2020-08-24

Misogyny Detection in Social Media on the Twitter Platform

Elena Shushkevich
Technological University Dublin

Recommended Citation

Shushkevich, E. (2020). *Misogyny Detection in Social Media on the Twitter Platform*. Dissertation. Technological University Dublin. doi:10.21427/d1jc-vj32

Misogyny Detection in Social Media

on the Twitter platform

Elena Shushkevich

Technological University Dublin

Master **2020**

Misogyny Detection in Social Media

on the Twitter Platform

A Thesis Presented For the Award of Master of Computer Science by

Elena Shushkevich

Technological University Dublin – Tallaght Campus

Department of Computing

For Research Carried Out Under the Guidance of

Dr. John Cardiff and Dr. Paolo Rosso

Submitted to Technological University Dublin

February 2020

DECLARATION

I certify that this thesis which I now submit for examination for the award of _____, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for graduate study by research of the Technological University Dublin (TU Dublin) and has not been submitted in whole or in part for another award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of the TU Dublin's guidelines for ethics in research.

(The following sentence may be deleted if access to the thesis is restricted according to Section 4.8 of the TU Dublin Research regulations)

TU Dublin has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature _____ Date _____

Candidate

Acknowledgements

I would like to thank John Cardiff, Paolo Rosso and Michael Alexandrov. Also, I would like to thank my parents, without whom I would not have been able to complete this research.

Abstract

The thesis is devoted to the problem of misogyny detection in social media. In the work we analyse the difference between all offensive language and misogyny language in social media, and review the best existing approaches to detect offensive and misogynistic language, which are based on classical machine learning and neural networks. We also review recent shared tasks aimed to detect misogyny in social media, several of which we have participated in. We propose an approach to the detection and classification of misogyny in texts, based on the construction of an ensemble of models of classical machine learning: Logistic Regression, Naive Bayes, Support Vectors Machines. Also, at the preprocessing stage we used some linguistic features, and novel approaches which allow us to improve the quality of classification. We tested the model on the real datasets both English and multilingual corpora. The results we achieved with our model are highly competitive in this area and demonstrate the capability for future improvement.

List of Publications Based on this Thesis

1. Shushkevich, E., Cardiff, J., Rosso, P., Offensive language recognition in social media, 7th Int. Symposium on Language and Knowledge Engineering, to be published in Computacion y Sistemas Journal of Computer Science and Engineering, 2020

2. Shushkevich, E., Cardiff, J., Automatic Misogyny Detection in Social Media: A Survey, Computacion y Sistemas Journal of Computer Science and Engineering, accepted for publication, 2019

3. Shushkevich E., Cardiff J., Rosso P., TUVD team at SemEval-2019 Task 6: Offense target identification.In Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval), pp 770-774, Association of Computational Linguistics, 2019

4. Shushkevich, E., Cardiff J.: Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team, EVALITA 2018, Evaluation of NLP and speech tools for Italian, 2018

5. Shushkevich, E., Cardiff J.: Classifying Misogynistic Tweets Using a Blended Model: The AMI Shared Task in IBEREVAL 2018, Evaluation of Human Language Technologies for Iberian Languages Workshop, CEUR Proceedings vol 2150, 2018

Table of Contents

Chapter 1

Introduction

Nowadays, the Internet has become a routine for almost all people. Only ten years ago it was the norm to receive news from newspapers and/or television programmes. Today, Internet platforms are the main source of information for many people. Of course, many of them provide the same content as the printed media and the articles are written by the same journalists. However, new online platforms are now appearing, which are organized according to the same principle of magazines (with sections devoted to various topics, such as news sections, as well as sections on health, fashion or travel), but authors of these articles can be people who do not have professional education, but who may have deep knowledge in certain areas and have a their own position.

Not only the news, but also the statement of the author's position in the article can lead to discussions on the Internet (which were absent at the time of the print media existence only), during which the participants in the discussion can allow themselves to use offensive language and insulting remarks about other people. This problem - abusive language and hate speech on the Internet - appears more and more often, as more and more people use social media platforms in order to express their opinion.

The same problem is increasingly manifested with the increase in the number of users of social networks: now it is not necessary to go into the "general" discussion of any news or person, the user can use his personal Twitter or Facebook to comment on a situation, and often such comments can be offensive.

Offensive language is commonly defined as any communication that disparages a person or a group of people on the basis of a wide range indicators such as gender

(misogyny), race, ethnicity, sexual orientation, nationality, religion, color amongst other features. Misogyny in social media as part of the hate speech in general has become a big problem too. Such messages offend users and can cause them moral harm, so we need to pay more attention to this problem and try to identify this type of message and prevent them from spreading further. With the increasing importance of social media networks, the number of misogynistic messages also increases, so the problem of identifying such messages has become particularly relevant.

The motivation of misogyny detection in social media is to protect users from the harm of such type messages.

The problem of automatically detecting such messages has only recently begun to receive attention from the research community. A current problem is that researchers do not know at the moment which of the approaches to finding such vocabulary in social media is the best. Traditionally, there are two classical approaches to text classification: one of which is based on classical machine learning models and the other approach is based on neural networks. Today, while the area of recognition abusive language in social media has just begun to develop, it is important to understand which approaches work better, and under which circumstances.

Another issue is an absence of information - balanced datasets and challenges devoted to offensive and misogyny language recognition in social media - which are very important in cases when researchers want to test their models for offensive language recognition on real data. As more datasets become available over time, more research can be undertaken in the field of offensive and misogyny language detection and classification.

In our work we are going to make the detailed research of offensive language in social media and of existed approaches which could help to recognize it, and create the model which allow to identify such messages and classify them with high accuracy.

1.1 Objectives and Task

Here, we highlight the objectives and tasks under consideration in our work. In this thesis, we try to address the following questions:

- What is the meaning of offensive language in social media and why is it important nowadays?
- What are the differences between offensive language and misogynistic language, and what are the main characteristics of each?
- How the issue of detecting and disseminating such vocabulary currently being addressed? What are the main approaches to offensive and misogynistic language recognition?
- What are the best approaches for the misogyny detection in social media? How can we create a model which could compare with state-of-the-art models of offensive and misogynistic language detection in social media? Which approaches and features should we use to improve the quality of existing models?
- How can we demonstrate the quality and competitiveness of our approach?

To find the answers to these research questions we will review existing research in the field of offensive language and misogyny detection in social media, and also mathematical modeling tools and linguistic features that are applicable for creating systems that can identify and classify misogyny in social media.

We will propose a model developed for the task of misogyny detection in social media based on existing approaches and we will add new additional features which will help the model to be more effective in case of misogyny message recognition.

The model we will present is based on ensemble of classical machine learning models: Logistic Regression, Support Vector Machine, Naive Bayes and the combination of Logistic Regression and Naive Bayes Models. Each model will show the probability of a message belonging to a particular class (in other words, for any message, the model shows the probability of belonging to different classes) . The ensemble of models sums up the probabilities obtained and makes a final decision about which class the message belongs to, based on which class the sum of probabilities for all models was higher. The main new additional feature we created - adding additional texts from the links in messages - allows us to achieve better results.

We will show that our model achieves competitive results in case of misogyny detection in social media in comparison with results of other similar models using existing datasets from shared tasks devoted to misogyny identification problem and that our model achieves competitive results on multilingual corpora (English, Italian and Spanish datasets).

1.2 Document Outline

The thesis consists of 6 Chapters: Chapter 1, the current chapter, describes the motivation for our work - why this topic is urgent and what are the specific objectives and tasks in our work.

In Chapter 2 we discuss the difference in offensive language and misogyny language and describe the main approaches to offensive and misogyny language recognition in

social media and measures which help to analyse the quality of the classification. Also, in this Chapter we analyse the most significant works devoted to offensive and misogyny language recognition and classification. We also discuss recent shared tasks of offensive language classification, and present the best approaches and models which were applied to these tasks.

Chapter 3 presents the datasets which we used for our research into misogyny language recognition in social media, and which we used to train and test our models.

In Chapter 4 we present the models we exploited for our purposes, including Logistic Regression, Support Vector Machines, Naive Bayes, and combinations of these models. Also, in this Chapter we present the preprocessing techniques we applied, and new features we created especially for the thesis.

Chapter 5 present all of the results we achieved with our models, including experiments with additional data from links and experiments with multilingual corpora, and comparison the results.

Chapter 6 is conclusion in which we analyse and discuss the achieved results and highlight some important steps for future work.

Chapter 2

Offensive and misogyny language recognition

As mentioned in Chapter 1, offensive language recognition become a serious challenge, and to deal with it we should understand clearly what offensive language and misogyny in social media actually means. We explain the meaning of these words and the difference between offensive language and misogyny in social media.

We also describe some interesting approaches for offensive language detection and misogynistic messages detection and some challenges which aim was to create automatic models for misogyny detection.

This chapter is organised as follows: in Subsection 1 we give an explanation of offensive language in social media in general and in Subsection 2 we describe the existing methods and tools of offensive language in social media recognition and classification. In Subsection 3 we present and analyse relevant results achieved by other researchers. Subsection 4 is devoted to the study of automatic misogyny identification and the best approaches used in this area and Subsection 5 contains information about recent misogyny identification shared tasks named IberEval, Evalita, SemEval and an analysis of the approaches proposed by the teams achieved the best results at this challenges. Finally, Subsection 6 presents the summary of existing approaches of misogyny identification in social networks.

2.1 Offensive language in social media

So what does offensive language mean? Intuitively, these are messages containing swear words against other users. However, there are situations when the message does not contain profanity, but its essence is offensive to a specific person or a group of other people. Thus, hate speech is a generalized designation of linguistic features of

expressing the negative attitude of "opponents" based on religion, nationality, cultural or more specific subcultural values.

Hate speech can be expressed not only through a direct or veiled violence and discrimination, but also to justify cases of violence and discrimination (for example, historical cases), as well as reasoning about the criminality, superiority, disproportionate material wealth of any ethnic, religious group or social sign. Hate speech also includes statements that create a negative image of a group, mentioning it in a negative or offensive context, or quoting an offensive statement without any comment that would indicate a difference in the position of the author of the offensive quote and the person who reposted it.

Because of the growth of social media platforms such as Twitter, the volume of offensive messages has become really huge. While in the past it was practical to detect this type of message and just delete it from the platforms and forums, nowadays it is impossible to do so manually, and consequently we have the challenge of automatic identification offensive messages arises.

2.2 Methods and Tools for Automatic Detection of Offensive Speech

It is important to describe some methods and tools that have successfully been applied to solve the problem of text classification, a special case of which is the task of offensive language identification in social media. From the mathematical point of view, we can try to establish which models function better than others.

There are some different methods for the problem of text classification exist and we will describe them in more detail. Nowaday besides unsupervised models for the

offensive language classification [1;2] there are two main principal approaches which are commonly used: the neural network approach [3;4] and the classical machine learning approach [5;6]. Also, there are exist some combinations of methods - ensembles - which mean the creation of different types of models (for example, a combination of different classical machine learning models or a combination of both classical machine learning models and models based on neural networks approach) with a subsequent choice of the best models [7]. In this section we will describe these approaches in more detail.

2.2.1 Classical machine learning approach

Models based on the classical machine learning approach are popular for solving the problem of text classification and they show good results in it. In classical machine learning, researchers use a relatively small amount of data and determine which most important functions are in the data that the algorithm needs to predict. By models based on the classical machine learning, we mean a set of methods used to create models that can learn from observations and make predictions. Such models use algorithms, regression, and related sciences to understand data. These algorithms can usually be considered as statistical models.

In the following, we describe some of the most popular models which have been applied to text classification.

Logistic regression

One of the most popular and efficient types of machine learning models is logistic regression models. This method is well applicable for binary classification problems (that is, problems where we get one of two classes at the output) [8;9]. Logistic regression is used to predict the likelihood of a certain event from the values of many features. To do this, a dependent variable y is introduced, which takes one of two possible values - 1 (if the event happened) and 0 (if the event did not happen). Also, many dependent variables (predictors) are introduced, based on the values of which a conclusion is made about the value of the dependent variable. The logistic function (sigmoid) takes the form:

$$F(z) = 1/(1 + e^{-z}) \quad , where \ z = a_0 + a_1 x_1 + ... + a_n x_n \ ,$$

x and a - column vectors of values of independent variables and parameters (regression coefficients).

There are a lot of published works which confirm the effectiveness of logistic regression in case of texts classification. The authors of [10] achieved good results for solving the problem of toxic context recognition in social media. The work [11] shows that logistic regression could achieve better results in comparison with neural network in case of handwritten character recognition. Also, logistic regression shows good results in case of the recognition of similar shaped characters [12].

Naive Bayes

Another popular approach to creating a classification model is based on the use of the Naive Bayes classifier [13].

