Detection of Social Network Toxic Comments with Usage of Syntactic Dependencies in the Sentences

Serhiy Shtovba[0000-0003-1302-4899], Olena Shtovba[0000-0003-1418-4907],
Mykola Petrychko[0000-0001-6836-7843]
Vinnytsia National Technical University, Khmelnytske Shose, 95, Vinnytsia, 21021, Ukraine
shtovba@vntu.edu.ua
olena.shtovba@yahoo.com
petrychko.myckola@gmail.com

Abstract. Social networks sometimes become a medium for threats, insults and other components of cyberbullying. A huge number of people are involved in online social networks. Hence, a protection of network users from anti-social behavior is an important activity. One of the major tasks of such activity is automated detecting the toxic comments with threats, insults, obscene etc. The bag of words statistics and bag of symbols statistics are typical features for the toxic comments detection. The effect of syntactic dependencies in sentences on the quality of detection of the social network toxic comments is studied in the article for the first time. Syntactic dependences are relationships with proper nouns, personal pronouns, possessive pronouns, etc. Twenty syntactic features of sentences have been verified in the total. The paper shows that 3 additional specific features significantly improve the quality of toxic comments detection. 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". 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. 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 deviance criterion.

Keywords: natural language processing, syntactic dependencies, toxic comments, social network, machine learning, features selection, balanced accuracy, decision tree.

1 Introduction

Social networks sometimes become a place for threats, insults and other components of cyberbullying. A huge number of people are involved in online social networks. Hence, a protection of network users from anti-social behavior is an important activity. One of the major tasks of such activity is automated detecting the toxic com-

ments. Toxic comments are textual comments with threats, insults, obscene, racism etc.

The various techniques are used for human-free detecting the toxic comments. Bag of words statistics and bag of symbols statistics are typical source information for the toxic comments detection. Usually the following statistics-based features are used: length of the comment, number of capital letters, number of exclamation marks, number of question marks, number of spelling errors, number of tokens with non-alphabet symbols, number of abusive, aggressive, and threatening words in the comment, etc. [1]. High count of bad words in the comment increases a chance to classify it as toxic. However, there are some difficulties with usage of the bad words statistics. Some out-of-vocabulary words are produced by typos and by spelling errors. Often authors of toxic comments distort their bad words purposely. They convert the bad words to phonetically identical forms by replacing letter combinations *oo* to *u*, *for* to *4*, *too* to *2* etc. Another variant is to distort to visual similar forms, for example, *5h1t*, *b!tch*, *b1tch*. Scientists develop special technologies for detecting the masked bad words [2, 3], but vandals have a reserve in time and in persons. In addition to analyzing the separated keywords, some methods take into account the order of the words in sentences. For example, authors of [4, 5] used n-grams-based approach, but such modeling does not reflect the whole relations in sentences.

The aim of the paper is to study an effect of syntactic dependencies in sentences on the quality of detecting the social network toxic comments. Syntactic dependences are relationships with proper nouns, personal pronouns, possessive pronouns, etc. Opposite to n-gram method and naive Bayesian approach, the model based on the syntactic dependencies does not directly tie with the training set vocabulary. All the various proper names, personal pronouns, possessive pronouns are allocated into separate groups. It allows to use the vocabulary-free generalized features in the model. Another instance from this group in the test set will not affect the simulation negatively. We use the information technology from [6] for extraction the syntactic features from the data set. We compare the results of toxic comments detection on two sets of features. The first set is typical features that based on bag of words statistics and bag of symbols statistics. The second one is extended set that contains typical features together with syntactic features. The experiments are performed on the “Toxic Comment Classification Challenge” data set.

2 Data sets and preprocessing

Data set “Toxic Comment Classification Challenge” is collected by Conversation AI team, a research initiative founded by Jigsaw and Google, both a part of Alphabet. The data set is used in kaggle-competition [7]. The data set consists of 159751 Wikipedia comments which have been labeled by human raters for toxic behavior. Most of the comments are English [8].

