ISSN 2007-9737

Offensive Language Recognition in Social Media

Elena Shushkevich[1], John Cardiff[1], Paolo Rosso[2], Liliya Akhtyamova[1]

[1] Technological University Dublin,
Social Media Research Group,
Ireland

[2] Universitat Politècnica de València,
Spain

e.shushkevich@yandex.ru, john.cardiff@tudublin.ie, prosso@dsic.upv.es

Abstract. This article proposes an approach to solving the problem of multiclassification within the framework of aggressive language recognition in Twitter. At the stage of preprocessing external data is added to the existing dataset, which is based on information in the links in dataset. This made it possible to expand the training dataset and thereby to improve the quality of the classification. The model created is an ensemble of classical machine learning models included Logistic Regression, Support Vector Machines, Naive Bayes models and a combination of Logistic Regression and Naive Bayes. The obtained value of macro F1-score for one of the experiments achieved 0.61, which exceeds the state-of-art published value by 1 percentage point. This indicates the potential value of the proposed approach in the field of hate speech recognition in social media.

Keywords. Hate speech, ensemble of models, logistic regression, support vector machine, naive Bayes.

1 Introduction

Nowadays social media has become an important part of people's lives, where everyone can read news, communicate with friends and share their opinions. These advances also bring new challenges and risks associated with new technologies. With the increasing influence of social networks and online discussions, there is a problem of aggressive language in users' messages increasing dramatically. It is no secret that on the Internet, in the territory where there is no real contact, people feel more free and allow themselves to use statements that can bring moral harm to other users.

Aggressive, offensive, and hate speech in social media can relate to various social aspects: the problems of immigration, race, gender, weight, and religion of other people. Hate speech messages often contain insults, but situations also rise when the message does not contain swear words and profanity, but the meaning of the message is offensive and humiliating to a group of people or a particular person.

However, as the intention is normally the same, we treat offensive and hate speech as synonyms.

Based on these facts, the problem of hate speech recognition in social networks in order to protect other users from such messages is very important one.

Currently, more and more attention is paid to solving this problem and experts from different fields (including computer scientists, psychologists and linguists) are making efforts to create approaches and technologies that are able to recognize offensive messages in social networks with a maximum accuracy and in the shortest period.

In this article, we present our approach to hate speech recognition based on the classical models of machine learning and the results we achieved by the experiments with two Twitter datasets. One of the experiment had the aim to identify the offensive language in messages, and another one was aimed to recognize the target of the offensive messages: an individual person or a group of people. This paper is organized as follows.

Some relevant related works in the area of hate speech recognition are described in Section 2.

Section 3 describes the datasets we used for our experiments. Section 4 presents the preprocessing stage and the methodology for the models we used. In Section 5, the results are described and analyzed. In Section 6, we summarize our work and plan some steps for the future research.

2 Related Work

In order to classify a message as hate speech, it is necessary to understand clearly how to distinguish an offensive message from a non-offensive one. It seems most obvious to call the text offensive if it includes swearing, but this is not always correct, since the author may not have the aim to offend users, and he may have used profanity to express emotion or an another reason.

In [11], the authors created a dataset to work with aggressive language identification in the context of two types of abuse: sexism and racism.

They collected and annotated 16,914 tweets over the course of two months: 3,383 of which (written by 613 unique users) were labeled as sexists, - 1,972 tweets by 9 users were racist and 11,559 tweets by 614 users were neither sexist nor racist.

The authors indicated some rules that help draw a clear line between offensive and no offensive tweets. The most important rules which help to indicate an offensive message are (i) a sexist or racial slur were used in a message, (ii) a minority were attacked or there were negatively stereotypes a minority, or there was the aim to seek to silence a minority in a message, and (iii) hate speech or violent crime were promoted but did not directly use or xenophobia or sexism were defended in a message.

Additionally, the authors extracted some features using the meta-data of messages. They highlighted a gender of users while looking at names in profiles and found that about half of all messages were written by men, 2.26 percent - by women and 47.64 percent - undefined users.