This classifier is based on the Bayes theorem with the assumption that the parameters of each attribute are independent. In this case, the classification model is defined as follows:

$$v_{NB}(a_1, a_2, ..., a_n) = argmax_{v \in V} p(x = v) \prod_{i=1}^{n} p(a_i | x = v),$$

where v is a class of a message (for example, “insult” / “not insult”), a- attributes (words). Due to the assumption of independence (naivety), the parameters of each attribute can be trained separately, and this greatly simplifies training, especially in cases where the number of attributes is large.

It should be noted that the Naive Bayes classifier makes a very bold and not entirely correct assumption: in the classification of texts we assume that different words in the text on the same topic appear independently of each other, although this is not entirely true. However, despite the fact that the attributes are, sure, dependent, their dependence is the same for different classes and is reduced mutually when assessing probabilities.

There are a lot of successful examples of Naïve Bayes classifier implementation for the aim of text classification. In the work [14] authors used Naïve Bayes for the classification of abusive comments from YouTube, and in [15] authors designed a classifier using Naïve Bayes as a machine learning approach to determine the opinion expressed both in English and Bangla. In [16] it was shown that Naive Bayes classification can be used to identify non-native utterances of English, and in [17] authors built a classifier based on Naive Bayes, that was able to determine positive, negative and neutral sentiments for a message from Twitter. In [18] five

different versions of Naive Bayes were considered, and compared on six new, non-encoded datasets, that contain ham messages of particular Enron users and fresh spam messages.

Support Vector Machines

A popular method is the Support Vector Machine (SVM), which solves the classification and regression tasks by constructing a nonlinear plane separating the solutions [19;20;21]. Due to the nature of the feature space where the boundaries of the solution are constructed, SVM has a high degree of flexibility in solving classification problems of various levels of complexity.

The SVM is based on the concept of hyperplanes that define the boundaries of hypersurfaces. A separating hyperplane is a hyperplane that separates a group of objects that have different class affiliations. The main idea of the method is to translate the original vectors into a space of higher dimension and then to search a separating hyperplane with a maximum gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane separating the classes. The separating hyperplane is that hyperplane that maximizes the distance to two parallel hyperplanes. The algorithm works under the assumption that the greater the difference or distance between these parallel hyperplanes, the smaller the average error of the classifier.

A good example of SVM implementation with the target of texts classification is presented in [22], where the authors created a classifier for active learning, which can let the system ask for labels only on the documents which will most help the classifier learn. In [23] an SVM classifier shows best results for topic classification. In case of

offensive language and cyberbullying detection, classifiers based on SVM also show high results, as presented in [24;25].

Ensembles of models

Ensembles of models have proved to be successful in solving the problems of text classification. An ensemble is a certain aggregate, parts of which form a single whole. The idea of using an ensemble of models is that for the classification several simpler models are used, and then the results obtained in the course of such classification are aggregated into a single final result. As an example, when we combine three different models in one ensemble and two of them identify a message as Class 1, while the last model identify the same message as Class 2, the ensemble identifies this message as Class 1.

As a successful implementation of ensembles of models we could mention [26] where the authors used two hybrid ensemble based models (bagging and bayesian boosting based) for a positive/negative opinion classification of digital camera and works [27;28] where ensembles were created for sentiment classification.

2.2.2 Neural Network (NN) approach

Another class of methods for offensive language detection is based on neural networks usage. Neural networks (NN) allow us to find hidden connections and patterns in texts, but these connections cannot be represented in an explicit form. The increased attention of researchers to neural networks is due to several reasons. Firstly, the use of neural networks improves significantly the quality of solving some standard

text classification problems and sequences. Secondly, the use of neural networks reduces the complexity of working directly with the texts. Thirdly, neural networks allow us to solve new problems (for example, to create chat bots). At the same time, neural networks cannot be considered a completely independent mechanism for the linguistic problems solving.

From a formal point of view, a neural network is a directed graph of a given architecture, the vertices or nodes of which are called neurons. At the first level of the graph the input nodes are situated, and on the last one are output nodes, the number of which depends on the task. As an example, for binary classification, one or two neurons can be placed on the output level of the network, and k neurons for a classification into k classes. All other levels in the graph of the neural network are called hidden layers. All neurons that are on the same level are connected by edges with all neurons of the next level and each edge has a weight. Each neuron is assigned an activation function that simulates the work of biological neurons: they "remain silent" when the input signal is weak, and when its value exceeds a certain threshold, the input value is triggered and transmitted further along the network.

The task of training the neural network with examples (that is, with the pairs "object" - "correct answer") is the task to find the weights of the edges that predict the best correct answers. It is clear that the architecture - the topology of the structure of the neural network graph - is the most important parameter. Although there is no formal definition for "deep networks", it is customary to consider as deep all neural networks consisting of a large number of layers or having "non-standard" layers (for example, containing only selected connections or using recursion with other layers).

Convolutional neural networks (CNN)

A known difficulty in text classification is the variable length of the input: sentences in texts can be of arbitrary length, so it is not clear how to apply them to the input of a neural network. One approach is taken from the field of image analysis and consists in the use of convolutional neural networks (CNN) [29;30;31].

A sentence in which each word is already represented by a vector (vector of vectors) is submitted at the input of the convolutional neural network. Typically, pre-trained word2vec [32;33] models are used to represent words with vectors. Word2vec takes a large text corpus as input and maps each word to a vector, giving the coordinates of the words in the output. First, it generates a corpus dictionary, and then calculates the vector representation of words, learning in the input texts. The vector representation is based on contextual proximity: words occurring in the text next to identical words will have close vectors. The resulting vector representations of words can be used for natural language processing and machine learning. The convolutional neural network consists of two layers: a “deep” convolutional layer and an ordinary hidden layer. The convolution layer, in turn, consists of filters and the “downsampling” layer. A filter is a neuron whose input is formed using windows that move through the text and select a certain number of words (for example, a window of length “three” will select the first three words, words from second to fourth, third to fifth, etc.).

At the output of the filter, one vector is formed, which aggregates all the vectors of words that are included in it. Then, on the subsampling layer, one vector is formed corresponding to the entire sentence, which is calculated as the component-wise maximum of all the output filter vectors. Convolutional neural networks are easy to learn and implement. For their training, a standard algorithm for the back propagation

of errors is used, and due to the fact that the weights of the filters are evenly distributed (the weight of the ith word from the window is the same for any filter), the number of parameters of the convolutional neural network is small. From the point of view of computer linguistics, convolutional neural networks are a powerful tool for the aim of classification. As an example of usage CNN in case of texts classification we could mention the study of health-related topics on social media for the early detection of the different adverse medical conditions, in particular in cases related to the treatment of mental diseases. In [34], convolutional neural networks with word2vec embedding were used to classify user comments on Twitter. The aim of the classification was to reveal adverse drug reactions of users.

Recurrent neural networks (RNN)

Another type of neural networks are recurrent neural network (RNN) [35]. It is necessary to have large corpora to study language models, so the larger the training corpus, the more pairs of words the model "knows". Using neural networks to develop language models reduces the amount of data stored. Depending on the number of hidden layers and the number of neurons on them, the trained network can be stored as a number of dense matrices of relatively small dimension. However, such a neural language model does not allow to take into account the long connections between words. This problem is solved by RNN in which the internal state of the hidden layer is not only updated after a new word arrives at the input, but is also transferred to the next step. Thus, the hidden layer of the recurrent network accepts two types of inputs: the state of the hidden layer in the previous step and the new word. If a RNN processes a sentence, then the hidden states allow it to remember and

transmit long connections in sentences. As an example of successful implementation RNN we could mention the work [36] where implemented RNN using characters as input instead of words, which achieved an increase of approximately 8% in average class accuracy. Also, the authors of [37] showed the successful RNN implementation in a case of classification Greek language messages from Facebook.

Long-Short-term memory (LSTM)

The last type of neural networks which should be noted in case of texts classification is Long-Short-term memory (LSTM) [38;39] which is a specific kind of RNN architecture, and it is capable of learning long-term dependencies. This type of neural networks are suitable for solving a number of various problems and are now used widely. The authors of [40] applied LSTM to predict polarities of tweets and gained 1% better accuracy comparing to the standard RNN model. A bidirectional LSTM, consisting of two LSTMs that are run in parallel, achieved good results in text classification [41]. In [42] the authors developed a variant LSTM model that is based on a tree topology, and this model showed superiority for sentiment classification than the standard LSTM.

LSTMs are designed specially to avoid long-term addiction problems. Storing information for long periods of time is their usual behavior, and not something that they are struggling to learn.

The structure of LSTM looks like a chain also, but the modules look different. Instead of one layer of the neural network, they contain four layers, which interact in a special way. The key component of LSTM is the cell state. The state of the cell resembles a conveyor belt. It passes directly through the entire chain, participating in only a few

linear transformations. Information can easily flow through it without being changed. However, the LSTM can remove information from the state of a cell and this process is governed by structures called gates.

Filters allow it to skip information based on certain conditions which consist of a sigmoidal neural network layer and a pointwise multiplication operation. The sigmoid layer returns numbers from zero to one that indicate how much of each block of information should be skipped further down the network. Zero in this case means "do not let anything", one - "let everything". LSTM has three such filters to protect and monitor the cell state.

An advanced RNN model, bidirectional LSTM with attention mechanism which adds weights for importance of each input, was proposed in [43;44] and achieved good results in case of hate speech classification.

Despite the fact that the use of neural networks seems very promising for the task of texts classifying, in our case there is a problem - the topic of misogynistic messages classification is relatively new and there are few datasets for training the neural network at the moment. For this reason, we have chosen classic machine learning models in our research. However, in the future, with the availability of more suitable datasets, we also plan to use neural networks to improve the quality of our model.

2.2.3 Measures

Various measures are used to evaluate the results obtained in the data analysis. Although several exist, we describe in more detail here the features of the measures

which we have used in this thesis. These are accuracy [45], F1-score [46] and macro-F1-measure [47;48].

Accuracy

Accuracy is used to evaluate the performance of the algorithm, i.e. the proportion of documents for which the classifier made a correct decision. It is the ratio of the number of documents for which the classifier made a correct decision to the size of the training sample.

This metric has one feature that needs to be considered. It assigns the same weight to all documents, which may not be correct if the distribution of documents in the training sample is strongly biased towards one or more classes.

One way to deal with this problem is to train the classifier on a specially prepared, balanced corpus of documents. The disadvantage of this solution is that in this case we take from the classifier information about the relative frequency of documents.

Another way is to change the approach to formal quality assessment using precision and recall, which we will describe further.

F1-score and macro F1-score

In this case, precision (a proportion of documents actually belonging to this class relative to all documents that the system has attributed to this class) and recall (a proportion of documents found by the classifier belonging to the class relative to all documents of this class in the test sample) are metrics that are used in assessing most of the algorithms for extracting information.

It is clear that the highest precision and recall are the best. But in real life, maximum precision and recall are not achievable at the same time and we have to look for a balance. F1 and macro-F1 scores are calculated for this purpose.

To evaluate the results obtained by the modeling, we used the macro F1-score, which is well suited in the case of text 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:

$$F = 2 \times Precision \times Recall \div (Precision + Recall)$$

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.

2.3 Approaches for Hate speech identification

Hate Speech has a key characterizations like virality and presumed anonymity which make it potentially more harmful in comparison with communication offline, so the challenge of hate speech identification became really critical now. Although approaches for hate speech controlling are different in different countries and depend of the local laws, it is obvious that such type of expressions must be taken under control and prevented.

It is important to mention some works where researchers had the aim to identify offensive language and achieved quite good results. In this section we describes some interesting researches.

2.3.1 Approach of Zhang and Luo

Some interesting results for the hate speech detection were shown in an article by Zhang and Luo [49]. Firstly, authors analyzed data from Twitter and made a conclusion that detecting hateful content compared to non-hate using linguistic characteristics is quite difficult because of absence unique discriminative features. Secondly, the authors proposed two new models for hate speech identification based on a deep neural networks approach.

For the analysis they used seven public Twitter datasets: five of them included three different types of tweets labels 'sexism', 'racism' and 'neither' and the number of messages in each dataset was 6559 - 18593 tweets. Another two datasets were contacted on 2435 tweets and 24783 tweets divided by 'hate' and 'non-hate' classes.

At the preprocessing step the researchers made spelling corrections, elongated word normalisation, segmentation hashtags for words and unpacked contractions, and then they lemmatised each word to return its dictionary form.

For the first experiments a special measure named ‘uniqueness’ score was created, which indicated the number of 'unique' words corresponding to each class (i.e. not occurring in other classes) which 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 and takes a value from 0 to 1. They then

created a scale of values for this measure in increments of 0.1 (the meanings 0, 0-0.1, 0.1-0.2, 0.2-0.3 and so on) and checked each tweet in all datasets 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 and it means that there is no discriminative features which could indicate hate speech because of the fact that people can write an offensive messages using different words.