Each comment is manually categorized with 6 binary labels: toxic, severe toxic, obscene, threat, insult, and identity hate. Some comments have toxic multiplicity. In this case a comment belongs to 2, 3, and even 6 toxic categories simultaneously (Fig-

ure 1). Also a comment may be neutral, i.e. it does not belong to any toxic category. For example, the following comment *"Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be banned."* is neutral. Comment *"Hi! I am back again! Last warning! Stop undoing my edits or die!"* is toxic and threated, and comment *"Would you both shut up, you don't run Wikipedia, especially a stupid kid."* is toxic and insult.

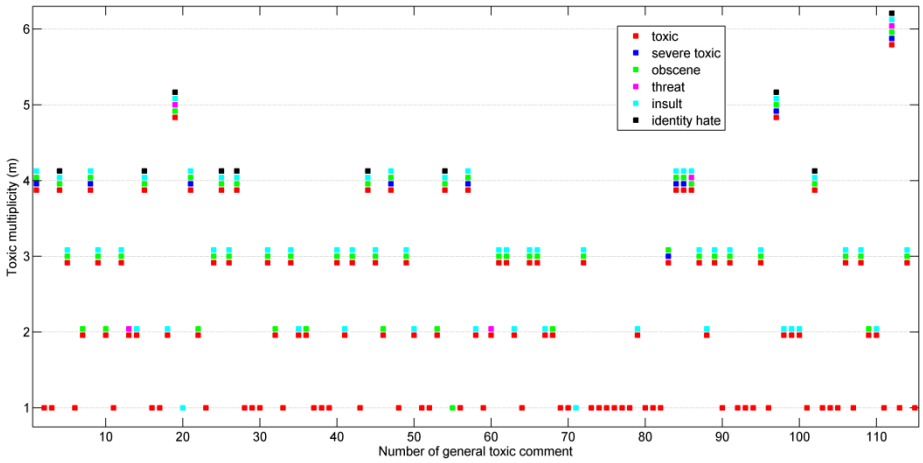

Fig. 1. Categories of the first 115 non-neutral comments

16225 comments have the toxic labels. The rest of the comments are neutral. A distribution of the comments on toxic multiplicities is presented on Figure 2. It shows that only comments with high toxicity multiplicity are rarely encountered. Most of toxic comments (60.8%) belong to several toxic categories *(m*>1).

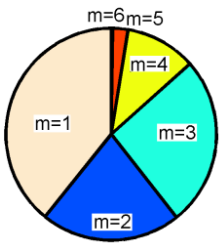

Fig. 2. Distribution of multiplicity (*m*) of toxic comments

Figure 3 shows the combinations of toxic categories in one comment. The set of toxic comments with the same category is represented by a color square. Each toxic category represents the corresponding color. The area of the square equals to the number of comments with the same toxic category. The intersection of squares reflects the number of comments that belong to two relevant toxic categories simultaneously. Figure 3 shows that all the severe toxic comments also belong to toxic cate-

gory – the blue square is completely inside the red square. Also, almost all the severe toxic comments are obscene and insult. There are 3 very low intersecting categories: severe toxic, threat, and identity hate. Few comments belong simultaneously to two out these three categories. Figure 3 also shows the degree of similarity for two finite sets in form of Jaccard index (k_i). It is calculated as the cardinality of the intersection of the sets divided by the cardinality of the union of the sets. For our case Jaccard index corresponds to the ratio of the area of intersection of two squares over the area of the union of two squares.

