclipse environment, using Maven. Specific features were automatically calculated for 106590 comments, this represents 66.8% of the volume of the initial data set. Part of the comments were not processed due to the another language and great number of out-of-vocabulary words. Distortion of the words occurs as a result of the misprints and grammar errors. There are many cases of the intentional distortion of the words to the phonetically similar forms. For this purpose English letter combinations *oo* are changed into *u*, *for* – for *4*, *too* – for *2,* etc. Another variant – intentional distortion to visually similar forms, such as *5h1t*, *b!tch*, *b1tch*. Such distortion occurs by replacing of visually similar symbols: *i* and *1, i* and *!*, *S* and *5* etc.

In a new data set the frequency of the neutral comments slightly reduced – from 89.8% to 87.2%. Distribution by the categories of the toxicity did not change greatly (Table 1).

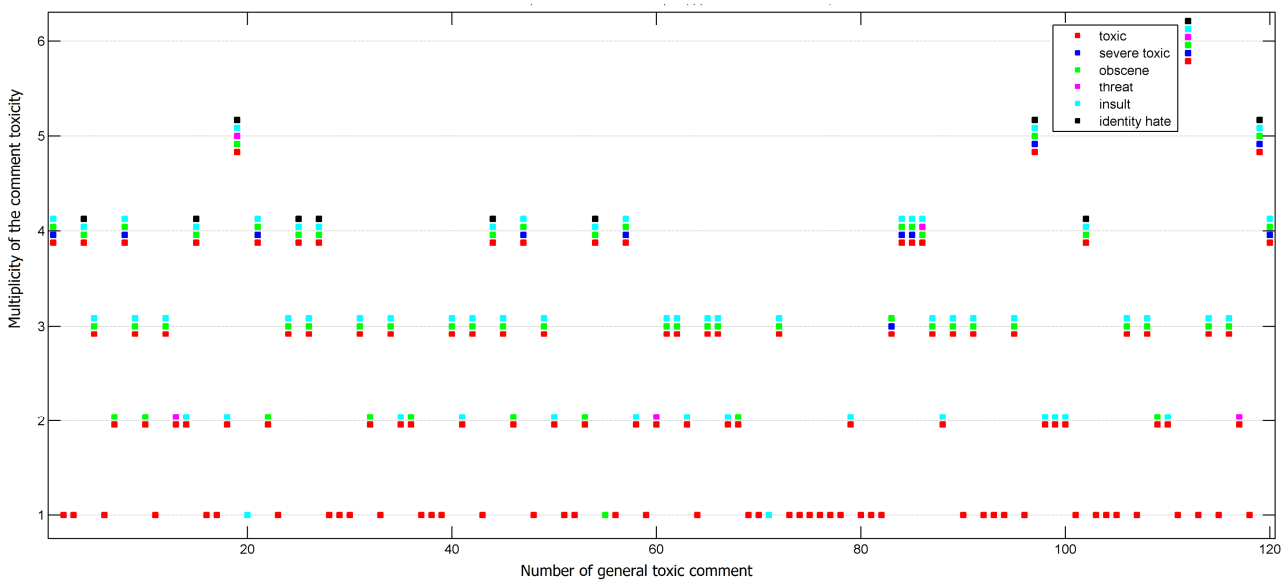

Fig. 1. Categorization of the first 120 toxic comments

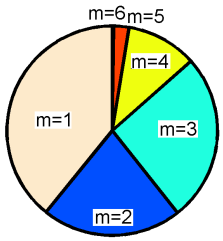

Fig. 2. Distribution of the multiplicity (m) of the toxic comments

Table 1

Distribution by the categories of the toxicity

Category	Comments in the initial data set	Comments in the new data set	Share of the initial data set, %
Toxic	15294	12948	84.7
Severe toxic	1595	1492	93.5
Obscene	8449	7303	86.4
Threat	478	442	92.5
Insult	7877	6943	88.1
Identity hate	1405	1251	89

Features of the comments and the metric of the identification quality

Formalized description of each comment will be performed by means of the following features:

x_1 – number of words;

x_2 – number of unique words;

x_3 – share of the unique words;

x_4 – number of the tokens without taking into account the stop-words;

x_5 – number of grammar errors;

x_6 – number of uppercast words;

x_7 – share of uppercast words;

x_8 – length of the comment;

x_9 – number of capital letters;

x_{10} – number of exclamation marks;

x_{11} – number of question marks;

x_{12} – number of punctuation marks;

x_{13} – number of masking symbols *, &, $, %.