In the second part of the research, the authors present two new models based on deep neural network approach. At the beginning they took standard CNN with three convolutional layers and afterwards added a new Gated Recurrent Unit (GRU) layer [50]. GRU is similar to LSTM, but the latter has three gates (input, output and forget gates), whereas GRU has only two gates (reset and update gates). This simpler structure allows it to train and generalise better on small data. The next model combine CNN and 'skipped CNNs'. The idea of 'skipped CNNs' is to ignore inputs at certain (consecutive) positions of the window. For example, applying a 2-gap to a size 4 window will produce [O,X,X,O] shape, where 'O' indicates an activated position and 'X' indicates a deactivated position in the window. The results of modeling authors compared with state-of-the-art results of CNN modeling and the results showed by SVM model. The results showed that this two models (CNN+GRU and CNN+'skipped CNNs') achieved the best F1-score on all datasets.

The authors also analyzed the errors which were indicated after the model's creation: the first type of error is connected with the situation in which the user writes a message which contains the potentially offensive word, but the meaning of all post is not offensive. The second type of error was when the text of the message was not insulting, but there was a link to an offensive content (making it an offensive tweet).

The last type of error was a misunderstanding between the authors and the compilers of the dataset: the authors believed that the post was not offensive, while in the dataset it was marked as offensive one. It is important to keep in mind the errors that may occur in order to minimize the probability of their repetition in further work.

2.3.2 Approach of Park and Fung

In [51], Park and Fung present three different models based on neural networks which show quiet good results in hate speech identification. The first of these is CNN-based CharCNN which is the character-level convolutional network, the second model, WordCNN, is the convolutional network in which a sentence segmented into words on input and converted into a 300-dimensional word2vec embedding trained on 100 billion words from Google News. The last model named HybridCNN is a combination of CharCNN and WordCNN with two inputs: characters and words. The idea of creating this model was in an observation that offensive tweets often contain either purposely or mistakenly misspelled words. All three models had 3 layers. The authors compared their results with the results of models based on Logistic Regression, Fast Text and Support Vector Machines using the F1-measure. The dataset for experiments was constructed from sexist tweets, racist messages and neither racist no sexist tweets.

The authors present two type of classification: one-step and two-step. The one-step classification was a classification for three different classes: 'racism', 'sexism' and 'none' and in this case the best results was shown by HybridCNN model with 0.827 F1-score, while the best result by classical model (LogReg) was 0.814. The two-steps

classification required splitting into two classes 'Abusive' and 'None' at the first stage and then dividing the class 'Abusive' into classes 'Sexist' and 'Racist' in the second step. In this type of classification the best was achieved by the LogReg model with 0.826 F1-score, while HybridCNN presented 0.807 F1-score. In this case, the authors combined HybridCNN and LogReg and improved the F1 to 0.818.

The work demonstrated that the combination of neural network based model and classic machine learning approach allows the achievement of very good results for the two-step classification and this result looks encouraging.

2.3.3 Approach of Badjatiya, Gupta, Gupta and Varma

Another interesting work presented machine learning methods and linguistic features for the aim of hate speech identification is [52]. The goal of the authors was to investigate how the application of deep learning methods could improve the results of classifying tweets as racist, sexist or neither. As baseline methods three broad representations were chosen: char n-grams, Term Frequency - Inverse Document Frequency (TF-IDF)[1] and GloVe[2] which are used often for the purpose of text classification.

In the first part of experiment authors created different type of models which combined tweet semantic embeddings and multiple classifiers such as Logistic

[1] https://scikit-learn.org/0.21/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[2] https://nlp.stanford.edu/projects/glove/

Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs)[3]. The best result was achieved by the combination TF-IDF + SVM.

The second experiment involved the use of three deep learning architectures: FastText (which is similar to the static Bag of Words model, but use the updates of the word vectors through back-propagation during training), Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs). For each approach there were 2 methods of word embeddings (random embeddings and GloVe embeddings) applied. The best combination of this type of modeling was CNN + GloVe.

In the third experiment the authors made an attempt to use the average of word embeddings learned by Deep Neural Networks as features for multiple classifiers. This results was the best one, the combination LSTM + Random Embedding + GBDT achieved the highest results with 0.930 F1-score. This result demonstrated that this type of combination (classical machine learning and deep learning) has a good perspective for the hate speech classification tasks.

2.3.4 Approach of Saleem, Dillon, Benesch and Ruths

The authors of the work [53] presented an alternative approach to hate speech identification. As opposed to the traditional key-words based techniques which aim to caught some slurs of traditional offensive language, they focussed on group conversations - hateful or not. The idea was in the fact that hate speech could be expressed not in slurs, it could be normal conversation, but with offensive meaning.

---

3 https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

The researchers chose three types of communities which could be the target for offensive language: African-American (black), plus-sized (plus) and women. They took the data from the Reddit[4] social network: for each target a group of active support and a group of haters. They then used TF-IDF [54] for preprocessing and deleted all URLs, stopwords, numerals and punctuations. For the modeling they used Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machines (SVM). Three main experiments were conducted to check the suggestion that approach based on user groups (support or haters) could identify hate speech with good results. In the first experiment the authors used data from haters' Reddit communities and random comment from Reddit (not hate speech) with 10-fold cross validation and achieved quite good results for all three classifiers (the best results for 'black' with SVM: 0.81 accuracy). For the next experiment, they used data only from Reddit support and hate communities and all three classifiers showed good results, but NB performed slightly better then others (for 'black' - 0.8 accuracy). The results showed that the approach based on specific groups of comments can allow us to distinguish hate speech from negative cases.

For the third experiment, authors add the data from another source to the research: they chose the Voat[5] platform (and other web forums) for adding comments and checking all three target groups. As negative comments there were messages from Reddit groups of haters like in the previous experiments and Logistic Regression was used for modeling. The results were reasonably good, and from these results we can conclude that it is possible to use data from other websites with the aim of classifying messages in cases where the dataset is balanced.

[4] www.reddit.com
[5] https://voat.co

2.3.5 Approaches summary

In this section we summarise the surveyed works devoted to hate speech identification in terms of features used, classification algorithm, and main results. The summary is presented in table:

Table 2.1 Summary of hate speech identification approaches

Authors of the approach	Features	Classification Algorithm	Main results
Zhang Z., Luo L.	spelling corrections, elongated word normalisation, segmentation hashtags for words and unpacked contractions, lemmatization of each word to return its dictionary form, 'uniqueness' score	CNN + Gated Recurrent Unit (GRU) layer, combination of CNN and 'skipped CNNs'	detecting hateful content compared to non-hate using linguistic characteristics is quite difficult because of absence unique discriminative features + two new models for hate speech identification based on a deep NN
Park J.H., Fung P.	word2vec, wordsegment library	CharCNN, WordCNN, HybridCNN (combination of CharCNN and WordCNN), Logistic Regression, Fast Text, Support Vector Machines	the combination of neural network based model and classic machine learning approach allows the achievement of very good results for the two-step classification
Badjatiya P., Gupta S., Gupta M., Varma V.	char n-gram, TF-IDF, GloVe	Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees, FastText, CNN, LSTM	combinations of classical machine learning and deep learning models achieve the best results for the hate speech classification tasks
Saleem H., Dillon K.P., Benesch S.,	TF-IDF, all URLs, stopwords, numerals and	Logistic Regression, Naive Bayes, Support Vector Machines	the approach based on specific groups of comments can allow us to

Ruths D.	punctuations were deleted		distinguish hate speech from negative cases, it is possible to mix datasets from different sources in case if datasets are balances

2.4 Automatic Misogyny Identification

The specific form of hate speech in social media being investigated in this thesis is misogyny, and its various manifestations. Misogyny is a concept that means hatred, hostility, an ingrained prejudice against women. It can manifest itself in different ways, and one of the most common manifestations is sexism, that is, a set of stereotypes and bias towards people based on gender. The ideology of sexism divides people into men and women, contrasts them with each other and directly or indirectly affirms the superiority of men over women. It explains the economic, social and political inequality between them by differences in their nature.

Misogyny can also be expressed in the form of sexual objectification of women in relation to another person exclusively as an instrument (object) for their own sexual satisfaction. Such objectification can be manifested both at the society level and at the level of individual communication.

Misogyny also includes violence against women and humiliation of women - that is, the direct consequences of hatred. Despite the fact that in this case we are talking about virtual space - about Internet platforms – rather than direct physical contact, it should be noted that messages containing signs of misogyny are insulting and can harm the addressee of such a message. Therefore it is important to protect users from

misogynistic posts. The development of approaches to performing Automatic Misogyny Identification has become an urgent necessity in social media, because of the speed at which such messages is very high and it is impossible to control and delete misogynistic posts 'by hand'.

There are a number of approaches which have been published in recent years which can help to detect misogynistic messages. We will describe the most important approaches for automatic misogyny identification in more detail in the following.

2.4.1 Approach of Waseem and Hovy

Waseem and Hoy [55] created a dataset consisting of sexists tweets. As part of their work, they proposed a definition of the conditions under which messages were regarded as misogynistic. We use this classification of misogynistic messages in our work. A post is labeled as misogynistic if:

1. a sexist slur were used
2. a minority were attacked
3. there was the aim to seek to silence a minority.
4. a minority criticized without a well founded argument
5. hate speech or violent crime were promoted but didn't directly use
6. a minority was criticized and it was used a straw man argument.
7. truth was misrepresented blatantly or there was the aim to seek to distort views on a minority with unfounded claims.
8. there was a support of problematic hashtags.
9. there were negatively stereotypes a minority.

10. sexism were defended.

11. an offensive screen name was used.

The authors extracted a number of features using message metadata. They highlighted the gender of users by looking at names in profiles and found that about half of all messages were written by men 2.26% were written by women and 47.64% had no indication. Using the time zones which were marked in tweets, the authors created geographic distribution feature and, also noted the length of tweets.

A model was created based on logistic regression which showed best results in cases where features of gender and geographic location were counted. It is interesting that in cases where gender, geographic location and length of tweet were taken into account at the same time the result were not as good.

This work allows us to understand better the process of hate speech datasets constructing and features from metadata extraction.

2.4.2 Approach of Fasoli, Carnaghi, and Paladino

Interesting linguistic patterns are presented in [56]. Here, the authors of the article analyzed two different group of slurs: Sexist Derogatory Slurs (e.g., b*tch) 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 (SDSs) was the class of slurs with the aim of derogate woman in a context of stereotypes, sexual looseness and promiscuity, while Sexist Objectifying Slurs (SOSs) was the class of words which reduced women

to objects of man's sexual interests. The main goal was to analyze the offensiveness and the perceived social acceptability of SDSs and SOSs.

There were 39 participants in the survey and 13 Italian words (bona (foxy), bagascia (cunt), baldracca (floozy), bambola (doll), figa (pussy), gnocca (hot-chick), pupa (babe), puttana (bitch), sbarbina (a term referring to young girls), sgualdrina (tramp), troia (whore), zoccola (slut), velina (showgirl)). For the first experiment participants were asked to evaluate how pleasant or derogatory the words were in three positive adjectives (i.e., pleasant, gratifying, respectful) and three negative adjectives (i.e., offensive, humiliating, derogatory). Also, participants were asked about how frequently this terms used and how they socially acceptable. In the second experiment participants were asked to evaluate this slurs in specific social settings (affective relationships, working relationships - same status and working relationships - higher status; user is man or woman). At the end of the survey participants were asked about they gender, age, political orientation and level of education.

From the study, the researchers created an index of offensiveness (higher the score, the more offensive the slur) and an index of social acceptability (higher the score, the more acceptable the slur). It was shown by this indexes that the SOSs and SDSs exactly two different classes of slurs. The authors then chose three the most frequent slurs for each group and found that people use SOSs more often then SDSs and that the gender of the user did not matter. They showed that people tend to evaluate SDSs as more offensive, compared to SOSs, and females judge slurs as more offensive than males. Correlational analyses was performed that the highest level of frequency of use is connected to the highest level of social acceptability. The results of the evaluation in the specific social settings showed that the slurs have more social acceptability in a

context of an affective relationships then in a equal work-related context and a higher work-related context. It is interesting that the SOSs had the lowest social acceptability in the work-relation situation with unequal positions (supervisor-subordinate).

The main result of gender factor was in the fact that slurs from women are more acceptable then from men.

This work is very useful, because it was shown that there are different significant groups of slurs (for example, SOSs and SDSs) and it is necessary to take into account this fact then we speak about offensive speech. Also, we have to consider the relationships between users, because how it was demonstrated some groups of slurs are more acceptable then another.

2.4.3 Approach of Hardaker and McGlashan

Another interesting work connected with linguistic features was presented in [57]. The authors had two main aims of their research: firstly to investigate the language surrounding sexual aggression on Twitter, and secondly to find any communities in response to that sexually aggressive language. They found that offensive language had mostly the aim of insulting woman, using sexual aggressive words (abuse, rape, threats had high frequency) and were used often with words such as getting, received, receiving, and adjectives such as awful, cowardly, disgraceful, despicable, graphic, hateful, and horrendous. It is also interesting that in this type of abusive messages the grammatical actor (who is performing the abusive action) was absent as usual, and the focus of the message was the person being vilified.

