NIT-Agartala-NLP-Team at EXIST 2022: Sexism Identification in Social Networks

Amaan Rizvi[1,*,†], Anupam Jamatia[2,†]

[1]National Institute of Technology, Agartala, Tripura, India, 799046

Abstract

The paper describes the results submitted by the 'NITAgartala-NLP-Team' at Exist 2022. A dataset of 6,977 tweets for training and 3,386 tweets for testing was provided by the task organizers to train and test our models. Our models include a Logistic Regression, Support Vector Machine, and Multinomial Naïve Bayes for text classification. For Task-1 and Task-2 we attained the rank of 38/47 and 28/31 respectively. We discuss our approach to handling raw text and issues we encountered during text preprocessing and try to give our solutions.

Keywords

Text Classification, Logistic Regression, Support Vector Machine, Multinomial Naïve Bayes

1. Introduction

With the rapid development of internet technology and mobile communication technology, social media has become one of the largest source of data. But internet is no more an equal space for all. The past few years have seen a rise in concerns about the disproportionate levels of abuse experienced by women in social media platforms[1]. Hate speech on Twitter aimed at female politicians, journalists, and those engaging in feminist debate, has been documented across different countries. Recently, Amnesty International published a report where they describe Twitter as a 'toxic place' for women. According to this report, Twitter is promoting violence and hate against or threaten people based on their gender. Online gender-based violence can have significant psychological, social, and economic impacts. Most directly, it affects women's freedom of expression. One study showed that women who experience online abuse often adapt their online behaviour, self-censor the content they post and limit interactions on the platform out of fear of violence and abuse. By silencing or pushing women out of online spaces, online violence can affect the economic outcomes of those who depend on these platforms for their livelihoods. It can also lead to loss of employment and societal status, in cases where online violence impacts their reputation (for e.g. in cases involving revenge porn or non-consensual pornography)[2]. In addition to directly impacting the women who are present online, online gender-based violence could be predictive of violent crimes in the physical world. A study in

IberLEF 2022, September 2022, A Coruña, Spain.

*Corresponding author.

amaan.rizvi39@gmail.com (A. Rizvi); anupamjamatia@gmail.com (A. Jamatia)

https://github.com/amaanrzv39 (A. Rizvi); https://anupamjamatia.github.io (A. Jamatia)

0000-0001-6244-8626 (A. Jamatia)

the United States found that cities with a higher incidence of a certain kind of racist tweets reported more actual hate crimes related to race, ethnicity, and national origin[3].

Who are the victims?

Needless to say, similar to offline gender-based violence, marginalized women are more likely to experience online violence. It is important to note the inter-sectionality of online violence while developing regulations or interventions to address these issues. African-American women are likely more at risk when compared to their White counterparts. A Muslim woman might get more hate, owing to the discrimination she faces based on her gender as well as her religion. In terms of age, young girls are observed to be at much higher risk, perhaps owing to the greater engagement online. Studies have indicated that intimate partner violence can also lead to online violence – men posting indecent photos of, or hateful content against their partners, as revenge. Identification of such tweets has become a key area in field of Natural Language Processing and text classification. Our aim in Task-1 is simply the detection of sexist words in a tweet (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour). Task-2 involves, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours. In this paper, we aim to understand how sexist behaviours, beliefs and attitudes are expressed in Twitter conversations. We focus on tweets written in English and Spanish[4]. We propose to identify sexism in social networks using an automatic system based on machine learning. This is a new age technology based initiative towards the goal of creating novel mechanisms to detect and alert from abusive and sexist behaviours against women in social media.

2. Related Work

Substantial work has been devoted to the detection of hate speech in recent years, including tasks such as racist or xenophobic content detection, but few works have faced sexism detection and, in particular, they have dealt with sexism as the detection of hate speech against women. Consequently, they have worked with hostile and explicit sexism, overlooking subtle or implicit expressions of sexism. However, some ideas and techniques from hate speech detection may apply to our problem. Therefore, in this section, we briefly review related work in the hate speech field along with previous works on sexism and misogyny detection.