Using the time zones, which were marked in tweets authors, created geographic distribution feature and they calculated the length of tweets.

The authors created the model based on logistic regression, which showed best result in case when features of gender and geographic location were counted (F1-score 0.7362). It is interesting that in case when gender, geographic location and length of tweet were taking into account the result were not so good (F1-score 0.7347).

It was shown in [14] that detecting hateful content using linguistic characteristics is quite difficult because of the absence of unique discriminative features. Seven public Twitter datasets were analyzed, five of which included three different types of tweets labels: sexism, racism and neither, and another two datasets were contacted of tweets divided by hate and non-hate classes.

A uniqueness score measurement was created, which indicated the number of unique words corresponding to each class (i.e. not occurring in other classes) were included in the message. This measure is found as the intersection of words in a message with unique words from this class divided by the number of all words in that message.

The meaning of this measure takes a value from 0 to 1. The authors checked each dataset using this measure. They found that about half of all tweets did not contain the unique words of their classed or contained very few this words (the meaning of the measure less than 0.5) and it means that there are no discriminative features which could indicate hate speech because of the fact that people can write an offensive messages using different words.

In [2], the authors investigated the relationships between the user who posted an offensive message and the user who was the target of this message. The authors of the article analyzed two different groups of slurs: Sexist Derogatory Slurs (e.g., bitch) and Sexist Objectifying Slurs (e.g., hot chick) in case of different relationships (e.g., friends, partners, work-related context) and the gender of the user (man or woman) in the Italian language. Sexist Derogatory Slurs was the class with the aim of denigrating women in the context of stereotypes, sexual looseness and promiscuity, while Sexist Objectifying Slurs was the class of words, which reduced women to being objects of male sexual interests.

The authors showed that people tend to evaluate Sexist Derogatory Slurs as being more offensive when compared with Sexist Objectifying Slurs. Slurs directed at women were judged to be

more offensive those directed at men. Correlation analyses was performed that the high level of frequency of use is connected with the high level of social acceptability. The results of evaluation in the specific social settings shows that the slurs have more social acceptability in a context of affected relationships than in an equal work-related context (a conversation between colleagues) and a higher work related context (a conversation between a superior and a subordinate). It is interesting that the Sexist Objectifying Slurs had the lowest social acceptability in the work-relation situation with unequal positions (supervisor- subordinate).

It should be noted that although it was not possible to indicate special linguistic characteristics for aggressive speech messages, the authors of [1] conducted a study, which revealed some grammatical and lexical features, which are typical for aggressive posts. The article demonstrated the importance of functional linguistic variation in a corpus of racist and sexist Tweets. The authors analyzed 628 sexist tweets and 858 racist tweets and tried to establish the role of lexical and grammatical features using MDA (multidimensional analysis) in three different dimensions (interactive, antagonistic and attitudinal).

The first interactive dimension showed how interactive or informative the message was, the antagonistic dimension presented the attitude of the user to the reader: whether he agrees with them or not, and the attitudinal dimension represented the degree of attitudinal judgment exhibited by a tweet. Finally, the authors compared the difference for racist tweets and sexist ones along each of three dimensions.

The results showed that the sexist messages were more interactive and more attitudinal then racist ones, but had the same measure in the antagonistic dimension, and the most popular linguistic feature in offensive language were question marks (there were a lot of questions in this messages) and question DO (when a sentence begins with the word do).

When we speak about available approaches for models created for aggressive speech recognition it is important to note classical machine learning approaches. Models based on Support Vector Machine, Naive Bayes and Logistic Regression approaches and some ensembles of this classical models achieved quite good results [5, 9] in shared tasks AMI@IBEREVAL-2018 [3] and EVALITA-2018 [4]. The aim in these cases was to detect misogynistic behavior in English and Italian tweets, where the task was to multiclassify the messages according to the type of offensive language.