The authors investigated that there were a group of users with a lot of messages where the image of 'real man' was incompatible with abusive and threatening behaviour. They selected three different types of users (high-risk, low-risk, no-risk) and tried to compare their messages to find if high-risk users employ more offensive language. High-risk users contained in their Twitter profiles evidence of: intent to cause fear of (sexual) harm, harassment and potentially illegal behavior. Low-risk users had at their profiles little evidence of offensive material, insults, ridicule. No-risk users had not got any evidence of this facts. It was shown that the task to choose any group of users is very complicated and it is necessary to pay attention to the dynamic behavior of Twitter accounts and the fact that the type of account could change from low-risk to high-risk over time.

2.4.4 Approach of Clarke and Grieve

The article [58] shows the investigation of importance of functional linguistic variation in a corpus of sexist Tweets. The authors analyzed the role of lexical and grammatical features using MDA (multi-dimensional analysis), which is a method based on multi factor analysis and can reveal the grammatical and lexical features which are measured across a text corpus. They chose 81 features that occurred in at least 1% of tweets and subjected them to a multiple correspondence analysis (MCA) in R using FactoMineR, thus they had a positive and a negative scales for each linguistic feature which presented the relationships between frequencies of use this feature in sexist message in three different dimensions named interactive, antagonistic and attitudinal.

The first 'interactive' dimension showed how interactive or informative the message was. Was the aim of the message to involve reader in a discussion or to inform reader about any facts? The features with the highest score were 'Question mark' (when there is a symbol “?” in a message) , 'Question do' (when a message starts with a word “do”), 'Accusative case' (when an accusative case is used in a message) and the the lowest score had 'Existentials', 'Place adverbials', 'BE as main verb', and it showed that the most interactive tweets had a lot of questions, while informative message tended to present some facts.

The second 'antagonistic' dimension presented the attitude of the user to his readers: does he agree with them or not. The most frequent features in case when the user was antagonist for another users were 'Question DO', 'Question marks', '2nd person pronouns', and in situations when the user were agree with his readers were 'Subject pronouns', '1st person pronouns', 'Auxiliary B’'. It should be noted that 'agree' means that messages still were sexist, and the user had a communication with his friends who shared his point of view.

The last dimension was 'attitudinal' and presented the interpretation as representing the degree of attitudinal judgment exhibited by a tweet. The most frequent features were 'Predicative adjectives', 'Existentials', 'absence of Prepositions' which showed the users opinion, and features with the lowest score were 'Auxiliary BE', 'Progressive aspect', 'Hashtags' which indicated that user told some story or recounted any facts.

From this article we can make a conclusion that the most 'popular' linguistic feature in offensive language are 'question marks' and 'question DO'. Authors of sexist messages tend to write more personal tweets, and we should pay more attention for this construction in future research of sexist messages.

2.4.5 Approaches summary

In this section we summarise the surveyed works devoted to automatic misogyny identification in terms of features used, classification algorithm, and main results. The summary is presented in table:

Table 2.2 Summary of automatic misogyny identification approaches

Authors of the approach	Features	Classification Algorithm	Main results
Waseem Z., Hovy D.	marked gender of users, geographic distribution feature was created and, also noted the length of tweets	Logistic Regression	best results in cases where features of gender and geographic location were counted
Fasoli F., Carnaghi A., Paladino M.	index of offensiveness and index of acceptability	Factor Analyses	the slurs have more social acceptability in a context of an affective relationships then in a equal work-related context and a higher work-related context. slurs from women are more acceptable then from men.
Hardaker C., McGlashan, M.	lists of positive and negative keywords	Corpus Linguistics	offensive language had mostly the aim of insulting woman, using sexual aggressive words (abuse, rape, threats had high frequency) and were used often with words such as getting, received, receiving, and adjectives such as awful, cowardly, disgraceful, despicable, graphic, hateful, and horrendous. In this type of abusive messages the grammatical actor (who is

			performing the abusive action) was absent as usual, and the focus of the message was the person being vilified
Clarke I., Grieve J.	81 linguistic and grammatical features	multi-dimensional analysis, multiple correspondence analysis	the most 'popular' linguistic feature in offensive language are 'question marks' and 'question DO'. Authors of sexist messages tend to write more personal tweet

2.5 Shared tasks on Misogyny Identification

The first concerted efforts aimed at increasing the activities of researchers in the field of misogyny detection commenced in 2018. Such challenges are important because they allow us to identify patterns in this kind of messages on real data, as well as provide an opportunity to understand which methods and models give the best results for solving this problem.

In this work, we describe three Shared Tasks whose aim was to create a model that allows to reveal misogyny in social networks based on datasets received from Twitter. The author of this thesis participated in each of these tasks, the approaches and results of which will be described in subsequent chapters.

2.5.1 AMI@IberEval

The first challenge to be held on misogyny detection was the Automatic Misogyny Identification (AMI) shared task, held in conjunction with the IberEval 2018 workshop [59] (it is a workshop aiming at encouraging and promoting the development of Human Language Technologies). The aim of this challenge was to indicate misogyny behavior on the Twitter platform, written in English or Spanish. Separate training datasets were made available for each language, and these are described in Chapter 3. The challenge contained two different subtasks:

- Subtask A - Misogyny Identification: to separate misogynistic tweets from non-misogynous using binary classification.
- Subtask B - Misogynistic Behavior and Target Classification.

The purpose of Subtask A is clear. Regarding Subtask B, the idea of the target classification was to identify whether a misogynous tweet offends a specific person or a group of people. It was a binary classification task: one class containing tweets with active (or individual) offenses which were sent to or referenced a specific person. The other class consisted of texts with passive (or generic) offenses with were posted with the aim of offending a group of people.

The misogynistic behavior task was intended to divide misogynistic tweets to different groups, included four categories:

- Stereotype & Objectification - a description of women's physical and/or comparisons to narrow standards,
- Dominance - an assertion the superiority of men over women,
- Derailing - abuse of a woman,

- Sexual Harassment & Threats of Violence - actions as sexual advances, requests for sexual favors, harassment,
- Discredit - slurring over women with no other larger intention.

Some examples of different types of misogynistic tweets are presented in Table 2.3.

Table 2.3 Examples of tweets for each misogyny category

Misogyny Category	Tweet
Discredit	@melaniatrump stupid fuc**ing bi**ch
Stereotype and Objectification	Women are good only into the kitchen..#makemeasandwich Big te**ts and no brain..this is what I want as wife!
Sexual Harassment and Threats of Violence	I wanna touch your te**ts..you can't immagine what I can do on your body. Stupid bi**ch I'll put you down on the floor and I'll rape you! You should be scared!
Dominance	Women are inferior to men..so shut up please!
Derailing	@yesallwomen wearing a tiny skirt is "asking for it". Your teasing a (hard working, taxes paying) dog with a bone. That's cruel. #YesAllMen

In this challenge, 11 different teams from five countries took part in Subtask A and a total of 32 runs were received for English and 24 runs for Spanish datasets. For Subtask B, 9 teams for the English dataset participated, presenting 28 runs, and 6 teams for the Spanish dataset, submitting 20 runs. Subtask A and Subtask B were evaluated separately for each of the languages.

Table 2.4 presents achieved by the top 3 participating teams, showing the results for the English and Spanish datasets for Subtask A.

Table 2.4 The best results for the Subtask A (IberEval)

English		Spanish	
Team	Accuracy	Team	Accuracy
14-exlab	0.913	14-exlab	0.815
SB	0.871	JoseSebastian	0.815
AnotherTeam	0.793	SB	0.813

Table 2.5 presents the best results in Subtask B for the English and Spanish datasets.

Table 2.5 The best results for the Subtask B (IberEval)

English		Spanish	
Team	macro F1-measure	Team	macro F1-measure
SB	0.442	14-exlab	0.446
14-exlab	0.369	SB	0.441
Resham	0.351	JoseSebastian	0.433

2.5.2 AMI@Evalita

The next AMI challenge was held shortly afterwards, at the Evalita 2018 workshop [60]. Evalita is an evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. It was organised by several of the same team who had proposed the previous AMI shared task, and had the same Subtasks A and B as its predecessor. In this case, there were datasets in the English and Italian

languages. This allowed participants to check the stability of their models in the classification. A detailed description of the datasets we used will be presented in the next Chapter.

There were 13 submissions for Italian and 26 runs for English submitted respectively from 6 and 10 teams for the Subtask A and 11 submissions by 5 teams for Italian and 23 submissions by 9 teams for English for the Subtask B.

The best results of the Subtask A for the English and Italian datasets are presented in Table 2.6.

Table 2.6 The best results for the Subtask A (Evalita)

English		Italian	
Team	Accuracy	Team	Accuracy
Hateminers	0.704	Bakarov	0.844
Resham	0.651	CrotoneMilano	0.843
Bakarov	0.649	14-exlab	0.839

The best results for the Subtask B using the English and Italian datasets are presented in Table 2.7.

Table 2.7 The best results for the Subtask B (Evalita)

English		Italian	
Team	Accuracy	Team	macro F1-measure
Himani	0.406	CrotoneMilano	0.501
CrotoneMilano	0.369	Bakarov	0.493
Hateminers	0.369	14-exlab	0.485

2.5.3 SemEval

SemEval is an ongoing series of evaluations of computational semantic analysis systems, intended to explore the nature of meaning in language. As part of SemEval-2019,two challenges relevant to misogyny detection were held (and in which we participated). These are SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter (HatEval[61]) and SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media shared task (OffensEval [62]).

HateEval 2019 consisted of two subtasks, one of which was the binary classification between offensive and non-offensive messages in cases of hate speech detection against immigrants and women. The second task proposed to make a aggressive/non-aggressive and individual target/a group target classification on the offensive messages.

Some examples of different types of messages are presented in Table 2.8.

Table 2.8 Examples of tweets with the HatEval dataset

Type of tweet	Text of 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 b*tch? Did I ever deny that? Nope Next.
Group target	The German Government Pays for 3 Week Vacation for Refugees to Go Home Muslim Immigration 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.

The second task at the SemEval 2019 challenge is OffensEval 2019. The challenge had 3 different subtasks:

SUB-TASK A - Offensive language identification:

- Not Offensive - This post does not contain offense or profanity;

- Offensive - This post contains offensive language or a targeted offense.

SUB-TASK B - Automatic categorization of offense types:

- Targeted Insult and Threats - an insult or treat to an individual, a group, or an organization,

- Untargeted - non-targeted profanity and swearing.

SUB-TASK C - Offense target identification:

- Individual - The target of the offensive post is an individual,

- Group - The target of the offensive post is a group of people considered as a unity,

- Other - The target of the offensive post does not belong to any of the previous categories (e.g., a situation, an event, or an issue, but in the challenge the group).

Some examples of different types of tweets are presented in Table 2.9.

Table 2.9 Examples of tweets with the OffensEval dataset

Type of tweet	Text of 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

For the HatEval challenge there were a total of 108 submitted runs for Subtask A and 70 runs for Subtask B. 74 different teams submitted their runs, of which 22 teams participated to all the subtasks for the two languages.

2.5.4 Approaches for AMI@IberEval open challenge

In this subsection, we will summarise the approaches taken by the most successful participation teams in AMI@IberEval. For Subtask A, the best results were achieved by the *14-exlab* [63] team. The team used SVM models: with Radial Basis Function (RBF) kernel for the English dataset and SVM with a linear kernel for the Spanish dataset. The team *AnotherOne* used just SVM for modelling.

The best lexical features for the English corpus were shown by the *14-exlab* team and included Swear Word Count (a representation of the number of swear words contained in a tweet), Swear Word Presence (a binary value representing the presence of swear words), Sexist Slurs Presence (a small set of sexist words aimed towards women was used), Hashtag Presence (a binary value equals 0 if there is no hashtag in the tweet or equals 1 if there is at least one hashtag in the tweet).

For Subtask B (tweet classification by different types of misogyny and target classification: active or passive types) the best results were achieved by the *SB team* [64] with 0.44 average F-Measure. The teams created the best models using SVM with linear kernel and an ensemble model which combined SVM, Random Forest and

Gradient Boosting classifiers. Also, the team created lists of specific lexicons concerning sexuality (p*ssy, c*ncha), profanity, femininity (some words which could be used in negative sense like gallina, blonde) and the human body (having a strong connection with sexuality) and Abbreviations and Hashtags lists which included typical for the Internet slang words like 'smh'.

It should be noted that the best results of evaluation were achieved for different datasets (English and Spanish ones) using different approaches. For example, the *JoseSebastian team* [65] achieved 10th position for the English dataset, but the top result for the Spanish one with 0.82 accuracy for the binary classification in Subtask A using the SVM model. They replaced all hashtags with the keyword HASHTAG and some of them which are known as misogynistic ones with keyword MISO_HASHTAG. The authors note that the large difference between the results for the English and Spanish datasets could have a close connection with the choice of misogynistic hashtags in different languages.