x_{14} – number of smile signs;

x_{15} – share of exclamation marks;

x_{16} – share of question marks;

x_{17} – share of spacings;

x_{18} – share of capital letters;

x_{19} – share of small letters;

x_{20} – number of the comment words, which are in the list of the suspicious words on the site https://www.cs.cmu.edu/~biglou/resources/bad-words.txt;

x_{21} – number of the comment words, which are in the abusive words list on the site http://www.bannedwordlist.com;

x_{22} – number of the comment words which are in the ban-list of the Facebook on the site https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/;

x_{23} – number of the comment words, which are in the ban-list of Google on the site https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/;

x_{24} – number of the comment words, which are in the list of the suspicions words on the site https://gist.github.com/ryanlewis/a37739d710ccdb4b406d;

x_{25} – number of the comment words, which are in the combined dictionary of bad words from five mentioned-above lists;

x_{26} – number of the dependences with the proper names in singular;

x_{27} – number of the dependences with the proper names in plural;

x_{28} – number of the dependences with the personal pronouns;

x_{29} – number of the dependences with the possessive pronouns;

x_{30} – number of the dependences with the negations (with the words never or not);

x_{31} – number of the dependences with the negations, where proper names in singular are used;

x_{32} – number of the dependences with the negations where proper names in plural are used;

x_{33} – number of the dependences with the negations, where personal pronouns are used;

x_{34} – number of the dependences with the negations, where possessive pronouns are used;

x_{35} – number of the dependences between the proper names in singular and words, used in the connections with the negations;

x_{36} – number of the dependences between the proper names in plural and words, used in the connections with the negations;

x_{37} – number of the dependences between the personal pronouns and words, used in the connections with the negations;

x_{38} – number of the dependences connections between the possessive pronouns and words, used in the connections with the negations;

x_{39} – number of the dependences , where bad words are used;

x_{40} – number of the dependences with the negations, where bad words are used;

x_{41} – number of the dependences between the proper names in singular and bad words;

x_{42} – number of the dependences between the proper names in plural and bad words;

x_{43} – number of the dependences between the personal pronouns and bad words;

x_{44} – number of the dependences between the possessive pronouns and bad words;

x_{45} – number of the dependences between the pronouns and bad words.

The feature x_{13} takes into account the presence of the symbols *, &, $, %, which are sometimes used to replace the separate letters for masking the bad words, for instance, a$$, $hit etc. The need to take into account such symbols is stipulated by the fact that the users quickly generate new variants of the distorted swear words, which fail to enter the corresponding dictionaries.

Specific features x_{26} - x_{45} are studied for the first time for the problem of the identification of the toxic comments. For the verification of the importance the new features, the initial kaggle-problem of the comments categorization we will reduce to the classification problem with two classes. The first class – neutral comment and the second class – toxic comment. The data set is unbalanced – the ratio between classes is approximately 9 to 1. That is why, it is not expedient to verify the quality of the identification by the frequency of the classification errors. As the metric of the identification quality we will apply the mean value of the frequencies of the classification errors of each type:

$$Q_{aver} = \frac{P_{12} + P_{21}}{2},$$

where P_{12} – is the frequency of the errors of 1→2 types, when the neutral comment is recognized as toxic; P_{21} – is the frequency of the errors of 2→1 type, when toxic comment is recognized as neutral.

Q_{aver} – is a simple metric for the verification of the classifiers on the unbalanced data set It is suitable for the investigation problem i. e., for the determination of the expediency of taking into consideration the syntactic dependences in the sentences for the synthesis of the toxic comments classifiers.

Experimental studies

The classifier is realized in the form of the decision tree. Our choice is stipulated by the following reasons: 1) synthesis of the decision tree even on the large data sets is rather fast, this enables to carry out numerous experiments; 2) in the process of the synthesis of the decision tree the feature selection is carried out, this enables to verify their expediency. The data set will be divided into training and testing ones. In the test set each sixth comment will be included, the rest of the comments will be included into the training set Thus, the testing set contains 17765 comments and training set– 88825. Decision trees are synthesized on the training set and it will be prune so that to minimize Q_{aver} on the testing set . The studies will be carried out on two sets of the input features: typical – x_1 - x_{25} and expanded – x_1 - x_{44}.

For equalizing the balance we will perform the sampling of the training set increasing the weight of the observations of the minor class. It should be taken into account, that the correct classification of the comment with the multiple toxicity is more important than the comment, that belongs only to one toxic category. Weight w of the toxic comment C is suggested to define in such heuristic way:

$$w(C) = b + \sqrt{m(C)},$$

where b – is the basic weight of the toxic comment; $m(C) \in \{1, 2, ..., 6\}$ – is multiplicity of the C comment toxicity.