The classes for the classification included Stereotype and Objectification (a widely held but fixed and oversimplified image or idea of a woman), Dominance (an assertion the superiority of men over women), Derailing (a justify woman abuse), Sexual Harassment and Threats of Violence (a sexual advance, an intention to physically assert power over women through treats of violence) and Discredit (a slurring of women with no other larger intention) groups of misogynistic messages.

It should also be noted that the models, which are based on the approach of deep neural networks, have great prospects in the problems of identifying hate speech in messages. In [8], three different models were presented based on neural networks, which shows quite good results in hate speech identification.

These were a CNN-based model, which is the character-level convolutional network, a convolutional network where a sentence was segmented into words on input and finally a model, which combined the previous two with two inputs: characters and words. The idea behind creating this model was an observation that offensive tweets often contain either purposely or mistakenly misspelled words. All three models had 3 layers.

The classification was for three different classes: racism, sexism and none and in this case the best results was shown by the hybrid model of convolutional networks with 0.827 F1-score, while the best result by classical model (Logistic Regression) was 0.814.

In addition, authors created the combination of neural network based model and classic machine learning approach (Logistic Regression) and shown very good results for the multi-classification and this result looks encouraging.

Table 1. Examples of tweets with the HatEval dataset

Type of Tweet	Tweet
Hate Speech	He real truth is after Cologne and in the Nordic countries and Others no one trusts any refugees a better life for them doesn't mean 1
Non-hate Speech	NY Times: 'Nearly All White' States Pose 'an Array of Problems' for Immigrants
Individual target	You seem like a hoe Ok bitch? Did I ever deny that? Nope Next.
Group target	The German Government Pays for 3 Week Vacation for Refugees to Go Home MuslimImmigration No the German government isn't paying, the German taxpayers are paying!The German government is robbing native Germans to finance the Islamization of Germany.

3 Datasets

For our experiments, we used two datasets consisting of messages from Twitter. The first dataset was available in the framework of multilingual detection of hate speech against immigrants and women in Twitter (HatEval-2019 - one of the tasks in the frames of SemEval-2019 challenge[1]). The task consisted of two subtasks, one of which was the binary classification between offensive and non-offensive messages in case of hate speech detection against immigrants and women, and the second task proposed to make an aggressive/non-aggressive and individual target/a group target classification on the offensive messages.

Although were there two datasets (one in English, the other in Spanish) we used only the English one for our experiments. The training dataset included 10,000 tweets, of which 4,210 were labeled as hate speech tweets and 5790 not, and 1,560 tweets had individual target and 2,650 tweets had group target. The testing dataset contained of 3,000 tweets. Some examples of different types of messages are presented in Table 1.

The second dataset we chose was from the Identifying and Categorizing Offensive Language in Social Media shared task (OffensEval-2019 - one of the tasks in the frames of SemEval-2019 challenge).

The challenge had 3 different subtasks: the first two were binary classification for offensive language identification (is message offensive or not) and automatic categorization of offense types (is offensive message insult a person or a group of people or it is non-targeted profanity and swearing), and the last subtask had the aim to make the target classification for three groups: individual, group or other (in this case the target of the offensive post did not belong to any of the previous categories, e.g., a situation, an event, or an issue) target.

The training dataset we used consisted of 13,200 tweets, and 4,400 of them were offensive in the ratio: 2,407 tweets - individual target, 1,074 tweets - group target. Some examples from different types of tweets are presented in Table 2.

There were 860 tweets for testing in case of hate speech identification and 213 from them tweets for the target classification. It should also be noted that in the OffensEval dataset all references were anonymized and replaced with the string URL.

Although the datasets are not well balanced, this distribution can be perceived as the present state of affairs and the frequency of occurrence of such messages in reality.