'Misogyny' and 'sexism' are frequently considered interchangeable, though both terms have different nuances. However, the most widely accepted definition of misogyny implies the expression of hostility and hatred towards women. In contrast, sexism comprises any form of oppression or prejudice against women and therefore may be hostile (as in the case of misogyny) or subtle. Thus, sexism includes misogyny but is not limited to it[5]. Current studies on the identification of sexism are related to hate speech detection. Introducing sexism as the classification task was first proposed by Waseem, 2016[6]. He annotated 16 thousand tweets and categorized them as racist, sexist and neither. Waseem collected tweets around the famous Australian TV show such as 'My Kitchen Rules' using the hashtag #mkr. He tried different methods such as character level grams and word grams and employed logistic regression with 10-fold cross-validation. Sharifirad and Jacovi presented a categorization of sexism that included indirect, sexual, and physical sexism[7]. A more recent study by seeks to categorize accounts

of sexism. Because the growing interest of hate detection towards women, other tasks to protect women from hate on the internet have emerged. For instance, sexist meme detection and classification of sexist advertisements. We can find in the literature previous works that have specifically faced the automatic detection of misogyny in text as well as some datasets annotated with misogynist expressions. ElSherief *et al.* compiled Hate Lingo, an English dataset that comprises hate speech tweets that include hatred expressions towards people based on some intrinsic characteristics of the person, including their gender, class, ethnicity or religion. Similarly, Ousidhoum *et al.* creates a multi-lingual corpus that included expressions of hate towards women in English, French and Arabic. In this work, we discuss the machine learning models like Logistic Regression (LR), Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) as they gives better results on Exist 2021 dataset among other machine learning models.

3. Corpus

The task organizers IberLEF 2022 have provided a dataset (*EXIST 2021*). The EXIST 2021 dataset consists of 6977 tweets for training and 3386 tweets for testing. There were two tasks given by organizers:

1. **Task-1: Sexism Identification** - The first subtask is a binary classification problem. The model has to classify whether or not a given tweet contains sexist expression. The following tweets show examples of sexist and not sexist messages.
    - **SEXIST:**
        - *"Mujer al volante, tenga cuidado!"*
        - *"People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don't need a fat ass to get a man. Never have."*
    - **NOT SEXIST:**
        - *"Mujer al volante, tenga cuidado!"*
        - *"People really try to convince women with little to no ass that they should go out and buy a body. Like bih, I don't need a fat ass to get a man. Never have."*
2. **Task-2: Sexism Categorization** - The second subtask is a multiclass classification problem. The model has to categorize the given sexist tweet according to the type of sexism. In particular, we propose a five-classification task:
    - **IDEOLOGICAL AND INEQUALITY:** The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
        - *"Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron."*
        - *"I think the whole equality thing is getting out of hand. We are different, thats how were made!"*

- **STEREOTYPING AND DOMINANCE:** The text stereotypes women's role in society, suggesting they are more suitable to fulfill inhouse roles like mother, wife, family caregiver, faithful, tender, loving, submissive, etc., and unsuitable for manly tasks like driving, leadership, etc. It intends to establish that men are somehow superior to women.
  - *"@Paula2R @faberacuria A las mujeres hay que amarlas...solo eso... Nunca las entenderás."*
  - *"Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to."*
- **OBJECTIFICATION:** The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles like compliance with beauty standards, sexualization of female attributes, women's bodies at the disposal of men, etc.
  - *"Pareces una puta con ese pantalón" - Mi hermano de 13 cuando me vio con un pantalón de cuero."*
  - *"Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring."*
- **SEXUAL VIOLENCE:** The text reports harassment of a sexual nature like rape or sexual assault.
  - *"#MeToo Estas 4 no han conseguido su objetivo.El juez estima que se abrieron de patas https://t.co/GSHiiwqY6Aánta lagartona hay en este."*
  - *"fuck that cunt, I would with my fist."*
- **MISOGYNY AND NON-SEXUAL VIOLENCE:** The text expressses hatred and violence towards women.
  - *"Las mujeres de hoy en dia te enseñar a querer… estar soltero."*
  - *"Some woman are so toxic they don't even know they are draining everyone around them in poison. If you lack self awareness you won't even notice how toxic you really are."*

Sample distribution of English and Spanish tweets are given in following tables:

Table 1
English tweets samples distribution for Task 1

Category	No. of samples	Percentage(%)
Sexist	1636	47.61
Non-Sexist	1800	52.39

Table 2
English tweets samples distribution for Task 2

Category	No. of samples	Percentage(%)
Non-Sexist	1800	52.39
Ideological Inequality	386	11.23
Stereotyping Dominance	366	10.65
Sexual Violence	344	10.01
Misogyny Non Sexual Violence	284	8.26
Objectification	256	7.45

Table 3
Spanish tweets samples distribution for Task 1

Category	No. of samples	Percentage(%)
Sexist	1741	49.17
Non-Sexist	1800	50.83

Table 4
Spanish tweets samples distribution for Task 2

Category	No. of samples	Percentage(%)
Non-Sexist	1800	50.83
Ideological Inequality	480	13.55
Stereotyping Dominance	443	12.51
Sexual Violence	401	11.32
Misogyny Non Sexual Violence	244	6.89
Objectification	173	4.88

We observed that Sexist and Non-Sexist tweets are evenly distributed in the data set, but in case of multiple classes for Task 2 there is significant imbalance in data set. For e.g. category *'Objectification'* has very less samples for both English and Spanish tweets, *256* and *173* respectively. The retributions and possible solutions for target class imbalance problem will be discussed in *Error Analysis and Discussions* section.

4. Preprocessing and System Overview

We begin by segregating training and test dataset into English tweets and Spanish tweets. Since both languages are different and their words might have different meaning for different language, so separating different sources keep our machine learning models simple and clean. We train machine learning models on newly created datasets for both the languages. Here is a glance at distribution of tweets based on language:

Table 5
Distribution of English and Spanish tweets

Dataset	English	Spanish
Training	3436	3541
Test	2208	2160

Our first step was to clean the text data inorder to get a better vector representation of text data, we applied following text preprocessing techniques to clean text:

1. Removing web addresses from text e.g. *"Incredible! Beautiful! But I laughed so much when I read about you drifting in your wheelchair.I can just picture it https://twitter.com/i/status/1335010901649395714"*
2. Removing emoticons from text.
3. Removing unrecognized characters, emojis and stickers from text
4. Removing special characters.
5. Removing repeating patterns like 99, aaaaa, bbbbb, 00 etc.
6. Removing one character length word like l, 9, 1, B etc.
7. Fixing contractions, e.g converting words like I'll to I will
8. Stemming words using Snowball stemmer[8]

We decided to keep stop words[9] for both English and Spanish tweets, as our experiment yield slightly better F1 score by keeping them.

5. Experimental Setup and Results

To apply any machine learning model to text data we need to convert text to a vector representation, the mostly used techniques for converting text to vector are *Bag-of-words*, *TF-IDF*, *Word2Vec*, *TF-IDF Weighted-Word2Vec*:

- **Bag-of-words**: Bag-of-words (BoW)[10] approach simplifies bodies of text by considering them as unordered collections of words. Clearly, this has the disadvantage of ignoring sentence structure and semantic relationships between sentence elements (as if shuffled inside of a "bag" of words). Nonetheless, despite its strong assumptions, it has been shown to obtain good results and has seen wide use. We ran Logistic regression, SVM and Naive Bayes algorithm on this representation, results were fair but they were superseded by the results of TF-IDF representation, which we discuss next.

- **TF-IDF**: In information retrieval, TF-IDF(term frequency–inverse document frequency)[11] is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is calculated as follows:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t_i \in d} f_{t_i,d}}$$

where $f_{t,d}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document d.

$$idf(t,D) = log\frac{N}{1 + |d \in D : t \in d|}$$

where N is total number of documents in the corpus and $|d \in D : t \in d|$ is number of documents where the term t appears. Then tf–idf is given by

$$tfidf(t,d,D) = tf(t,d) * idf(t,D)$$

We built TF-IDF (Term Frequency – Inverse Document Frequency) features using unigram and bi-gram. Our machine learning model gave better result for this representation.

- **Word2Vec**: Word2Vec[12] is a method to construct a word embedding. A word embedding is a learned representation for text where words that have the same meaning have a similar representation. It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems. Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network, and hence the technique is often lumped into the field of deep learning[13]. There are two proposed architecture for Word2Vec:
  - **Continuous Bag-of-words (CBOW)**: predicts the middle word based on surrounding context words. The context consists of a few words before and after the current (middle) word. This architecture is called a continuous bag-of-words model as the order of words in the context is not important.
  - **Skip-Gram**:predicts words within a certain range before and after the current word in the same sentence.
- **TF-IDF weighted Word2Vec**: In TF-IDF weighted Word2Vec [13] representation we obtain vectors by multiplying TF-IDF value of words with its corresponding Word2Vec value.