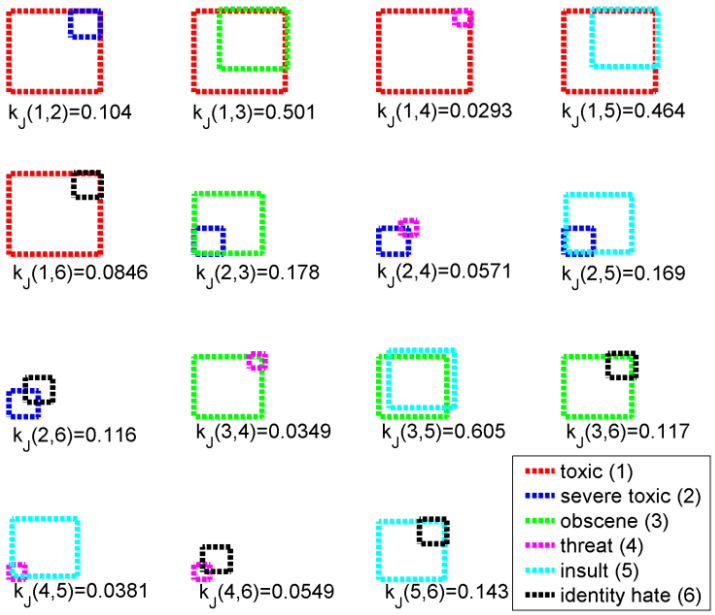

Fig. 3. Jaccard similarity indexes for various toxic categories

We propose to add several specific features to the typical feature set that based on statistics of a bag of the words and statistics of a bag of the symbols. The specific features are taking into account some syntax dependencies between words in comment. The specific features extraction was done using the technology from [6]. The specific features were extracted automatically for 106590 comments. Features extraction for some comments was unsuccessful due to non-English text and out-of-vocabulary words. As a result, the modified data set consists 66.8% of the source data set. Neutral comments compose 87.2% of the modified data set. It is slightly less than in the source data set where the neutral ratio is 89.8%. Distributions of the comments on toxic categories are almost equal for two data sets (Table 1).

Table 1. Source data sets and modified data sets

Category	Comments in source data set	Comments in modified data set	Share of source 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

3 Features and quality metric

The following features are used for formalized description of each comment:

x_1 is a number of words;

x_2 is a number of unique words;

x_3 is a ration of unique words;

x_4 is a number of tokens without the stop-words;

x_5 is a number of spelling errors;

x_6 is a number of all-caps words;

x_7 is a ratio of all-caps words;

x_8 is a length of the comment;

x_9 is a number of capital letters;

x_{10} is a number of explanation marks;

x_{11} is a number of question marks;

x_{12} is a number of punctuation marks;

x_{13} is a number of masking symbols (*, &, $, %);

x_{14} is a number of happy smiles;

x_{15} is a ratio of explanation marks;

x_{16} is a ratio of question marks;

x_{17} is a ratio of spaces;

x_{18} is a ratio of capital letters;

x_{19} is a ratio of lowercase letters;

x_{20} is a number of the comment's words that included into the bad word list at https://www.cs.cmu.edu/~biglou/resources/bad-words.txt;

x_{21} is a number of the comment's words that included into the swear word list at http://www.bannedwordlist.com;

x_{22} is a number of the comment's words that included into facebook black list at https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/;

x_{23} is a number of the comment's words that included into google blacklist at https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/;

x_{24} is a number of the comment's words that included into the naughty word list at https://gist.github.com/ryanlewis/a37739d710ccdb4b406d;

x_{25} is a number of the comment's words that included into 5 mentioned lists;

x_{26} is a number of dependencies with proper nouns in the singular;

x_{27} is a number of dependencies with proper nouns in the plural;

x_{28} is a number of dependencies with personal pronouns;

x_{29} is a number of dependencies with possessive pronouns;

x_{30} is a number of dependencies with denial (with words never or not);

x_{31} is a number of dependencies with denial that contain proper nouns in the singular;

x_{32} is a number of dependencies with denial that contain proper nouns in the plural;

x_{33} is a number of dependencies with denial that contain personal pronouns;

x_{34} is a number of dependencies with denial that contain possessive pronouns;

x_{35} is a number of dependencies between proper nouns in the singular and the words from dependencies with denial;

x_{36} is a number of dependencies between proper nouns in the plural and the words from dependencies with denial;

x_{37} is a number of dependencies between personal pronouns and the words from dependencies with denial;

x_{38} is a number of dependencies between possessive pronouns and the words from dependencies with denial;

x_{39} is a number of dependencies that contain the bad words;

x_{40} is a number of dependencies with denial that contain the bad words;

x_{41} is a number of dependencies between proper nouns in the singular and the bad words;

x_{42} is a number of dependencies between proper nouns in the plural and the bad words;

x_{43} is a number of dependencies between personal pronouns and the bad words;

x_{44} is a number of dependencies between possessive pronouns and the bad words;

x_{45} is a number of dependencies between pronouns and the bad words.