For the HatEval dataset and OffensEval dataset we made experiments with the aim of hate speech identification and the target of hate speech recognition on the training data (10,000 tweets and

[1]http://competitions.codalab.org/competitions/19935

Table 2. Examples of tweets with the OffensEval dataset

Type of Tweet	Tweet
Offensive tweet	DrFord DearProfessorFord Is a FRAUD Female @USER group paid for and organized by GeorgeSoros URL
Non-offensive tweet	@USER @USER Obama wanted liberals amp; illegals to move into red states
Individual target	@USER @USER @USER @USER LOL emoji Throwing the BULLSHIT Flag on such nonsense!! PutUpOrShutUp
Group target	4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence! emoji ! emoji

Table 3. Examples of tweets with additional data from the referenced link

Tweet	Tweet from the link
Thinking she a pretty decent bitch but she a hoe prolly	Thought she was a pretty ricky bitch but she like you gotti
First of all sebody find a boyfriend for @USER. She is so f* lonely .. when you don't	Believe in your stand but have other reasons influencing your thoughts.. you come up with these statements. Unbelievably unsmart.
Shes right..he is pretty awesome! @USER ..dont you agree?	GUYS! @USER is the coolest reporter around and the coolest guy I know

13,200 tweets accordingly) and after we tested our created model on the testing datasets (3,000 tweets and 860 tweets accordingly).

4 The Model

In this section, we explain two main steps of our experiments: preprocessing and modeling. Each stage of model creation was important for us because both preprocessing and modeling make a great contribution to the quality of the constructed classifiers and to the results of the research.

4.1 Preprocessing

The preprocessing stage is very important, because at this stage, we can work with data from the dataset directly and we can try to identify certain patterns that occur in messages. In the analysis of the data and their subsequent study, we have taken the following steps that allowed us to represent messages in a more convenient format for subsequent processing:

- we replaced all references to Twitter users (i.e., terms commencing with the @ symbol) with the term USER;
- we labeled some combinations of symbols with were used often in messages such as !!!,??? and replaced them with the term emoji;
- we added to the training dataset the texts of the messages referred to by the users in the original messages.

It is necessary to explain this last point in more detail. Table 3 provides examples of such messages. The left column shows the original tweets from the dataset (user names have been changed and links have been removed for privacy reasons), and the right hand column contains the text of the tweet that was referenced in the original messages. The first two examples are offensive messages, while the third one is a non-offensive message.

These examples reinforce the contention that if a message is offensive, there is a large probability that the original referenced message was itself

abusive, and when the message is not offensive, the linked message was non-offensive as well. In our work with the HatEval dataset, we used not only the texts of the original messages, but also the texts that were extracted using links.

We did this on the basis that, where such referenced data was available, that the data for training is expanded, which would in turn improve the classification results.

The OffensEval dataset did not include links, so we had not an opportunity to expand the training dataset using this feature.

It should be noted that this feature reflects the dynamic nature of social networks and it can make different contributions to the modeling results at different times.

For example, if the dataset is fresh and all links are active, we can actually expand the original dataset with many referenced posts.

However, over time, the linked tweets are blocked or deleted for various reasons, and consequently the texts of the message are no longer available. This means that if today we were able to extract additional data using links, there is no guarantee that we will be able to use the same additional information tomorrow.

We have made the replacement for all links which did not help with an extracting any additional information (then it was a link to the blocked or the external content) with the term URL. We used TF-IDF[2] (where TF is term frequency and IDF inverse document frequency). It is a statistical measure that is used for the evaluation of the importance of a word in a context. The weight of a word is proportional to the frequency of this word use in the message and inversely proportional to the frequency of this word use throughout the context, so this measure helps in a process of texts analysis.

4.2 Modeling

At the modeling stage, we constructed an ensemble of models based on the classical machine learning approach.

As we noted above, such models allow us to achieve sufficiently high results in solving the problem of recognition of hate speech. Our ensemble was based on four different models:

– Logistic Regression (LR) [6, 12] this type of classifier applies an exponential function to a lineal combination of objects, which we could extracted from the data. This type of classifiers is very popular because of the speed of data handling and ease of use.