For our experiment we used TF-IDF based feature representation and trained Logistic Regression, Naïve Bayes and SVM models on it. We used precion, recall and F1-score[14] as main metric for our error report which are summarized in tables below:

Table 6
Metric for English tweets on Task 1

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7196	0.72	0.72	0.72
Naive Bayes	0.6757	0.69	0.68	0.67
SVM	0.7052	0.71	0.70	0.70

Table 7
Metric for Spanish tweets on Task 1

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7287	0.73	0.73	0.73
Naive Bayes	0.7092	0.72	0.71	0.71
SVM	0.7226	0.73	0.72	0.72

Table 8
Metric for English tweets on Task 2

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.5756	0.49	0.46	0.47
Naive Bayes	0.48	0.54	0.17	0.12
SVM	0.5855	0.54	0.41	0.45

Table 9
Metric for Spanish tweets on Task 2

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.6208	0.61	0.46	0.49
Naive Bayes	0.4865	0.53	0.18	0.13
SVM	0.6101	0.66	0.41	0.46

6. Error Analysis and Discussions

Figure 1: Confusion Matrix of Task 1 results

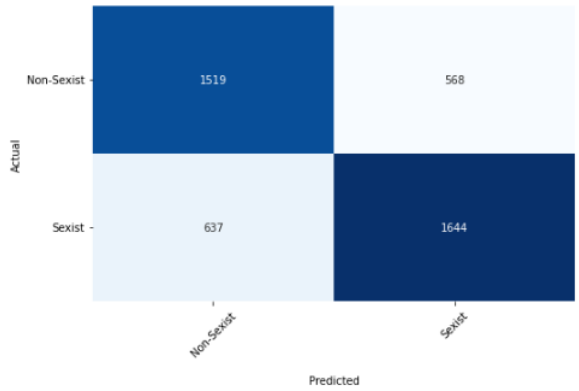
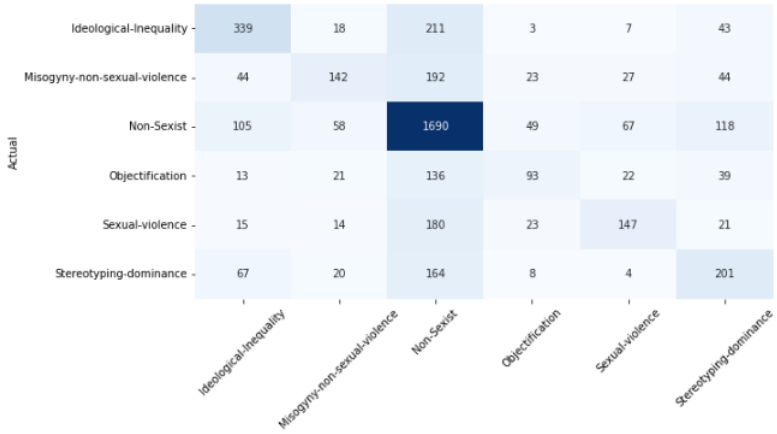


Figure 2: Confusion Matrix of Task 2 results



We can observe from above confusion matrices that our model performs well for Task1, but there are significant errors in our model for Task2. We manually check few of the examples that our model could not correctly identify to find the root cause of those errors, we found some interesting pattern for those errors which are discussed below:

- Some tweets actually had offensive words but they were labelled as 'non-sexist'. e.g. *"If I get more fans, my boyfriend will join me! You'll get videos/photos of my pussy/mouth filled with cum, my tits/ass/face covered with cum, and shots of him fucking my face/pussy https://t.co/iWemENmL8T https://t.co/olTLzy5nAj"*, this should be a sexist tweet but it was labelled non-sexist in the data. Few more examples of tweets labelled as non-sexist but should have been otherwise, *"some tits in the dms https://t.co/RLYBT6PCrr"*, *"@thickiepops_art BE A WHORE!!"*. Correct annotation of these tweets during data preparation will help get better accuracy for machine learning and deep learning models.

