Sexism Detection in English and Spanish Tweets

Ritesh Kumar[1], Soumya Pal[2], and Rajendra Pamula[2]

[1] Department of Computer Science and Engineering
National Institute of Technology Jamshedpur, India
ritesh.cse@nitjsr.ac.in
[2] Department of Computer Science and Engineering
Indian Institute of Technology (Indian School of Mines), Dhanbad, India
{soumyapal0601,rajendrapamula }@gmail.com

Abstract. Sexism is an ancient evil that society struggles with till date. The freedom of speech and ease of anonymity granted by social media has also resulted in incitement to hatred. This presents the need for automatic detection of sexist posts or tweets on social media. In this paper, we present the machine learning models that can detect sexism. Specifically, we describe the model submitted for the shared task on sEXism Identification in Social network at IberLeF 2021. The problem concentrates on sexism detection in two languages: English and Spanish. The challenge is divided into two tasks of different granularity: (1) a binary classification problem to discover the instances of sexism in the post and (2) to predict one the five types of sexism present. Overall, our performance is good but it needs some improvement, our scores are encouraging enough to work for better results in future.

Keywords: SVM, LSTM, Random Forest, Misogyny, Social Media

1 Introduction

The digitalization of the world also means the digitalization of sexism. Sexism can be referred to as individuals' attitudes, beliefs, behaviors, organizational, institutional and cultural practices that either reflect negative evaluations of individuals based on their gender or support unequal status of women and men. With rapid advancement of the internet and social media, such beliefs can spread rapidly. Thus, it is very essential to detect such behaviour. The amount of data generated on social media sites can be estimated from the fact that, every second, on average, around 6,000 tweets are generated. Content moderation of such a huge data is difficult to achieve exclusively through man power. Social networking sites are struggling with content moderation. Artificial Intelligence and different Machine Learning techniques can be exploited for Sexism Detection.

In this paper, we present the models to detect sexism, specifically the system we submitted for EXIST 2021 shared task [3]. Sexism can be expressed directly,

[3] *IberLEF 2021, September 2021, Málaga, Spain.*

implicitly and can be hidden in the form of humor. Automatically detecting the subtle instances of sexism is much harder than discovering the misogyny and violence aimed towards women. We try to identify the instances of both implicit and explicit sexim and future classify the sexism into one of the five categories which are as follow: 1. Ideological and Inequality, Stereotyping and Dominance, Objectification, Sexual Violence, Misogyny and Non-sexual Violence.
As per requirement of EXIST 2021 [5], we submitted three runs for task1 and two runs for task2. We extracted different lexical and non lexical features from the text for the classification. Our best run in task1 achieved an accuracy of 71%. For task2, our best run was with an F1 score of 0.59.

The organization of the rest of the paper is as follows. Section 2 describes about Related Work. In Section 3, we describe about Task & Dataset. Section 4, describes about System Description i.e. Test Preprocessing, Feature Extraction and Machine Learning Models. In Section 5, we discuss about our Results. Lastly, Conclusion and future direction of our research work is presented in section 6.

2 Related Work

Several works have been proposed to detect misogyny and hate speech against women across social platforms. Resham Ahluwalia et al. [1] proposed how ML models can detect misogynist tweets. Pamungkas et al. [6] proposed a cross language study for misogyny detection in tweets. Jose et al. [2] identified cases of aggressiveness and hate speech towards women. Goenaga et al. [3] used RNN approach using a Bidirectional Long Short Term Memory (Bi-LSTM) with Conditional Random Fields (CRF) and evaluated the proposed architecture on misogyny identification task (text classification). Shushkevich et al. [8] combined several simpler classifiers into one more complex blended model, which classified the data taking into account the probabilities of belonging to classes calculated by simpler models. They used the Logistic Regression, Naive Bayes, and SVM classifiers. Liu et al. [4] used three classifiers that were trained by using SVM with a linear kernel, random forests (RF) and gradient boosted trees (GBT).

The dataset we used, misogynist tweets are considered as a sub-category of sexist tweets. As most of these works focus on misogyny and hate speech detection but not the subtle instances of sexism. We try to identify these subtle sexism in tweets.

3 Task and Dataset Description

In this section, we describe the automatic sexism detection shared task and the dataset provided to the participants [7].