– Support Vector Machine [7]. This classifier is based on the principle of constructing optimal hyperplanes, which could separate the data that are supposed to be linearly separated. Such hyperplanes will be as far away from all sample elements as possible and thus most clearly divide the space into classes.

– Naive Bayes (NB) [13]. In this case, the maximum likelihood function is calculated for each class, this function is applied to the classified object, and after applying the function of the conditional probabilities are calculated. The object belongs to the class with the highest calculated conditional probability.

– The combination of Logistic Regression and Naive Bayes models (LR+NB). In [10] it was shown that the combination of generative and discriminative classifiers demonstrates a strong and robust result in the task of texts classification. In the article, it was presented a model variant where an SVM is built over NB log-count ratios as feature values, because in short sentiment tasks NB has better results in comparison with SVM model, which achieve better results in the work with longer reviews. We used the interpolation between LR and NB with the coefficient of interpolation as a form of the regularization: in practice, it means that in this type of modeling we trust NB unless the LR is very confident.

We then created the ensemble of models that includes all of the above models:

Logistic Regression, Support Vector Machine, Naive Bayes and the interpolation model between Naive Bayes and Logistic Regression.

This construction was built using the idea that the more models will classify the message as a

[2]https://scikit−learn.org/stable/modules/generated/sklearn.feature extraction.text.TfidfVectorizer.html

Table 4. Results for each model with training HatEval dataset

Model	Macro F1-score for Hate Speech identification	Macro F1-score for Target identification
Logistic Regression (LR)	0.52	0.65
Naïve Bayes (NB)	0.60	0.69
LR+NB	0.65	0.70
Support Vector Machine (SVM)	0.61	0.69
Ensemble of models	0.67	0.72

Table 5. Results with HatEval dataset

Type of classification	Macro F1-score with the training dataset	Macro F1-score with the testing dataset
Hate Speech	0.67	0.58
Target	0.72	0.64
Aggressiveness	0.68	0.60

particular group, the higher a probability that the message really belongs to the selected class.

All models had an equal contribution to the classification. In order to find the tweet class, we summarized the probabilities, which we found using each model and divided this value by the number of models participating in the classification. Then we compared the obtained averages and choose the class if the average value for it was the maximum.

5 Results

To evaluate the results obtained by the modeling, we used the macro F1-score[3], which is well suited in the case of texts classification. This metric is a combination of precision and recall into an aggregated quality criterion. F1-score is a harmonic mean of precision and recall.

F1 score is calculated as the resulting precision and recall of the classifier for each class, and then it is considered the average. This measure reaches a maximum when precision and recall are equal to one, and is close to zero if one of the arguments is close to zero.

The results obtained from the experiments with HatEval dataset are presented in Table 4. The experiments include hate speech recognition and the target of hate speech identification.

Results for the task of hate speech recognition shows that the ensemble of models we created achieves the best results in comparison with Logistic Regression, Naive Bayes, Support Vector Machines models and the interpolation between Logistic regression and Naive Bayes.

Also, note that in case we did not use the information from the references in messages, the predicted F1-score for the modeling with HateEval dataset was 0.59, while with the addition of the data increased to 0.67.

Experiments for the target classification also shown that the ensemble of models achieves the best results with 0.72 macro F1-score on the training dataset. This results are quite better than F1-scores for hate speech identification, so we could say that the approach we propose is more useful in case of binary classification between offensive language especially, than the binary classification between offensive and non-offensive messages.

[3]https://scikit−learn.org/stable/modules/generated/sklearn.metrics.f1score.htmlsklearnmetrics−f1−score

Table 6. Results for each model with training OffensEval dataset

Model	Macro F1-score for Hate Speech identification	Macro F1-score for Target identification
Logistic Regression (LR)	0.63	0.57
Naïve Bayes (NB)	0.62	0.59
LR+NB	0.68	0.72
Support Vector Machine (SVM)	0.57	0.69
Ensemble of models	0.70	0.73

Table 7. Results with OffensEval dataset

Type of classification	Macro F1-score with the training dataset	Macro F1-score with the testing dataset
Hate Speech	0.70	0.68
Target	0.73	-

Table 5 presents our results for hate speech and the target of hate speech identification using training dataset in comparison with the results we achieved using the testing dataset.