The *Resham* team [66] used both an ensemble of models and neural networks for their modeling. They presented two approaches to deal with the challenge, the first of them was to create an ensemble of models including Logistic Regression, Support Vectors Machine, Random Forest, Gradient Boosting and Stochastic Gradient Descent models, and the second idea for modeling was to apply Word-level and Document-level Embedding and Recurrent Neural Network. The authors applied the Continuous Bag of Words (CBOW) approach to create words vectors, so they collected 20 words which are potentially misogynistic (like b*tch, sl*t) and downloaded 20,000 tweets which contained these words with the aim of finding the closest connection words. They achieved 100-dimensional word vectors and

300-dimensional word vectors for modeling. Also, they did the same for Document-level Embedding, presenting the whole tweet as a word (while with the Word-level approach each word was count as word) . The result of this unsupervised type of modeling were promising and the authors note that the accuracy could be higher in condition of using extended labeled dataset.

2.5.5 Approaches for the AMI@Evalita open challenge

The models with the highest accuracy for the Subtask A were presented by models of Logistic regression from the *Bakarov* team [67], and the ensemble of models from the *Resham* team [68], so we can conclude that the best models in this case are based on the classical machine learning approach.

In the case of Subtask B, the best models were also created using classical machine learning: ensemble of models by the *Himani* team, Support Vector Machines by the *CrotoneMilano* team [69] and Logistic Regression by the *Hateminers* team. Also, it should be noted that the *Bakarov* team made the text classification based on using semantic features obtained from vector space models of texts. They used a factorization of the term-document matrix (the method of singular value decomposition) and a normalization of factorized values. As an interesting feature, the *CrotoneMilano* team calculated the length of words and took it into accounting during the experiments.

It should be noted that the results achieved by the participating teams in AMI@Evalita were generally better than the results achieved for the previous IberEval challenge, and this difference can be explained by the fact that the collective

knowledge of the IberEval approaches was published at the time of model preparation, and consequently the quality of the datasets and modeling were higher.

2.5.6 Approaches for SemEval-2019 open challenges

For Task 5 (HatEval) the best result for Subtask A was achieved by the *Fermi* team [70] with 0.651 macro-averaged F1-score. Researchers used SVM model with RBF kernel only on the provided data, exploiting sentence embeddings from Google's Universal Sentence Encoder as features. Another models applied successfully were based on Neural Network models and, more specifically, Convolutional Neural Networks (CNNs) and Long Short Term Memory networks (LSTMs). For the Spanish dataset the best result 0.73 macro F1-score was achieved using a linear-kernel SVM trained on a text representation composed of bag-of-words, bag-of-characters and tweet embeddings by *Atalaya* team [71].

For Subtask B the best results were achieved using SVM and some special features such as sentiment lexicon and Word Count (LT3 team with 0.570 [72] and Logistic Regression (CIC-1 team with 0.568 for the English dataset and 0.705 for the Spanish one [73]).

For the OffensEval challenge, there were 104 participating teams at the first subtask, 71 teams at the Subtask B and 66 teams took part in the Subtask C. The best results were shown by *NULI team* [74] with 0.829 macro F1-score at the Subtask A, 0.716 - Subtask B and 0.569 - Subtask C. The model created were built using the Bidirectional Encoder Representation from Transformer (BERT), and also using a

number of preprocessing techniques such as hashtag segmentation and emoji substitution.

Also, good results were shown by *NLPR@SRPOL* [75] team with 0.803 macro F1-score at the Subtask A, 0.692 - Subtask B and 0.628 - Subtask C. The team ensembles of OpenAI GPT, Random Forest, the Transformer, Universal encoder, ELMo, and combined embeddings from fast-Text and custom ones. Also, they used external datasets to train the model.

The team *vradivchev_anikolov* [76] also had good results with 0.815 macro F1-score for the Subtask A, 0.667 - Subtask B and 0.660 for the Subtask C using the BERT model [77].

2.6 Conclusion

In this Chapter, we have described the problems of offensive and misogyny language recognition in social media, analysed the principal approaches to this issue, as well as the challenges devoted to this topic, which have been held recently. It is worth noting that such challenges have not held before, as the problem of insults in social networks, as it is only relatively recently become such a concern. It is critical that society, including the scientific community, reacts to the emergence of this problem and tries to find the best ways to counteract it.

We have analyzed the various subtasks proposed in the challenges, and we can say that the task of identifying misogynistic messages in social networks is not only a binary classification into offensive and non-offensive messages, but also there are

cases of multi classification, when the researcher needs to identify a certain type of misogyny.

We note that the methods we considered and which can be applied to the problem of offensive and misogyny language recognition in social media, such as models based on neural networks and models of classical machine learning, are heterogeneous and work better or worse depending on the situation, which was clearly demonstrated by the example of the models of challenge winners analysis.

In connection with this fact, it is worth noting that the use of models based on neural networks work best with a large amount of data. The training datasets for misogyny recognition in social media are just beginning to appear and they are typically quite small both in quantity and size. For this reason, we believe that at this time, the best approach to take is based on classical machine learning models. Furthermore, this allows us to reduce data processing time and to work productively with a small number of data.

Chapter 3

Datasets

In this chapter, we describe in more detail the datasets that were proposed for research in the challenges presented above. These datasets formed the main base for training models we created.

3.1 AMI@IberEval English and Spanish datasets

Three different approaches to sorting messages were used to create the training and testing IberEval datasets. Firstly, key words (profanity) were used to search for offensive tweets. Secondly, the accounts of potential victims of misogyny (for example, profiles of active feminists) were tracked. Thirdly, the posts of already identified misogynists were used. Messages in datasets cover the time period from July 20, 2017 to November 30, 2017. During the selection, 83 million English-language messages and 72 million Spanish-language posts were selected. Following this, the messages were annotated by two experts, and in cases of disagreement, a third expert was involved. The remaining tweets were marked by majority rule with the participation of CrowdFlower[6] platform.

As a result of the labeling, training datasets were created which consisted of 3251 English and 3307 Spanish messages. The testing datasets consisted of 831 posts for Spanish and 726 tweets for English. The data included the following fields:

- User's ID;
- tweet text;
- "misogynous" field, where it was 1 in case the tweet was misogynistic, and 0 if not;

[6] https://figure-eight.com/

- “misogyny category” field, which took on the values “stereotype”, “dominance”, “derailing”, “sexual_harassment”, “discredit” or 0 in case then the message was non-misogynous;

- “target” field, with values “active” for individual target of offense, “passive” for a generic target or 0 in case then the tweet was non-misogynistic.

The distribution of English tweets for different types is presented in Table 3.1.

Table 3.1. AMI@IberEval English dataset

Training dataset	Testing dataset	Type of tweet
1568	283	Misogynistic
1683	443	Non-misogynistic
943	123	Discredit
410	32	Sexual Harassment & Threats of Violence
29	28	Derailing
137	72	Stereotype & Objectification
49	28	Dominance
942	104	Active target
626	179	Passive target

The distribution of Spanish tweets for different types is presented in Table 3.2.

Table 3.2 AMI@IberEval Spanish dataset

Training dataset	Testing dataset	Type of tweet
1649	415	Misogynistic
1658	416	Non-misogynistic
978	287	Discredit
198	51	Sexual Harassment & Threats of Violence
20	6	Derailing
151	17	Stereotype & Objectification
302	54	Dominance
1455	370	Active target
194	45	Passive target

It should be noted that the data for classification according to the category of misogyny/non-misogyny are quite balanced. The prevailing type of misogyny is discredit, and the most common goal of misogynistic tweets is individuals.

3.2 AMI@Evalita English and Italian datasets

In order to construct the Italian and English datasets, the authors of the challenge took the following actions:

- messages containing relevant offensive language in English and Italian were downloaded (insults here acted as keywords);

- profiles of potential victims of misogyny were monitored (e.g. gamergate victims);

- for dataset creation, tweets of accounts that were already identified as misogynists before were used;

As a result, 10,000 tweets in each language were selected. Further, the data was annotated by six experts using the same platform as in the AMI@IberEval challenges. The inter-rater annotator agreement on the English dataset for the fields of "misogynous", "misogyny category" and "target" was 0.81, 0.45 and 0.49 respectively, and the inter-rater annotator agreement for the Italian dataset was 0.96, 0.68 and 0.76 respectively.

As a result of the selection, the final training datasets for English and Italian included 4000 messages, while the testing datasets consisted of 1000 posts for each language. The distribution of English tweets for different types is presented in Table 3.3 and the distribution of Italian messages is presented in Table 3.4.

Table 3.3 AMI@Evalita English dataset

Training dataset	Testing dataset	Type of tweet
1785	460	Misogynistic
2215	540	Non-misogynistic
1014	141	Discredit
352	44	Sexual Harassment & Threats of Violence
92	11	Derailing
179	140	Stereotype & Objectification
148	124	Dominance
1058	401	Active target
727	59	Passive target

Messages for various types of misogyny were presented in the same form as in the AMI@IberEval challenge, and the values for the "misogyny category" were "discredit", "sexual_harassment", "derailing", "stereotype", "dominance" or 0 in case of non-misogynous tweets. For the target there values were "active" in case of individual target of offence, "passive" in case of generic target or 0 if a message did not include misogyny.

Table 3.4 AMI@Evalita Italian dataset

Training dataset	Testing dataset	Type of tweet
1828	512	Misogynistic
2172	488	Non-misogynistic
634	104	Discredit
431	170	Sexual Harassment & Threats of Violence
24	2	Derailing
668	175	Stereotype & Objectification
71	61	Dominance
1721	446	Active target
107	66	Passive target

3.3 HatEval (SemEval Task 5) dataset

The HatEval datasets were compiled in order to detect texts against women and immigrants in social media. Twitter posts in English and Spanish were presented. Several strategies were used to select messages for the datasets. On a timeline, most posts were selected from July to September 2018, and tweets from earlier periods were also used.

To collect messages, firstly, the accounts belonging to potential victims of insults were monitored, secondly, the message history of previously identified haters was examined, and thirdly, all messages were filtered using keywords such as words, hashtags and stems.

In the case of selection of texts based on keywords, both neutral words and obviously offensive words were used, as well as highly polarized hashtags.

The entire dataset is composed of 19600 tweets, 13000 for English and 6600 for Spanish. They are distributed across the targets as follows: 9091 - immigrants and 10509 - women.

During the annotation process, offensive messages were marked with 1 if there was hate speech, 0 if not. If the target of offence was an individual, the tweet was marked with 1, and 0 if there was group target. Additionally, in cases then there was aggressive offence, tweets were marked with 1, and 0 if not.

There were at least three annotators for each message, so there were three independent judgments for each tweet. Also, the default F8[7] settings for assigning the majority label were adopted and the average confidence on the English dataset were 0.83, 0.7 and 0.73 (hate speech, target and aggressiveness respectively), while for the Spanish dataset the values were 0.89, 0.47 and 0.47 respectively. There were two additional judgments for each messages provided by native or near-native speakers of British English and Castilian Spanish crowdsourcing specialists. The final label for a message was based on majority voting from crowd, expert1, and expert2. Each post was identified with a special numerical label which substituted the original Twitter's IDs.

[7] http://www.figure-eight.com/

The distribution of messages for English and Spanish datasets is presented in Table 3.5.

Table 3.5 HatEval dataset

English		Spanish		Type of Tweet
Training dataset	Testing dataset	Training dataset	Testing dataset	
4210	1260	2909	660	Hate speech
5790	1740	2060	940	Non-hate speech
1560	522	1254	423	Individual target
8440	2478	3715	1363	Group target
1763	590	3308	474	Aggressive
8237	2410	1661	1126	Non-aggressive

3.4 OffensEval (SemEval Task 6) dataset

To create the training and testing OffensEval datasets, the Offensive Language Identification Dataset (OLID) dataset [78] was used. In this case, a three-stage hierarchical annotation model was proposed and each of the three levels was used for the OffensEval subtasks.

The first task was to separate offensive messages from non-offensive ones. If the message contained insults, threats, and posts containing any form of untargeted profanity, it was marked as “OFF” - offensive, in other cases the mark was “NOT” - non-offensive.

The second aim was to indicate if an offensive message had the target of offense: in case then a message contained an insult/threat to an individual, group of other it was marked as “TIN” - targeted insult, in case then a message contained swearing words and profanity, but had not got a specific target, the mark was “UNT” - untargeted.

The third task was to highlight the type of targeted offensive messages: if there was an individual target a post had label “IND”, and if a message insulted a group of people the label was “GRP”.

The distribution for the offensive/non-offensive messages and offensive posts with individual/group target is presented in Table 3.6.

Table 3.6 OffensEval dataset

Training dataset	Testing dataset	Type of tweet
4400	240	Offensive
8800	620	Non-offensive
2407	100	Individual target
1074	78	Group target

It should also be noted that in the OffensEval dataset all references were anonymized and replaced with the string URL.

Chapter 4

Models and Experiment Designs

In this chapter we present our approach to solving the problem of misogyny detection, and the experiments we undertook in order to demonstrate its viability and performance. There are two main steps of our experiments: preprocessing and classification using various models and model combinations. 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.