EXIST 2021 shared task was divided into two subtasks (Task1 and Task2). Task1 aims at finding the instances of sexism while the goal of Task2 is to categorize the sexism into one of the following 5 categories (shown in Table 1):
1. Ideological and inequality : Discrediting of the feminist movement, rejecting the precence of inequality between men and women, or presenting man as the

victim of gender bias.
2. Stereotyping and dominance : Expressing the false idea about women like they are more suitable for certain tasks and not suitable for certain others.
3. Objectification : Presenting women as objects.
4. Sexual Violence : Sexual suggestions, requesting sexual favours or sexual harrasment threats.
5. Misogyny and Non-sexual Violence : Indicating voilence and hate toward women of non-sexual nature.

Table 1. Categorization of Sexism with Example

Category	Example
Ideological and inequality	I think the whole equality thing is getting out of hand.
Stereotyping and dominance	Most women no longer have the desire.
Objectification	First thing I see are her slut nails.
Sexual Violence	fuck that cunt, I would with my fist.
Misogyny And Non-sexual Violence	Want to settle down now that you've gotten bored with cock carousel.

We used the dataset obtained from the shared task at IberLEF 2021. The dataset contains 11,000 short texts i.e. tweets and gab posts in English and Spanish languages. The dataset consists of 6,977 tweets for training and 3,386 tweets for testing with balanced distribution in both the languages. In addition the test data contains 492 gabs in English and 490 in Spanish from the uncensored social network Gab (gab.com). The data was labeled manually by human annotators. Data was annotated at two different levels of granularity. First, each text was labelled as sexist or non-sexist. Secondly, the instances labelled as sexist are further divided as ideological-inequality, stereotyping-dominance, objectification, sexual-violence, misogyny-non-sexual-violence. The EXIST dataset focuses on covering sexism in a broad sense, i.e, both explicit misogyny to implicit subtle sexism.

4 System Description

We splitted the annotated data into a training set with 85% of the annotated data and validation set with the remaining 15% of the instances. We built similar systems for both English and Spanish, and then results obtained by these two systems were combined and submitted as a run. Same splitting was used for both task1 and task2.

4.1 Test Preprocessing

We removed punctuation symbols, links and numbers. Also, we removed stop words. We used lemmatization for the English text and stemming for the spanish text to ping together the different inflected forms of a word. In particular, we used nltk wordnet for lemmatization and nltk snowball for stemming. The same preprocessing is used for both Task1 and Task2.

4.2 Feature Extraction

We extracted the following features from the text:

- Upper Case Count : Number of uppercase letters present in the text
- Hashtag Count : Number of hashtags present in the text
- Link Count : Number of hyperlinks in present in the text
- Tweet Length : Total number of characters in the text
- Slang Words Count : used in the English language

These features were extracted before preprocessing as hashtags and links were removed after preprocessing. We used the Slang word count feature only for the English text for both Task1 and Task2. All the rest features were used for both languages and both tasks. To convert the text into numerical features we used Tf-idf vectorizer. For LSTM, we used the tokenizer by keras library and for SVM and Random Forest we used Tf-idf vectorizer from scikit-learn library. Tokenizer by keras converts the text into either a sequence of integers or into a vector where the coefficient for each token could be binary, based on word count, based on tf-idf etc.

4.3 Machine Learning Models

For Task1, we submitted three runs based on three different algorithms, namely- SVM, Random forest and LSTM. We used the scikit-learn library for SVM and Random forest based models and Keras for LSTM. For SVM, we experimented with three kernels - linear, rbf (radial basis function) and polynomial. The final run we submitted was the linear kernel SVM. We used the following values of the parameter in scikit-learn:
1. For SVM, kernel coefficient gamma is 0.01, and penalty parameter is 5 (default parameters).
2. For Random Forest, maximum depth is 150, and number of forest is 1000.
3. For LSTM, we used mean squared error for loss,and batch size as 64.

For Task2, we submitted two runs based on -SVM and Random Forest. We treated Task2 as a multi-classification problem with 6 categories (five categories of sexism and one category as non-sexist) and not as an extension of Task1. That is, in case of Task2, we used the whole dataset for training and not just the text classified as sexist in Task1. The parameter values were the same as mentioned above. And SVM was implemented with a linear kernel.