Twenty specific features $x_{26} - x_{45}$ are examined for toxic comments detection for the first time. Let us modify the original kaggle-task of categorizing the toxic comments to the classification one with two alternatives: a neutral comment and a general

toxic comment. It allows to easy checkup the informative levels of the proposed syntactic features.

The data set is unbalanced with class proportion about 9 to 1. Hence, misclassification rate is not suitable metric for quality of the classifier. According to [9] we use balanced accuracy approach. The metric of quality of the classifier is as follows:

$$Q_{aver} = \frac{P_{nt} + P_{tn}}{2},$$

where P_{nt} denotes probability of $n \rightarrow t$ type classifying errors, when a neutral comment is recognized as a general toxic comment; P_{tn} denotes probability of $t \rightarrow n$ type classifying errors, when a general toxic comment is recognized as a neutral comment. Q_{aver} is mean of probabilities of each type misclassification. It is simple and interpretable metric for examination a classifier on unbalanced data set.

4 Computational experiments

A decision tree is used as a classifier. We choose this kind of classifier taking into account the following reasons: 1) a synthesis of the decision tree is a fast procedure even for large training set, hence, it is possible to carry out several experiments; 2) features selection is carried during the decision tree synthesis; it is easy to check the informative levels of the proposed syntactic features. We divide the data set on training data and test data. The test set consists of every sixth comment. The rest comments are in the training set. Thus, the test set contains 17765 comments and training set contains 88825 comments. We use the training data for decision tree synthesis. After this, the decision tree is pruned for minimization Q_{aver} on the test set. We check up two sets of the features: typical set – x_1 - x_{25} and extended set - x_1 - x_{44}.

Rebalancing the class distribution is yielded by a sampling in way of increasing the weight of minor class objects. We suppose that correct classification of the comment with high toxic multiplicity is more important than the comment with low toxic multiplicity. Weight w of toxic comment C is defined by the following heuristic formula:

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

where b denotes a bias of toxic comment weight; $m(C) \in \{1, 2, ..., 6\}$ denotes toxic multiplicity of comment C.

Figure 4 shows the dependences of the classifier quality under the bias of toxic comment weight. The decision trees were synthesized with two splitting rules: Gini index-based rule and deviance criterion-based rule. The experiments show that Gini index-based rule provides better decision trees. Q_{aver} is low, when the bias of toxic comment weight belongs to [4.5, 5.8]. Minimal value of Q_{aver} =0.118 is obtained for

$b \in [5.2, 5.5]$. Figure 4 shows that the extended set of features significantly improves the classifier quality.