As we can see, the results obtained with the training dataset are below the testing results by 1-9 percentage points for aggressive language recognition and the difference between the results on testing and training datasets is only 1 percentage point. In the first case (hate speech recognition), the best-published result was equal to 0.60 F1-score, while our result is slightly lower (0.58 macro F1-score).

In addition, we made the experiment of aggressiveness of speech recognition to compare our results with the published results. In the published results macro F1-score was defined as the average value for all classification types: macro F1 measures for hate speech, target and aggressiveness were summed and the resulting value was divided by 3.

For such an estimate, the best-published result is 0.60 macro F1-score, while our result reaches 0.61 macro F1-score.

The difference of 1 percentage point may indicate that the model we created is universal and it can show equally good results for different types of classification, not for any particular one class.

Table 6 shows the results we have achieved with the OffensEval training dataset for hate speech and the target of the hate speech recognition.

From the presented data we can see that the best results (0.70 macro F1-score for hate speech identification and 0.73 macro F1-score for target classification) are achieved using the ensemble of models that combines simpler models, as we expected in the modeling. Also, note that for the target classification the interpolation of Logistic Regression and Naive Bayes models shows high results using the interpolation between LR and NB with 0.25 coefficient of interpolation. Table 7 shows the classification results for the training and the testing OffensEval datasets. The results achieved using the ensemble of models in the task of hate speech identification are quite similar for the training and for the testing datasets and have a difference in two percent points only. In case of the target classification, it is not possible for us to compare the results on the training and testing datasets, because the golden data, which indicate a type of each message (individual or group), is not available for this moment.

It is interesting to compare this data in future with our training results to make a conclusion about the difference between messages real distribution and our predictions.

To sum up, the ensemble of models we propose achieves the best results in all types of classification both using HatEval and OffensEval datasets. The macro F1-score for all experiments was quite high, but in case of the target classification, the results are higher than the results of hate speech identification.

The results of hate speech identification are a little higher on OffensEval dataset, because in this case the data for classification was bigger than in HatEval dataset (13,200 tweets in the OffensEval dataset and 10,000 messages in the HatEval dataset). The results of the target classification were better for the OffensEval dataset with the difference of 1 percent point for the training datasets.

Now we would like to analyze some reasons because of which, in our opinion, the results of classification for the HatEval dataset were difference from the results with the OffensEval dataset.

First, there were more labeled messages in case of hate speech recognition for models training in the OffensEval dataset than in the training HatEval dataset, and it could affect the result. Despite the fact that the developed ensemble of models is able to make a classification using a small number of training data, an increase in the number of messages for our training model always leads to an increase in the accuracy of the classification.

Secondly, in case of the HatEval dataset we had an opportunity to insert additional messages in the training dataset using an external content obtained through links in the preprocessing stage, while in the OffensEval training dataset all links were closed (replaced by special characters).

Additional messages are not only the incremental amount of text, which improves the quality of classifiers work, but also an opportunity to catch messages related on the meaning, and to identify some themes and patterns, which can be potentially more frequent in the context of aggressive language detection. As shown above, the use of data from links in messages improves the classification quality of our model.

5 Conclusion

This article describes a possible approach to offensive language recognition. It involves a preprocessing step and a creation of models, which allows us to obtain quite good results in a solving the problem of a small number of messages classification with the aim to identify hate speech. The results were good for different types of classification: both in the case of hate speech identification, and in the target of hate speech identification. The results achieved by the proposed model are competitive in comparison with the best-published results, which were achieved in the processing of the same datasets.