This chapter is organised as follows: in Subsection 1 we give a brief explanation of the experiments we are going to do and the targets of the experiments. Subsection 2 explains the preprocessing steps which helps to prepare data to experiments. In Subsection 3 we explain the models we use for the experiments. Subsections 4 and 5 describes the additional experiments with external links and multilingual corpora respectively, and Subsection 6 summarizes the methods and approaches we use in our work.

4.1 Introduction

As noted previously, the problem of misogyny detection in social networks is quite new, but remains an extremely important issue to deal with., For the experiments, we have to determine clearly which type of classification we want to get in the modelling part of the work and which models and tools will serve our needs best. In this chapter we will describe in detail the purposes of our modeling and the methods that we use for this purpose.

In all the experiments we conducted, we set ourselves two goals of classification: the first goal (and the first type of experiments) was to determine whether a message contained misogyny or not.

The second goal (and the second type of experiments) was to identify what the target of the misogynistic message was: whether the insult was directed at a particular person or whether it was a more general target. All experiments were conducted firstly on a training dataset, and subsequently on a test dataset.

It should be noted that in the experiments under training and test datasets, we obtain the following datasets: the training dataset is the original dataset, broken down in the proportion of 80:20, where we train the model on the biggest part of messages, and then check the accuracy of the resulting model on the remaining messages. This allows us to test our original hypothesis of choosing the best model (with the highest results) of all the models that take part in the experiments. Then we apply the best model to the test dataset proposed by the challenges' organizers.

4.2 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. It is intuitively clear that a user name in this case does not carry a semantic load and can be replaced. Of course, sometimes a name can be

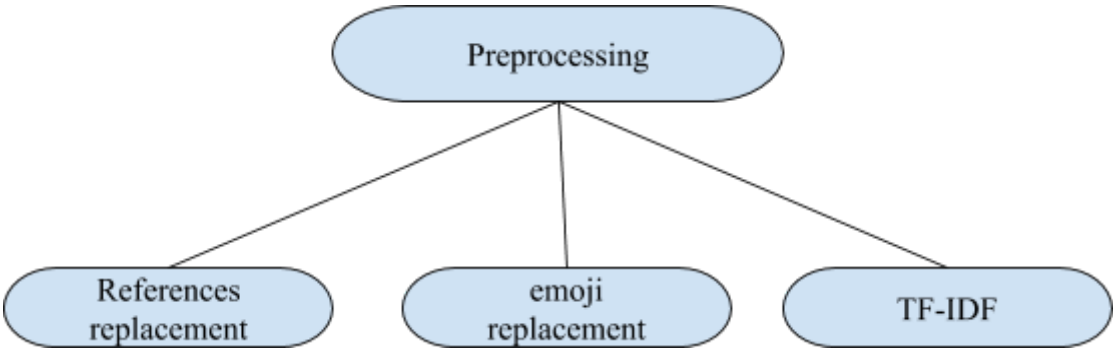
useful in analysis, for example, when it already contains an insult, but at the moment there is not enough research to confirm a correlation of this kind. Also note that in most cases, the user name was used only to personalize the message, so with the USER marker we emphasize that the message most likely had a target. This tagging can be particularly useful when we want to identify the purpose of the offending message - whether it was a particular individual.

- we labeled some combinations of symbols with were used often in messages such as !!!,??? and replaced them with the term emoji. Often users use such combinations as an imitation of offline speech - if in live communication the user would raise his voice and shouted or protested or in any other way showed aggression, in network communication he can probably use such combinations of characters, and thus we mark the strong emotionality of the message, which can be associated with offence.

We used TF-IDF (where TF is term frequency, and IDF inverse document frequency), a statistical measure, 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.

All steps of the preprocessing are shown in Figure 4.1.

Figure 4.1 The preprocessing steps



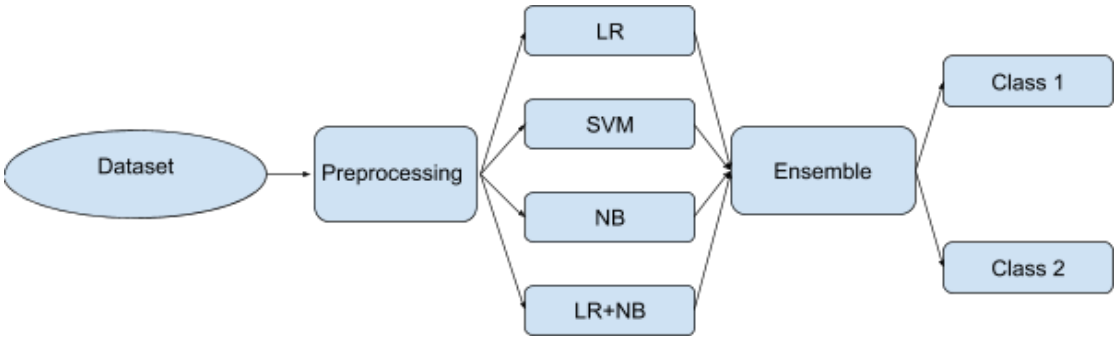
4.3 Core Experiments with Ensemble Model

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, Support Vector Machine, Naive Bayes and a combination of Logistic Regression and Naive Bayes models.

When constructing the ensemble, we compared the results obtained during modeling on the basis of each model, and the final result (the belonging of the message to a particular group) was determined by the majority rule: the message was assigned to the group when the majority of models voted for this group.

All stages of the modeling and experiments are shown in Figure 4.2.

Figure 4.2 The modeling and experiments stages



First, the preprocessing steps described above were applied to the original datasets, following which the simulation was carried out using the classical machine learning models. These were subsequently combined into an ensemble. As a result, the final labelling of messages (belonging to a group) was put down according to the rule of the majority of votes.

Due to the fact that we used some classical machine learning models, it is necessary to explain in more detail why these models were chosen by us.

4.3.1 Logistic Regression (LR)

As we noted above, classification models based on Logistic Regression are very popular when working with the analysis. The basic idea of a linear classifier is that a feature space can be divided by a hyperplane into two half-spaces, in each of which one of the two values of the target class is predicted.

If this is possible to do without errors, then the training sample is called linearly separable. Obviously, in our case, the probability of classification error is quite high, but we include this classifier in a number of models we used to create the ensemble.

4.3.2 Support Vector Machine (SVM)

The Support Vector Machine creates a hyperplane or set of hyperplanes in multidimensional or infinite-dimensional space that can be used to solve classification, regression, and other related problems.

When analyzing the results of the challenge on the definition of misogyny in social media, presented by us above, it was found that Support Vector Machines were used in many models, built by the winners of the challenges, and classifiers based on SVMs often achieve excellent results, and is also well applicable in the case of multi classification.

Accordingly, we also decided to include a classifier based on SVMs in the models from which we will create the final ensemble.

4.3.3 Naive Bayes (NB)

As it was mentioned previously, when analyzing the best approaches to solving the problem of offensive language recognition in social media, the Naive Bayesian classifier is quite simple and fast to use and is used as a reference point when comparing different methods. Its advantages include the fact that it is resistant to unknown words and thematic changes in documents, which is important in the case of text analysis in social media, where users prefer different, sometimes even non-existent words and terminology.

Based on these reasons, we chose the Naive Bayesian classifier as one of the models for misogyny recognition in Twitter.

4.3.4 The combination of Logistic Regression and Naive Bayes models (LR+NB)

In [79] it was shown that the combination of generative and discriminative classifiers demonstrates a strong and robust result in the task of texts classification. A model variant was presented in which 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 (which allowed us to achieve better results with our datasets in comparison with a combination of Support Vector Machines and Naive Bayes) 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.

4.3.5 Ensemble

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 our idea that the more models will classify the message as a 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.

We chose this method for determining whether a message belongs to a class, because this method allows us to avoid ambiguity in the estimation: for example, if we chose a summation method based on binary values (i.e., 1 if the model says that the message belongs to a class, and 0 if the model says that the message does not belong to this class), and given that 4 models are included in our ensemble, we could face a situation of a draw when voting (i.e., 2 models voted for one class and 2 models voted for another class), which would complicate our final choice of the class.

4.4 Experiments with Links

It is well known that classification of tweets is a particularly challenging task, due to their short informal nature [80]. We note that many tweets include a hyperlink, which is a URL to another message (in Twitter, this is usually another tweet). In order to improve the quality of the classification results, we hypothesized that the content of the referenced message is associated with the original message, and therefore by appending the referenced text to the original message, we can obtain a longer message which can improve the classification quality. Accordingly we added to the training datasets the texts of the messages referred to by the users in the original messages. This was possible in the cases of three datasets: AMI@IberEval, AMI@Evalita and HatEval. For the OffensEval dataset it was impossible, as all links in this dataset have already been replaced with the string @URL.

It is necessary to explain this feature in more detail. Table 4.1 provides examples of such messages we worked with. 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 also.

Table 4.1 Examples of tweets with additional data from the referenced link.

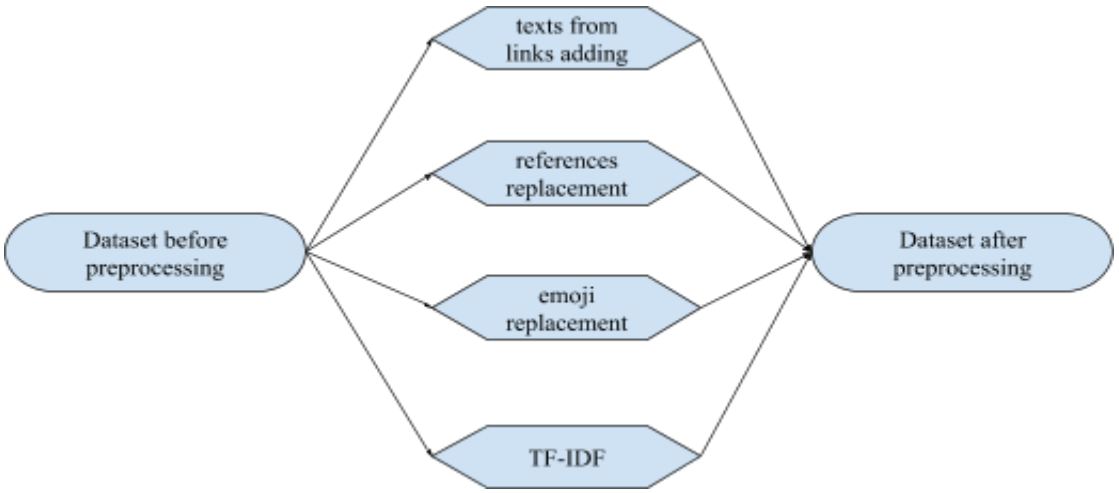
Tweet	Tweet from the link	Text to be classified
Thinking she a pretty decent b*tch but she a hoe prolly https://twitter.com/prettyindie/status/1180341053611794432	Thought she was a pretty ricky b*tch but she like yo gotti	Thinking she a pretty decent b*tch but she a hoe prolly Thought she was a pretty ricky b*tch but she like yo gotti
First of all sebody find a boyfriend for @USER. She is so f* lonely..where you don't https://twitter.com/shilpitewari/status/999352944565858310	believe in your stand but have other reasons influencing your thoughts.. you come up with these statements. Unbelievably unsmart.	First of all sebody find a boyfriend for @USER. She is so f* lonely..where 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?? https://twitter.com/natty_trolls/status/999317174408826884	GUYS! @USER is a coolest repotrer around and the coolest guy I know	Shes right..he is pretty awesome! @USER ..dont you agree?? GUYS! @USER is a coolest repotrer around and the coolest guy I know

In our work 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.

All steps of the preprocessing in this case are shown in Figure 4.3.

Figure 4.3. The preprocessing steps with the texts from links adding.



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 a lot of 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.

Therefore, in this category of experiments, we have made the replacement for all links which did not help with an extracting any additional information (the it was a link to the blocked or the external content) with the term URL.

4.5 Comparative Experiments with multilingual corpora

We tested the created model on multilingual datasets, Spanish and Italian, proposed as part of the AMI@IberEval and AMI@Evalita challenges mentioned above, to check the competitiveness of our model for different languages and also we used new additional feature to expand our existed datasets from the links in messages.

We want to design a fairly universal model that would allow us to achieve proper results regardless of the language in which the data is presented for analysis, and for this aim we use models and features (for example, adding data to a dataset using links in messages) that, in our opinion, should work equally well on datasets from different languages. With the help of experiments on the Spanish and Italian datasets and subsequent comparison of the results with the results obtained with the English datasets, we plan to test how universal the model we are able to create.

4.6 Conclusion

In conclusion of this Chapter we would like to summarize the methods and approaches that we use in experiments to identify misogyny in social networks and the targets of abusive messages.

First, we use the preprocessing step not only to process the input data for the subsequent modeling step, but also we use external data (via links in messages), which we assume will allow us to extend the original datasets and improve the accuracy of our classification.

Secondly, after analyzing the best approaches to the problem of offensive language identification in social networks, we propose to build an ensemble of classical

machine learning models to identify misogyny in Twitter. Although the use of neural networks for this purpose also allows to achieve good results, the size of our datasets is not large enough to speak about an achieving a classification with a stable high accuracy. Based on the analysis, we assume that classical machine learning models, and especially the ensemble of such models, will allow us to achieve higher results in classification.