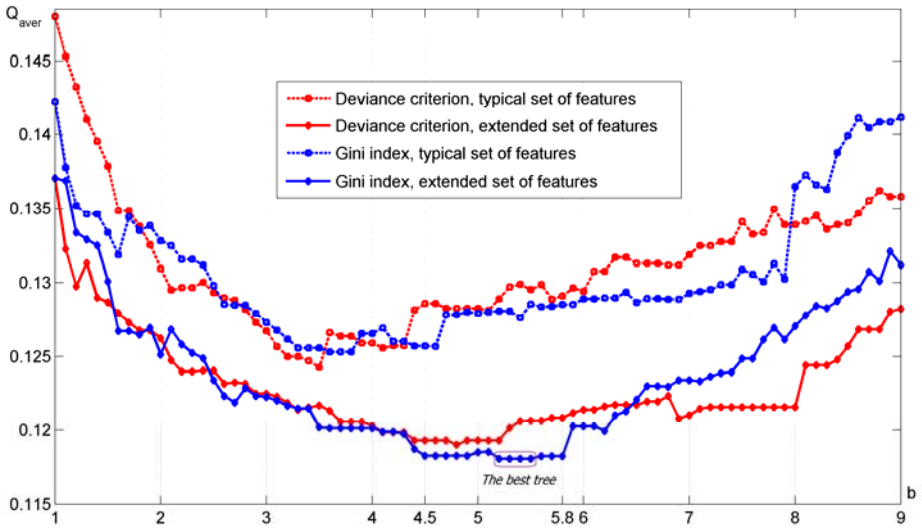

Fig. 4. Experimental dependencies of toxic comments classifier quality

The best model is a decision tree with minimal value of Q_{aver}. The best decision tree is presented on Figure 5. Misclassification rate for the best decision tree is Q=0.0987. The other metrics for the best decision tree are as follows: Q_{aver}=0.118, $P_{nt} = 0.0919$, and $P_{tn} = 0.1442$. The best tree correctly detects almost the all comments with high and average toxic multiplicities (Figure 6). The best tree correctly detects almost all the toxic comments with labels severe toxic, obscene, and identity hate (Figure 7).

Let us analyze 5 best trees. All the trees use the following features: x_3 - x_9, x_{15}, x_{17} - x_{19}, x_{22}, x_{24} - x_{26}, x_{39}, and x_{43}. 4 out 5 trees use feature x_1 additionally. Among their most important features are 3 new syntactic ones: a number of dependencies with proper nouns in the singular (x_{26}); a number of dependencies that contain the bad words (x_{39}) and a number of dependencies between personal pronouns and the bad words (x_{43}).

We also point to 4 following slightly less important features. Typical features x_2, x_{10}, and x_{12} are in 2 out 5 the best trees. Syntactic feature x_{28} is selected for 1 out 5 the best trees. The mentioned 4 extra features may be used for more complicated models for toxic comment detection.

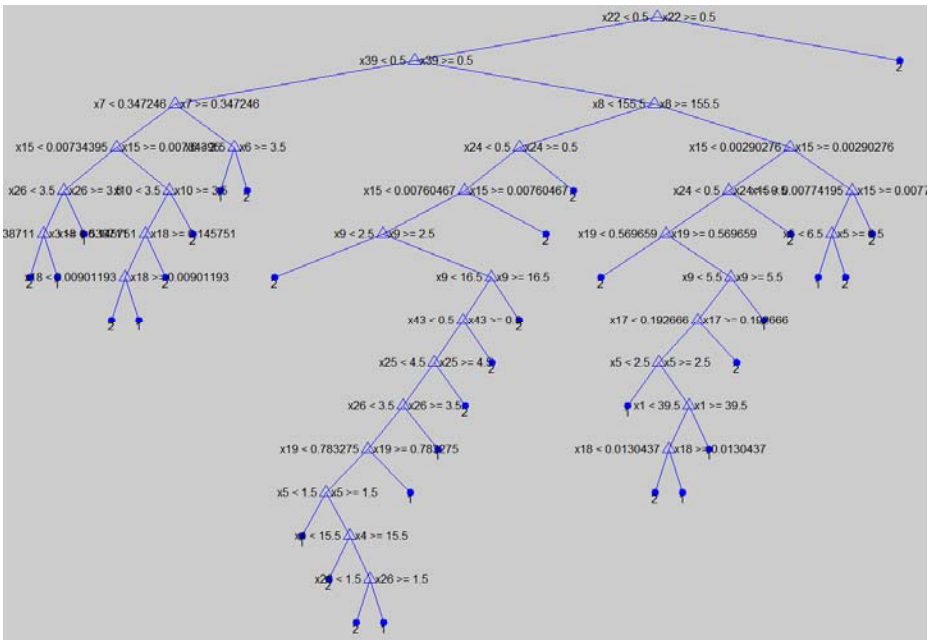

Fig. 5. The best decision tree (1 – neutral comment, 2 – general toxic comment)

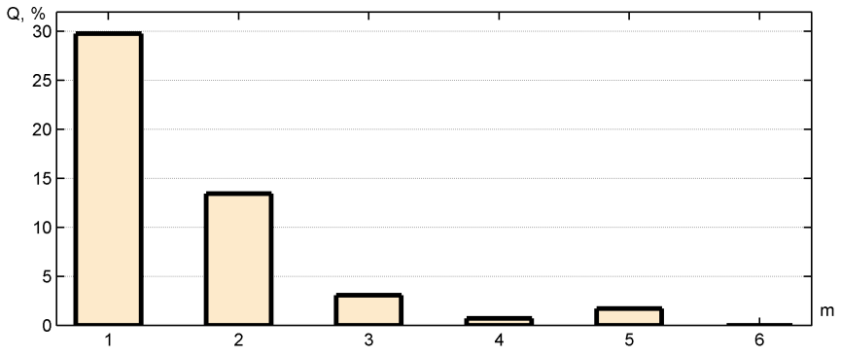

Fig. 6. Distribution of misclassification rate (Q) on the comments with various toxic multiplicity (m)