5 Results and Discussion

The results of Task1 are represented in terms of accuracy (shown in Table 2), while the results of Task2 are in terms of macro F1 score (shown in Table 3). The best score as accuracy, we get from Task1 is 0.7115. For Task2 we get best score as F1- measure is 0.4504. Table 2 and 3 shows the ranking of our submissions based on shared task official ranking. Our best system was ranked 38 in Task1

and 43 for task2. Also, *baseline* is provided by the organizers for both the task. These rankings indicate the combined results in English and Spanish.

For Task1, the Random forest system performed better than LSTM and SVM. But for Task2, SVM performed better than Random forest. The accuracy and F1 score obtained in Task2 was lower than that of Task1. The same trend could be found in results of all the teams. This can be attributed to the fact that classification of the sexist text into finer granularity is a much more difficult task than detecting instances of sexism. Also, as Task2 was multi-class classification problem it's F1 score was lower as compared to the binary classification problem in Task1 .Features like number of slang words, number of hashtags present, hyperlink count resulted in improved accuracy.

Table 2. Results for Task1- The official Evaluation measure is Accuracy. The best score obtained by us is mentioned in bold

Rank	Run	**Accuracy**	F1
38	**task1_Soumya_2**	**0.7115**	**0.7114**
45	task1_Soumya_1	0.7047	0.7045
55	task1_Soumya_3	0.6761	0.6761
52	task1_Baseline_svm_tfidf	0.6845	0.6832

Table 3. Results for Task2- The official Evaluation measure is F1. The best score obtained by us is mentioned in bold

Rank	Run	Accuracy	**F1**
43	**task2_Soumya_1**	**0.5923**	**0.4504**
46	task2_Soumya_2	0.595	0.4415
51	task2_Baseline_svm_tfidf	0.5222	0.395

6 Conclusion and Future Work

We evaluated the performance of different classification algorithms for sexism this year's shared task. The results, we achieved were average as compared to other submissions obtained in the EXIST 2021 shared task. We found that extracted features like number of slangs and hyperlink counts can result in better performance. We look forward to experimenting with different features. Also, fine tuning the parameters of the algorithm can help in improvement of the overall performance. And the results of more than one system can be combined to generate an overall better score. We shall be exploring these tasks in the coming days.

Bibliography

[1] Ahluwalia R, Shcherbinina E, Callow E, Nascimento AC, De Cock M (2018) Detecting misogynous tweets. In: IberEval@ SEPLN, pp 242–248

[2] Canós JS (2018) Misogyny identification through svm at ibereval 2018. In: IberEval@ SEPLN, pp 229–233

[3] Goenaga I, Atutxa A, Gojenola K, Casillas A, de Ilarraza AD, Ezeiza N, Oronoz M, Pérez A, Perez-de Viñaspre O (2018) Automatic misogyny identification using neural networks. In: IberEval@ SEPLN, pp 249–254

[4] Liu H, Chiroma F, Cocea M (2018) Identification and classification of misogynous tweets using multi-classifier fusion. In: Evaluation of Human Language Technologies for Iberian Languages: IberEval 2018, CEUR Workshop Proceedings, pp 268–273

[5] proceedings of IberLEF 2021 can be tentatively referred to as: Manuel Montes, Rosso P, Gonzalo J, Aragon E, Agerri R, Angel M, Carmona A, Mellado EA, de Albornoz JC, Chiruzzo L, Freitas L, Adorno HG, Gutierrez Y, Zafra SMJ, Lima S, de Arco FMP, (eds) MT (2021) Proceedings of the iberian languages evaluation forum (iberlef 2021). CEUR Workshop Proceedings

[6] Pamungkas EW, Cignarella AT, Basile V, Patti V, et al (2018) 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In: 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018, CEUR-WS, vol 2150, pp 234–241

[7] Rodriguez-Sanchez F, Carrillo-de Albornoz J, Plaza L, Gonzalo J, Rosso P, Comet M, Donoso T (2021) Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural 67(0)

[8] Shushkevich E, Cardiff J (2018) Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018. In: IberEval@ SEPLN, pp 255–259