- Sexism identification task required semantic and cognitive understanding of text, like this tweet *"i'm straight but i'm also allergic to women lol haha"* is a misogynist comment but as we can observe it does not have any offensive words in it. This type of relationship is almost impossible to capture in tf-idf vector representation and in fact even hard for most word embedding neural network models.
- It is not just the text that describes the kind of tweets but also the url's in the tweet. For e.g. *"https://t.co/Of8axF4XXj.NOT.ACCEPTABLE - what if it was your family member it was happening to? #think #harassed #ukrunchat #runners #femalerunner #harassedrunners #unacceptable https://t.co/uPjtSm0s8J"*, this tweet contains video that describes it as case of sexual violence. But as we remove url's in our data preprocessing, we can not infer the class of this tweet with only text data. It might be an idea to create multi-model task that captures images, videos and text information but for this task we sticking to only text data. So, this type of tweets turn to be a outliers for our algorithm. One possible solution could be to eliminate those tweets that contain only url's and very few words (*e.g. number of words* < 4) while training our model. But the problem will still persist with these kind of tweets while evaluating test data.
- And also presence of abusive GIF images and emoticons in tweets are hard to capture using normal natural processing techniques. For emoticons, we can create a regex pattern for abusive emoticons and keep them as feature, but GIF images like videos mentioned above still remain hurdle.

That was our summary of the type of errors and their root cause analysis. We also tried to provide solution for each of these problems, which can be vital in improving model performance.

7. Conclusion

In this paper, we provide an overview of models for text classification, we also explored the application of machine learning to understand how sexist attitudes and behaviours are expressed in social networks conversations. Our aim in this task was to detect sexism in a broad sense in Twitter. For this, we presented three machine learning models namely logistic regression, support vector machine and naive bayes. Regarding feature extraction, we used two traditional methods bag of words and tf-idf. The experimental results showed that tf-idf representation gives better result than bag of words and also logistic regression model outperforms other two giving 72% accuracy for Task 1 and 60%accuracy for Task 2.
We also tried to present vector representation for Spanish corpus and used snowbell stemmer for Spanish text, this as per our knowledge is one of the few experiments which applies natural language processing techniques for Spanish text.
The points that we did not discuss in this paper is word embedding for Spanish text[15], this is something we would try in our future experiments. While coming to an end, we would like to suggest the task organizers to collect more data such that every category for both English and Spanish text has proportionate amount of distribution, as low training examples of some categories can lead machine learning model to fit poorly for those category instances. We look forward to the new research in this topic and future task that the organizers provide.

References

[1] M. F. Wright, B. D. Harper, S. Wachs, The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition, Personality and individual differences 140 (2019) 41–45.

[2] C. R. Carlson, H. Witt, Online harassment of us women journalists and its impact on press freedom, First Monday (2020).

[3] C. Hardaker, M. McGlashan, "real men don't hate women": Twitter rape threats and group identity, Journal of Pragmatics 91 (2016) 80–93.

[4] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022).

[5] S. Frenda, B. Ghanem, M. Montes-y Gómez, P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, Journal of Intelligent & Fuzzy Systems 36 (2019) 4743–4752.

[6] Z. Waseem, Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138–142.

[7] S. Sharifirad, A. Jacovi, Learning and understanding different categories of sexism using convolutional neural network's filters, in: Proceedings of the 2019 Workshop on Widening NLP, 2019, pp. 21–23.

[8] M. F. Porter, Snowball: A language for stemming algorithms, 2001.

[9] C. Fox, A stop list for general text, in: Acm sigir forum, volume 24, ACM New York, NY, USA, 1989, pp. 19–21.

[10] Y. Zhang, R. Jin, Z.-H. Zhou, Understanding bag-of-words model: a statistical framework, International Journal of Machine Learning and Cybernetics 1 (2010) 43–52.

[11] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, volume 242, Citeseer, 2003, pp. 29–48.

[12] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[13] J. Lilleberg, Y. Zhu, Y. Zhang, Support vector machines and word2vec for text classification with semantic features, in: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2015, pp. 136–140.

[14] C. J. Van Rijsbergen, Information retrieval. 2nd. newton, ma, 1979.

[15] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2021) 506–518.