Fig. 3 shows the experimental dependences of the identification quality on the basic weight of the toxic comment. During the experiments decision trees were synthesized according to two splitting rules: on the base of the index of Gini and entropy criterion. The experiments proved, that better trees are synthesized by the Gini index. Small values of Q_{aver} are achieved when the basic weight of the toxic comment takes the values from 4.5 to 5.8. Minimum Q_{aver}=0.118 is achieved, when $b \in [5.2, 5.5]$. The frequency of the errors of the best decision tree in the whole test set is 0.0987, if $P_{12} = 0.0919$ and $P_{21} = 0.1442$.

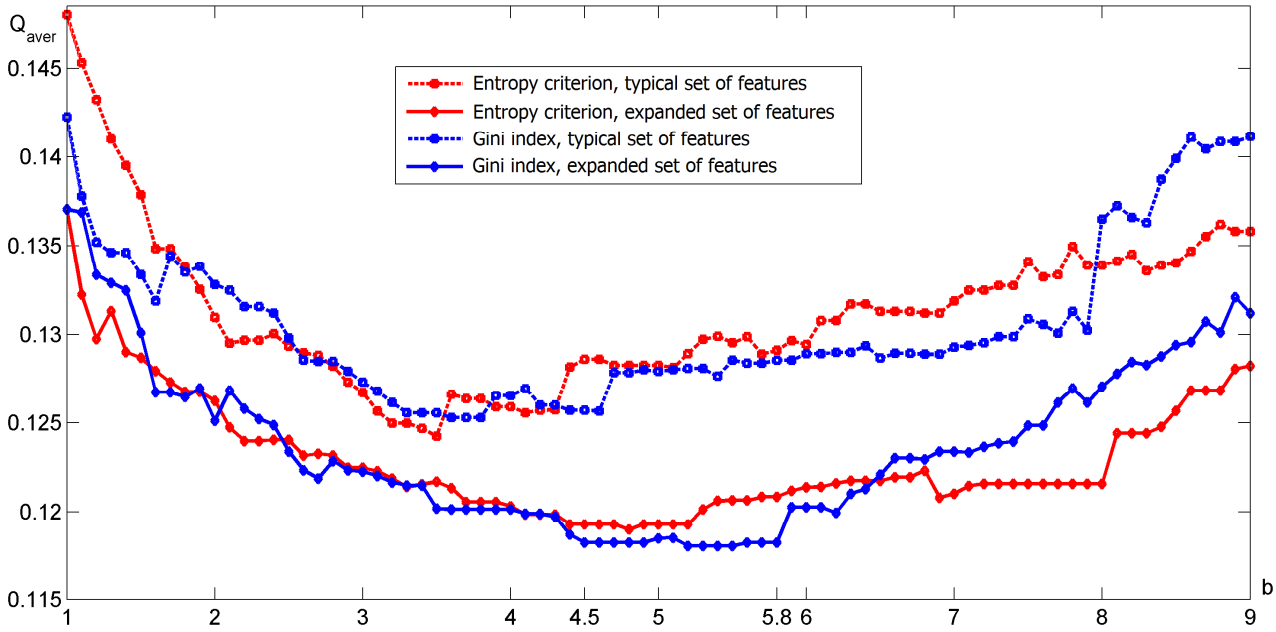

Fig. 3. Dependence of the decision tree quality on the basic weight of the toxic comment

Fig. 3 shows that the expanded set of features greatly improves the quality of identification. Verification of the five best decision trees showed that all of them use such features as: x_3-x_9, x_{15}, x_{17}-x_{19}, x_{22}, x_{24}-x_{26}, x_{39} and x_{43}. Four of five trees additionally use feature x_1. These features are the most informative. There are three new specific features among them: x_{26} – number of the dependences with the proper names in singular, x_{39} – number of the dependences where bad words are used and x_{43} – number of the dependences between personal pronouns and bad words.

Four less important features are used by several best decision trees. Typical features x_2, x_{10} and x_{12} are used by two out of five best models. Specific feature x_{28} – number of the dependences with personal pronouns is used by ONE model. These additional four features can be used for the synthesis of more complex models for detecting toxic comments.

Conclusions

The problem of the categorization of the comments in the social networks for detecting toxic texts is considered. The sample of the comments from kaggle-competition “Toxic Comment Classification Challenge” is taken as the experimental data. For detecting toxic comments the typical set of features on the base of the bag of words statistics and corresponding dictionaries of the bad words is used. The effect from the additional accounting of the specific features is verified in the paper. The additional set was created by 20 specific features, which describe syntactic dependences between the words in the comment.