Finally, we assume that the model we built will be versatile enough to show good classification results for datasets consisting of messages in different languages, not just in English one. To test this hypothesis, we use the final model on Spanish and Italian datasets.

It should be noted that model we created is very competitive in case of the task of misogyny detection in social media and it achieved good results in shared tasks devoted to the problem of misogyny recognition in social media. The new feature we used - additional information from links - allowed us to make the quality of results higher and it should be noted that nobody used this feature previously for this aims. We expect that the results will be better when the idea of misogyny detection will be more widespread and the datasets for this type of classification became to be bigger.

Chapter 5

Results and Analysis

In this chapter we present results we achieved with experiments on the English AMI@IberEval, AMI@Evalita, HatEval and OffensEval datasets. We also report and compare these with the results of the experiments on multilingual corpora and our experiments with the term expansion using text found in links. We present an overall analysis of these results.

5.1 Results for English Corpora

The largest amount of data that we used in the experiments were in English language, and it is necessary to consider in detail the results of the experiments for each dataset and analyze them.

5.1.1 Results for the English IberEval dataset

The classification results for the AMI@IberEval training dataset for the detecting misogyny and target classification are presented in Table 5.1.

Table 5.1 Results for the AMI@IberEval training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.78	0.73
Naive Bayes (NB)	0.60	0.59
LR+NB	0.79	0.75
Support Vector Machine (SVM)	0.72	0.70
Ensemble of models	**0.80**	**0.78**

It should be noted that the best results for both types of classification were achieved using an ensemble of models, including models of Logistic Regression, Naive Bayes, SVM and the interpolation between Logistic Regression and Naive Bayes.

The classification results obtained using the ensemble of models for the test dataset are presented in Table 5.2.

Table 5.2 Results for the AMI@IberEval testing dataset

Type of classification	macro F1-score with the training dataset	macro F1-score with the testing dataset
Misogyny	0.80	0.57
Target	0.78	0.55

For the AMI@IberEval dataset we should note that the results for the task of misogyny detection in messages turned out to be higher than the results obtained for the target classification, and this can be explained by the fact that the dataset for detecting misogyny was larger than the dataset for determining the purpose of the offensive message. Thus, we can conclude that our model begins to work better when the amount of data for training the model increases. We also note that in comparison with the work [81], where we used a smaller number of simpler models to construct the ensemble (we achieved the 9th place there with 0.758953 accuracy), a more complex ensemble of models turned out to be more promising for use. This suggests that we made the right choice in favor of increasing the number of models that we combine into the final ensemble.

5.1.2 Results for the English AMI@Evalita dataset

Table 5.3 shows the results of the misogyny classification and the target classification for the training Evalita dataset. In this case, we can also say that the ensemble of models shows the best results for both types of classification.

Table 5.3 Results for the Evalita training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.79	0.62
Naive Bayes (NB)	0.72	0.67
LR+NB	0.72	0.68
Support Vector Machine (SVM)	0.73	0.70
Ensemble of models	**0.79**	**0.70**

Table 5.4 shows the classification results of the training AMI@Evalita dataset using the ensemble of models.

Table 5.4 Results for the Evalita testing dataset

Type of classification	macro F1-score with the training dataset	macro F1-score with the testing dataset
Misogyny	0.79	0.58
Target	0.70	0.52

For the Evalita dataset, the results obtained using the test dataset, as well as in the case of experiments with the AMI@IberEval dataset, were lower than in experiments with the training dataset. Also, we note the similarity with the AMI@IberEval dataset: the results for classifying messages into misogynistic and non-misogynistic were

higher than the results obtained for the target classification. As we noted in [82] (where we achieved the 4th place with 0.638 accuracy), the results for the target classification can be improved in case when we carry out not an independent classification according to these two tasks, but sequential one ('non-independent-classification' means that on the first-step classification according to 'misogyny' - 'non-misogyny' we mark messages that were identified as misogynistic, and only then we make the target classification on the second step. 'Independent' classification means that we use created model to mark the messages as 'misogyny' - 'non-misogyny' type and by the type of the target independently).

5.1.3 Results for the English HatEval dataset

The results for the HatEval training dataset are presented in Table 5.5.

The experiments include misogyny recognition and the target of hate speech identification. Results for the task of misogyny 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.

Table 5.5 Results for the HatEval training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.52	0.65
Naive 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**

Experiments for the target classification also show that the ensemble of models achieves the best results with 0.72 score on the training dataset.

Table 5.6 presents our results for misogyny and the target of misogynistic messages identification using training dataset in comparison with the results we achieved using the testing dataset.

Table 5.6 Results for the HatEval testing dataset

Type of classification	macro F1-score with the training dataset	macro F1-score with the testing dataset
Misogyny	0.67	0.58
Target	0.72	0.64

As we can see, the results obtained with the training dataset are below the testing results by 1-9 percentage points for misogynistic language recognition and the difference between the results on testing and training datasets is only 1 percentage point.

The best published results for the HatEval dataset for the Hate Speech was 0.60 macro F1-score, while we achieved 0.58 macro F1-score, and for the Subtask B (where macro F1-score was calculated as (Misogyny + Target + Aggressiveness (our testing result equals 0.60))/3) our result were 0.61 macro F1-score, while the best published result was 0.60. Thus, the results we achieved show a competitiveness of the ensemble of models we created.

5.1.4 Results for the English OffensEval dataset

Table 5.7 shows the results we achieved with the OffensEval training dataset for misogyny and the target of the misogynistic messages detection. From the presented data we can see that the best results (0.70 F1-score for misogyny 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 the same best results using the interpolation between LR and NB with 0.25 coefficient of interpolation.

Table 5.7 Results for the OffensEval training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.63	0.57
Naive 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 5.8 shows the classification results for the training and the testing OffensEval datasets. The results achieved using the ensemble of models in the task of misogyny identification are quite similar for the training and for the testing datasets and have a difference in two percent point only.

Table 5.8 Results for the OffensEval testing dataset

Type of classification	macro F1-score with the training dataset	macro F1-score with the testing dataset
Misogyny	0.70	0.68
Target	0.73	(data unavailable)

We also note that, as was shown in [83], the model we proposed made it possible to achieve a result of 0.62 F1 score in the case of Subtask C in the OffensEval challenge (we achieved the 25th place of a total of 65 participants).

To compare the results of misogyny identification we achieved with HatEval and OffensEval datasets we can make a conclusion that the results 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 points for the training datasets.

Also note that the results achieved using the IberEval and Evalita datasets turned out to be lower for the same reason - these datasets contained less data for training models.

Also, with the increase of datasets from IberEval to OffensEval, the gap between the results achieved on the training and testing datasets was narrowing, while the quality of classification increased.

5.2 Results for the experiments with the links

Table 5.9 shows a comparison of the results obtained with the ensemble of models in case when we took the original datasets for the training and in the case when the messages that we extracted using the links were added to the original datasets.

Table 5.9 Results of the experiments with the links

Dataset	macro F1-score for misogyny identification without data from links	macro F1-score for misogyny identification with data from links
IberEval	0.58	0.80
Evalita	0.76	0.79
HatEval	0.59	0.67

It should be noted that we was able to use this additional feature only for IberEval, Evalita and HatEval English datasets, and for the OffensEval it was impossible because of the original type of datasets provided for experiments.

In the case of the IberEval dataset, there were 851 links and it was possible to extract information from 315 of tweets. For Evalita the dataset containing 1204 links, and it was possible to extract useful information from 122 links only. In the case of HatEval, the dataset had 1449 links, 523 of which were used for research.

Note that due to the dynamic nature of extracting available messages via links (until the data is blocked or the user has restricted access to them), it is impossible to predict how much information will be able to add to an existing dataset. For example, initially in the Evalita dataset there were more links than in the IberEval dataset, however, more information on them was obtained precisely for the IberEval one.

For all the datasets we experimented with, the results improved when we added data from the links. The largest increase was noted for the IberEval dataset. This can be explained by the fact that the largest amount (in percent) of link data was added to this particular dataset. We also note that the addition of data in any case leads to an increase in the classification quality, therefore, in the future it is necessary to study

this area of work in more detail. It would also be useful to try to use not only texts at accessible links, but also the meta-data contained in such messages (for example, mark messages that were blocked).

5.3 Results for Multilingual corpora

In this section we present the results we achieved on the multilingual datasets we had: Spanish IberEval dataset and Italian Evalita dataset and make the analysis of this results in comparison with the English ones.

5.3.1 Results for Spanish IberEval dataset

The results of experiments with the Spanish IberEval dataset are presented in Table 5.10.

Table 5.10 Results for the Spanish IberEval training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.79	0.71
Naive Bayes (NB)	0.61	0.58
LR+NB	0.68	0.71
Support Vector Machine (SVM)	0.73	0.73
Ensemble of models	**0.79**	**0.73**

The results achieved using the testing Spanish IberEval dataset in comparison with English IberEval dataset are presented in Table 5.11.

Table 5.11 Results for the Spanish IberEval testing dataset

Type of classification	macro F1-score with the Spanish training dataset	macro F1-score with the Spanish testing dataset	macro F1-score with the English training dataset	macro F1-score with the English testing dataset
Misogyny	0.79	0.55	0.80	0.57
Target	0.73	0.52	0.78	0.55

The results of the experiments on IberEval datasets show that for Spanish, as well as for English, the ensemble of models allows to achieve the best results for both types of classification: binary classification of misogynistic messages and classification for the purpose of insults.

In the case of experiments on the Spanish dataset, the results are slightly lower than in case of experiments with the Spanish dataset, but the same pattern is observed - a rather large gap between the results on the training and testing datasets. This can be explained by the fact that the Spanish and English IberEval datasets were approximately the same in size and did not contain a lot of messages, so the stability of the prediction is not high enough compared to large datasets.

5.3.2 Results with Italian Evalita dataset

The experimental results for the Italian Evalita training dataset in are presented in Table 5.12.

Table 5.12 Results for the Italian Evalita training dataset

Model	macro F1-score for Misogyny identification	macro F1-score for Target identification
Logistic Regression (LR)	0.75	0.60
Naive Bayes (NB)	0.71	0.64
LR+NB	0.72	0.67
Support Vector Machine (SVM)	0.72	0.70
Ensemble of models	**0.76**	**0.67**

The results for the testing Italian Evalita dataset in comparison of English Evalita dataset are presented in Table 5.13.

Table 5.13 Results for the Italian Evalita testing dataset

Type of classification	macro F1-score with the training Italian dataset	macro F1-score with the testing Italian dataset	macro F1-score with the training English dataset	macro F1-score with the testing English dataset
Misogyny	0.76	0.56	0.79	0.58
Target	0.67	0.51	0.70	0.52

The difference for the results achieved with testing English and Italian datasets was 2 percentage points in case of misogynistic messages classification and 1 percentage point in case of target classification. This confirms our assumption that at this stage of the work, the model we created is quite universal in working with different languages. In case of experiments with the Italian Evalita dataset, we also obtained lower results than when experimenting with the English dataset, but this difference is not large,

which may indicate that the model proposed by us is quite universal at this stage when we working with datasets in different languages.

Thus, our model shows fairly stable results regardless of the dataset language: English, Spanish, or Italian.

However, in the future it seems promising to use for the model improvement some linguistic features mentioned above that are not the same for different languages - for example, swearing words dictionaries, the uniqueness of which is obvious for each language. Despite the fact that this will affect the stability of our model for different languages, the use of such features can improve the classification results for the particular dataset.

5.4 Results overview

Table 5.15 shows all achieved results for all testing datasets using macro F1-score.

Table 5.14 Results overview for all datasets

Dataset	Type of Classification	
	Misogyny	Target
IberEval (English)	0.57	0.55
Evalita (English)	0.58	0.52
HatEval (English)	0.58	0.64
OffensEval	0.68	-
IberEval+links	0.80	-
Evalita+links	0.79	-
HatEval+links	0.67	-
IberEval (Spanish)	0.55	0.52
Evalita (Italian)	0.56	0.51

As we mentioned before, the results for misogyny identification are slightly better than the results for target identification, and the new feature - adding texts from links - allows us to achieve better results. The results achieved on the multilingual corpora (Spanish and Italian datasets) are quite similar with the results achieved on English datasets, which indicates the stability of our model.

Chapter 6

Conclusion

6.1 Achievements

In conclusion, it should be noted that the problem of hate speech detection in messages is a really important and urgent social problem. Misogyny, as a special type of hate speech, can occur in messages published on social media, and such cases can harm users of these Internet platforms.

The occurrence of this problem appeared only in recent years, and therefore the first systems that allow us to detect and classify misogynistic messages are only now being developed. In particular, for the construction of such systems, machine learning approaches are used, both classical models and models based on the deep learning approach, as well as models that are combinations of simpler ones.