The features which were used at the preprocessing stage indicate that the date preprocessing is very important and it is necessary to pay attention not only to the process of the model creating, but also to the analysis of the messages in the dataset and it is vital to try to identify certain patterns that occur in them.

Also, the use of messages that were referenced in the data improved the results of our model, so we can say that it makes sense to develop this area of research in the future and we should try to take into account not only the text of the message, which was referenced, but also other information, including:

– if a message contains a link to a blocked or deleted message, it makes sense to mark it as a separate marker, since a blocking or a deleting may indicate that the message was offensive and it was a part of an aggressive language;

– if a message contains a link to an external content, it makes sense to add some part of this content (for example, the number of characters which not exceeding the number of characters in the tweet) as data for analysis, and thus expand the existing dataset. It is also possible to enter an additional token that will indicate that the message has a link to external sources.

In addition, in the future we plan to use additional sources of information, such as dictionaries of words of different tonality and lists of

a swearing/harassment vocabulary for working with the identification of aggressive language.

Despite the fact that the message may contain, for example, swearing words, but be not offensive, it is intuitively clear that the combination of insult in the message - the message is offensive in nature is more common, so this direction in the continuation of the research seems promising to us.

In addition, as described above, the use of models based on deep machine learning, in particular CNN, and the combination of such models with models based on the classical machine learning approach, increases the accuracy of the research. In the future, we plan a developing of our ensemble model in this direction in order to improve the accuracy of the classification.

Acknowledgements

The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMISFAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

1. **Clarke, I. & Grieve, J. (2017).** Dimensions of abusive language on Twitter. *Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics,* pp. 1– 10.
2. **Fasoli, F., Carnaghi, A., & Paladino, M. (2015).** Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, Vol. 52, pp. 98–107. DOI: 10.1016/j.langsci.2015.03.003.
3. **Fersini, E., Anzovino, M., & Rosso, P. (2018).** Overview of the task on automatic misogyny identification at Ibereval. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages,* pp. 214–228.
4. **Fersini, E., Nozza, D., & Rosso, P. (2018).** Overview of the Evalita 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian,* pp. 59–66, Academia, 4497. DOI: 104000/books.
5. **Frenda, S., Ghanem, B., & Montes-y-Gomez, M. (2018).** Exploration of misogyny in Spanish and English tweets. *CEUR Workshop Proceedings,* CEUR-WS.org, Vol. 2150, pp. 260–267.
6. **Genkin, A., Lewis, D., & Madigan, D. (2018).** Large-scale Bayesian logistic regression for text categorization. *Proceedings of the NAACL Student Research Workshop,* Vol. 49, No. 3, pp. 291–304. DOI: 10.1198/004017007000000245.
7. **Joachims, T. (2002).** *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.
8. **Park, J. & Fung, P. (2017).** *One-step and two-step classification for abusive language detection on Twitter.* arXiv preprint aeXiv:1706.01206**.**
9. **Shushkevich, E. & Cardiff, J. (2018).** Classifying misogynistic tweets using a blended model: the AMI shared task in Ibereval 2018. *CEUR Workshop Proceedings, CEUR-WS.org,* pp. 255–259.
10. **Wang, S. & Manning, C. (2012).** Baselines and bigrams: simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, Vol. 2, pp. 90–94.
11. **Waseem, Z. & Hovy, D. (2016).** Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, pp. 88–93.
12. **Wright, R. (1995).** *Logistic regression.* L.C. Grimm.
13. **Zhang, H. & Li, D. (2007).** Naive Bayes text classifier granular computing. *GRC´07 IEEE International Conference*. DOI: 10.1109/GrC.2007.40.
14. **Zhang, Z. & Luo, L. (2018).** Hate speech detection: A solved problem? The challenging case of long tail on twitter. *Semantic Web*, Vol. 10, No. 5, pp. 925–945. DOI: 10.3233/SW-180338.

Article received on 29/10/2019; accepted on 05/03/2020.
Corresponding author is Elena Shushkevic