It is determined that taking into account of the specific features enables to improve considerably

the quality of the identification of toxic comments. Among the suggested 20 specific features the following 3 features are the most informative: number of the dependences with the proper names in singular, number of the dependences , where the bad words are used and number of the dependences between personal pronouns and bad words. The selection of 3 specific features enables to reduce considerably the computational complexity of the syntactic parsing of the comment, because the calculation of all 20 specific features requires a lot of resources. Accordingly, if three specific features are added to the typical set, then the identification of the toxic comments can be realized in real time.

For the improvement of the reliability of the toxic comments identification it is expedient to verify the effect of the replacement of the simple search of the bad words by special technologies, aimed at revealing of the masked insulting words [5] on the base of fuzzy similarity measures and Levenshtein distance. Promising is also the combining of the suggested models on the base of the statistic analysis of the text with the models of other types, which take into consideration the features of cooperation and general activity of the author of the comment [9]. In particular, it is expedient to verify the effect of taking into consideration such features of cooperation as: 1) distribution of the number of the answers (or other reactions) on the remarks of the specific participants of the discussion; 2) distribution of the number of the comments of one and the same participant of the discussion; 3) similarity of the name, registration time, IP-address and e-mail address of the author of the comment with the corresponding attributes of other users; 4) correlation of the activity of the author of the comment with other participants. Also it is expedient to verify the effect from taking into consideration such indices of cooperation as: 1) distribution of the duration of the user's reaction – time, during which his assessment of the comment, response to the address, etc. appears; 2) duration of the stay in the social network, volume of the created content and number of the assessing actions; 3) regularity of the stay in the network of the author of the comment.

The paper is written by the results of the realization of the state-sponsored scientific research program 46–Д–388 «Identification of the hidden dependences in the online social networks, based on the methods of fuzzy logic and computational linguistics».

REFERENCES

1. Fine-Grained Classification of Offensive Language / J. Risch, E. Krebs, A. Löser [et al.] // Proc. of GermEval 2018, 14th Conference on Natural Language Processing. Vienna, Austria, 2018. – P. 38 – 44.

2. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media/ J. Salminen, H. Almerekhi, M. Milenković [et all] // Twelfth International AAAI Conference on Web and Social Media. – 2018. – P. 330 – 339.

3. Srivastava S. Identifying Aggression and Toxicity in Comments using Capsule Network / S. Srivastava, P. Khurana, V. Tewari // Proc. of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). – 2018. – P. 98 – 105.

4. Sood S. O. Using Crowdsourcing to Improve Profanity Detection / S. O. Sood, J. Antin, E. F. Churchill // Association for the Advancement of Artificial Intelligence. Spring Symposium: Wisdom of the Crowd. – 2012. – P. 69 – 74.

5. Mohammad F. Is preprocessing of text really worth your time for toxic comment classification? / F. Mohammad // Proc. of Inter. Conference on Artificial Intelligence. CSREA Press. – 2018. – P. 447 – 453.

6. Bisikalo O. Development of the method for filtering verbal noise while search keywords for the English text / O. Bisikalo , A. Yahimovich , Y. Yahimovich // Technology Audit and Production Reserves. – 2018. – № 6. – P. 33 – 41.

7. Toxic Comment Classification Challenge. Available [Electronic resource] / Access mode : https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

8. Stop Illegal Comments: A Multi-Task Deep Learning Approach [Electronic resource] / A. Elnaggar, B. Waltl, I. Glasera [et all] // Software Engineering for Business Information Systems, Technische Universitat¨ Munchen, Germany. – 2018. – Access mode : https://arxiv.org/pdf/1810.06665.pdf.

9. Kumar S. Antisocial Behavior on the Web: Characterization and Detection / S. Kumar, J. Cheng, J. Leskovec // Proceedings of the 26th International Conference on World Wide Web Companion. – International World Wide Web Conferences Steering Committee. – 2017. – P. 947 – 950.

Editorial office received the paper 24.12.2018.
The paper was reviewed 18.02.2019.

Shtovba Serhiy – Dr. Sc. (Eng.), Professor, Professor with the Computer Control Systems Department, e-mail: shtovba@vntu.edu.ua.

Shtovba Olena – Associate Professor, PhD, Associate Professor with the Department of Management, Marketing and Economics, e-mail: olena.shtovba@yahoo.com.

Yahymovych Olexandr – PhD-student, Automation and Intelligent Information Technology, yahimovich.olexandr@gmail.com.

Petrychko Mykola – Student, Department of Computer Systems and Automation, petrychko.myckola@gmail.com.

Vinnytsia National Technical University.