We achieved the goals set at the beginning of the work, as the result:

- we researched the importance of the offensive language and misogyny language recognition in social media;
- we analysed the difference between offensive and misogyny language and the best existing approaches for detection and classification such language in social media;
- we analysed the recent shared tasks devoted to the problem of misogyny language recognition and highlighted the approaches which allow to achieve the best results in this case;
- we proposed an approach to the detection and classification of misogyny in texts, based on the construction of an ensemble of models of classical machine learning: Logistic Regression, Naive Bayes, Support Vector Machines. Also, at the preprocessing stage we used some linguistic features;

- we added additional texts from the links in the messages - which allowed us to improve the quality of our model;
- We demonstrated how efficiently the model we created worked not only in English datasets, but also in datasets in other languages, and we were convinced of the high quality of the results achieved. The model we created allows us to achieve competitive results compared with existing models of misogyny identifying and classifying based on real Twitter datasets.

6.2 Future Work

In the future, we plan to expand our research and improve the results, focusing on the following areas:

- To create a neural network and use it in the ensemble, since the use of neural networks looks promising solution to the problem of misogynistic messages identification and classification (as shown by other researchers in cases of offensive language recognition and classification).

- Our study revealed that adding information that can be extracted using links from existing messages increases the results of the classification. Also, it should be noted that the preprocessing step is really important, and that we could improve the results of our model by expansion of the new feature - texts from the links. We have to date just used the texts from the links when it is possible, but we can also use other information from the link - we could highlight situations when there is a link to a blocked tweet, external content or to an article/video and use these features at the preprocessing step as well.

- To add dictionaries with aggressive/harassment lists of words and to use other existed external linguistic resources - because, as was shown in the works mentioned above, the use of such features increases the classification accuracy - for us this is especially important in case of multilingual datasets.

In general, focusing on the model we have already created, we can say that the identification of misogyny is a difficult but feasible task in many cases, and we hope that this work and its further development will improve the process of identifying and classifying misogynistic messages in social media.

References

1 Nobata C., Tetreault J., Thomas A., Mehdad Y., Chang Y. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

2 Djuric N., Zhou J., Morris R., Grbovic M., Radosavljevic V., Bhamidipati N. 2015. Hate speech detection with comment embeddings. In *Proceedings of International World Wide Web Conference (WWW)*.

3 Gamback B., Sikdar U. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

4 Zhang Z., Robinson D., Tepper J. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.

5 Schmidt A., Wiegand M. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*, pages 1–10, Valencia, Spain.

6 Malmasi S., Zampieri M. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

7 Malmasi S., Zampieri M.. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.1

8 Genkin, A., Lewis, D., Madigan, D. 2007. Large-scale bayesian logistic regression for text categorization. In: P*roceedings of the NAACL Student Research Workshop. p. 49(3):291304. Technometrics.*

9 Wright, R. 1995. Logistic regression. *L.C. Grimm & P.R. Yarnold (Eds.) Reading and understanding multivariate statistics. Washington, DC: American Psychological Association,* 217-244.

10 Gaydhani A., Doma V., Kendre S., Bhagwat L. 2018. Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TF-IDF based approach. *ArXiv abs/1809.08651.*

11 Hazra T., Sarkar R., Kumar A. 2015. Handwritten English Character Recognition Using Logistic Regression and Neural Network. *5. 6-391. 10.21275/v5i6.NOV164228.*

12 Basu K., Nangia R., Pal U. 2012. Recognition of similar shaped handwritten characters using logistic regression. In: *Document Analysis Systems*, pp.200-204.

13 Zhang H., Li D. Naive bayes text classifier. granular computing. 2007. In: *GRC 2007 IEEE International Conference. pp. 708–708. IEEE.*

14 Awal M. A & Rahman M., Rabbi J. 2018. Detecting Abusive Comments in Discussion Threads Using Naïve Bayes. 163-167. *10.1109/ICISET.2018.8745565.*

15 Hasan K. M., Sabuj M., Afrin Z. 2015. Opinion mining using Naïve Bayes. *511-514. 10.1109/WIECON-ECE.2015.7443981.*

16 Mayfield L., Jones R. 2001. You're not from round here, are you? Naive Bayes detection of non-native utterance text. In *Proc. of the Second NAACL*, p. 239-246.

17 Pak A., Paroubek P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of LREC. 10.*

18 Metsis V., Androutsopoulos I., Paliouras G. 2006. Spam filtering with naive bayes - which naive bayes? In *Proceedings of CEAS,* 17, p. 28-69.

19 Joachims T. 2002. Learning to classify text using support vector machines: Methods, theory and algorithms. *Kluwer Academic Publishers.*

20 Scholkopf B., Burges C., Smola A. 1999. Advances in kernel methods : Support vector learning. *Cambridge: MIT Press.*

21 Vapnik V.1995. The Nature of Statistical Learning Theory. *Springer, New York.*

22 Schohn G., Cohn D. 2000. Less is more: Active learning with support vector machines. In *Proceedings of ICML*, pages 839–846.

23 Joachims T.1997. Text categorization with support vector machines: Learning with many relevant features. *Technical Report 23, Universität Dortmund, LS VIII*.

24 Chen Y., Zhou Y., Zhu S., Xu H.2012. Detecting offensive language in social media to protect adolescent online safety. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pp. 71–80. IEEE.*

25 Dadvar M., Trieschnigg D., de Jong F.2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: *Sokolova, M., van Beek, P. (eds.) AI2014. LNCS, vol. 8436, pp. 275–281. Springer, Cham.*

26 Vinodhini G., Chandrasekaran Dr. 2014. Opinion Mining using principal component analysis based ensemble methods. *CSIT. 2. 10.1007/S40012-014-0055-3*.

27 Xia R., Zong C., Li S. 2011. Ensemble of feature sets and classification algorithms for opinion classification. *Information Sciences 181:1138–115212*.

28 Li W., Wang W., Chen Y. 2012. Heterogeneous ensemble learning for Chinese sentiment classification. *Computing Sciences 9(15):4551–4558.*

29 Kim Y. 2014. Convolutional neural networks for sentence classification. 2014. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

30 Simard P., Steinkraus D., Platt J., et al. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3, pages 958–962.

31 Tompson, J., Goroshin, R., Jain A., LeCun, Y., Bregler C.. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656.

32 Mikolov T., Chen K., Corrado G., Dean J. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

33 Mikolov T., Yih W., Zweig G.2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL HLT.*

34 Akhtyamova, L., Cardiff, J., Alexandrov, M. 2017. Adverse Drug Extraction in Twitter Data using Convolutional Neural Network, In *28th International Conference on Database and Expert Systems Applications (TIR woskshop), Springer LNCS*, Vol 10438.

35 Tang D.,Qin B.,Liu T.2015. Document modeling with gated recurrent neural network for sentiment classification. In: *EMNLP*. pp. 1422–1432.

36 Mehdad Y., Tetreault, J.R. 2016. Do characters abuse more than words? In: *SIGDIAL Conference.* pp. 299–303.

37 Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993.*

38 Hochreiter, S., Schmidhuber, J.1997. Long short-term memory. *Neural Computation. 9 (8):* 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.

39 Gers, F. A., Schmidhuber, E. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. In: *IEEE Transactions on Neural.*

40 Wang, X., Liu, Y., Sun, C., Wang, B., Wang, X. 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory. In: *ACL (1). pp. 1343–1353.*

41 Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B. 2016. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*.

42 Tai, K.S., Socher, R., Manning, C.D. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

43 Gao, L., Huang, R. 2017. Detecting online hate speech using context aware models. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP.* p. 260-266.

44 Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.2017. Hate me, hate me not: Hate speech detection on facebook. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy*.

45 Metz, C.1978. Basic principles of ROC analysis. Semin Nucl Med. 8 (4): 283–98.

46 Sasaki Y. 2007. The truth of the F-measure.*Teaching, Tutorial materials, Version: 26th October.*

47 Lewis D.D., Schapire R., Callan J.P., Papka R.1996. Training algorithms for linear text classifiers. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 298-306.*

48 Yang Y., Liu X.1999. A Re-Examination of Text Categorization Methods. In *Proceedings 22nd Annual International SIGIR . Berkeley.*

49 Zhang Z., Luo L.2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *vol. 1, no. 0, pp. 1–5, 2018.*

50 Cho K., van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014).*

51 Park J.H., Fung P.2017. One-step and two-step classification for abusive language detection on Twitter. *arXiv preprint aeXiv:1706.012.*

52 Badjatiya P., Gupta S., Gupta M., Varma V.2017. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760.

53 Saleem H., Dillon K.P., Benesch S., Ruths D.2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces.*CoRR abs/1709.10159.*

54 Jones K. S. 2004. A statistical interpretation of term specificity and its application in retrieval. *MCB University: MCB University Press, 2004. — Vol. 60, no. 5. — P. 493-502.*

55 Waseem Z., Hovy D.2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL Student Research Workshop, pp. 88–93. Association for Computational Linguistics.*

56 Fasoli F., Carnaghi A., Paladino M.2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences, 52:98–107.*
57 Hardaker C., McGlashan, M.2015. Real men don’t hate women: Twitter rape threats and group identity. *Journal of Pragmatics, 91, pp.80-93.*

58 Clarke I., Grieve J. 2017. Dimensions of abusive language on twitter. In *Proceedings of the First Workshop on Abusive Language Online, 1–10.*

59 Anzovino M., Fersini E., Rosso P.2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In: *Proc. 23rd Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2018, Springer-Verlag, LNCS(10859), pp. 57-64.*

60 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 (EVALITA'18). Caselli, Tommaso and Novielli, Nicole and Patti, Viviana and Rosso, Paolo CEUR.org, Turin, Italy.*

61 Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel F., Rosso P., Sanguinetti M. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

62 Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

63 Pamungkas E.W., Cignarella A.T., Basile V., Patti V. 2018. 14-ExLab@UniTo for AMI at IberEval 2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. *CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain.*

64 Frenda S., Ghanem B., Montes-y-Gomez M.2018. Exploration of Misogyny in Spanish and English tweets. *CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain.*

65 Canós, J.S. 2018. Misogyny identification through SVM at IberEval 2018. *CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain.*

66 Ahluwalia R., Shcherbinina E., Callow E., Nascimento,A., De Cock M. 2018. Detecting Misogynous Tweets. *CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain.*

67 Bakarov A. 2018. Vector Space Models for Automatic Misogyny Identification. In Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. *Final Workshop (EVALITA 2018), Turin, Italy. CEUR.org.*

68 Ahluwalia R., Soni H., Callow E., Nascimento A., De Cock. M. 2018. Detecting Hate Speech Against Women in English Tweets. In Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. *Final Workshop (EVALITA 2018), Turin, Italy. CEUR.org.*

69 Basile A., Rubagotti C. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. I*n Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018), Turin, Italy. CEUR.org*.

70 Almatarneh S., Gamallo P., Pena F.J.R.2019. CiTIUS-COLE at semeval - 2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets. *In: the 13th international Workshop on Semantic Evaluation.*

71 Perez J.M., Luque F.M.2019. Atalaya at SemEval 2019 Task 5: Robust Embeddings for Tweet Classification.*In: the 13th international Workshop on Semantic Evaluation.*

72 Bauwelinck N., Jacobs J., Hoste V., Lefeve E. 2019. LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval). *In: the 13th international Workshop on Semantic Evaluation.*

73 Ameer I., Siddiqui M.H.F., Sidorov G., Gelbukh A. 2019. CIC at SemEval-2019 Task 5: Simple Yet Very Efficient Approach to Hate Speech Detection, Aggressive Behavior Detection, and Target Classification in Twitter. *In: the 13th international Workshop on Semantic Evaluation*.

74 Ping L. 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. *SemEval@NAACL-HLT*.

75 Seganti A., Sobol H., Orlova I., Kim H., Staniszewski J., Krumholc T., Koziel K.. 2019. NLPR@SRPOL at SemEval-2019 Task 6 and Task 5: Linguistically enhanced deep learning offensive sentence classifier. *arXiv:1904.05152.*

76 Radivchev V., Nikolov A., Nikolov-Radivchev. 2019. SemEval-2019 Task 6: Offensive Tweet Classification with BERT and Ensembles, *SemEval@NAACL-HLT*.

77 Devlin J., Chang M., Lee K., Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

78 Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*.

79 Wang S., Manning C. 2012. Baselines and bigrams: simple, good sentiment and topic classification. *In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. pp. 90–94. ACL.*

80 Pérez F., Cardiff, J., Pinto D., Rosso P. 2016. Prototype/Topic based clustering method for weblogs. *Intelligent Data Analysis, vol. 20, no. 1, pp. 47-65, IOS Press.*

81 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.*

82 Shushkevich, E., Cardiff J.2018. Misogyny Detection and Classification in English Tweets: The Experience of the ITT Team. *EVALITA 2018, Evaluation of NLP and speech tools for Italian.*

83 Shushkevich, E., Cardiff J., Rosso P. 2019. TUVD team at SemEval-2019 Task 6: Offense Target Identification. *SemEval@NAACL-HLT*.