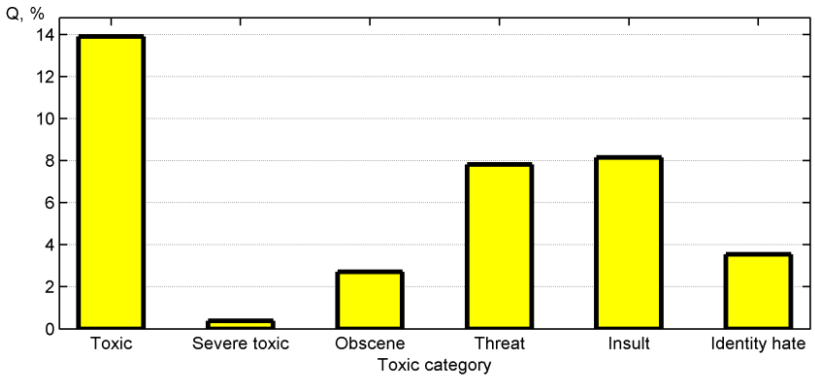

Fig. 7. Distribution of misclassification rates (Q) on various toxic categories

5 Conclusion

The problem of detecting the toxic comments in social networks was considered. For our experiments we used kaggle data set "Toxic Comment Classification Challenge". The bag of words statistics and bag of symbols statistics are typical features for detecting the toxic comments. The effect of syntactic dependencies in sentences on the quality of the social network toxic comments detection was studied in the article. Syntactic dependences are relationships with proper nouns, personal pronouns, possessive pronouns, etc. In total 20 syntactic features of sentences had been checked.

A novelty of the research consists of the experimental confirmation that 3 additional specific features significantly improve the quality of toxic comments detection. Those 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 selection of 3 specific features allows to significantly reduces the computational complexity of text comment preprocessing, since the calculation of all 20 specific features requires a lot of resources. Accordingly, with 3 specific features added to the typical set, the identification of the toxic comments can be done in real time with good quality.

Acknowledgements. Authors thank Olexandr Yahimovych for extraction the syntactic features from the data set of toxic comments. This research is supported by government scientific project 46–G–388 «Fuzzy logic and computational linguistics based the identification of hidden dependencies in online social networks».

References

1. Salminen, J., et al.: Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceeding of the

Twelfth International AAAI Conference on Web and Social Media, 2018, pp. 330-339 (2018).
2. Srivastava, S., Khurana, P., Tewari, V.: Identifying Aggression and Toxicity in Comments using Capsule Network. In: Proceeding of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 98-105 (2018).
3. Sood, S.O., Antin, J., Churchill, E.F.: Using Crowdsourcing to Improve Profanity Detection. In: Proceeding of Association for the Advancement of Artificial Intelligence. Spring Symposium: Wisdom of the Crowd, 2012, pp. 69–74 (2012).
4. Mohammad, F.: Is preprocessing of text really worth your time for toxic comment classification? In: Proceeding of International Conference on Artificial Intelligence, 2018, pp. 447-453 (2018).
5. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, 2012, pp. 19–26 (2012).
6. Bisikalo, O., Yahimovich, A., Yahimovich, Y.: Development of the method for filtering verbal noise while search keywords for the English text. Technology Audit and Production Reserves. 6(2): 33–41 (2018).
7. Toxic Comment Classification Challenge. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.
8. Elnaggar, A. et al.: Stop Illegal Comments: A Multi-Task Deep Learning Approach. arXiv preprint arXiv:1810.06665 (2018).
9. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In Proceedings of the 20th IEEE International Conference on Pattern Recognition, 2010, pp. 3121-3124 (2010).