CSECU-DSG at SemEval-2023 Task 10: Exploiting Transformers with Stacked LSTM for the Explainable Detection of Online Sexism

Afrin Sultana, Radiathun Tasnia, Nabila Ayman and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
afrin.sultana.cu@gmail.com, radia.tasnia.cu@gmail.com,
nabila.ayman.cu@gmail.com, and nowshed@cu.ac.bd

Abstract

Sexism is a harmful phenomenon that provokes gender inequalities and social imbalances. The expanding application of sexist content on social media platforms creates an unwelcoming and discomforting environment for many users. The implication of sexism is a multi-faceted subject as it can be integrated with other categories of discrimination. Binary classification tools are frequently employed to identify sexist content, but most of them provide extensive, generic categories with no further insights. SemEval-2023 introduced the Explainable Detection of Online Sexism (EDOS) task that emphasizes detecting and explaining the category of sexist content. The content of this paper details our involvement in this task where we present a neural network architecture employing document embeddings from a fine-tuned transformer-based model into stacked long short-term memory (LSTM) and a fully connected linear (FCL) layer . Our proposed methodology obtained an F1 score of 0.8218 (ranked 51st) in Task A. It achieved an F1 score of 0.5986 (ranked 40th) and 0.4419 (ranked 28th) in Tasks B and C, respectively.

Keywords: Sexism . Transformer document embeddings . Stacked LSTM . BERTweet . Fully connected layer

Content Warning: *This manuscript discusses examples of hateful and sexist content. The authors do not support the use of hateful and sexist language, nor any of the hateful and sexist representations quoted.*

1 Introduction

The ongoing progress and expansion of technology and social platforms are driving the growth of user-generated content. Online platforms (e.g., Facebook, Twitter, blogs, and forums) reflect a wide range of perspectives, facts, and experiences from a diverse set of people. In addition, social media, which offers an anonymous atmosphere, also enables users to display hostile sentiments that can result in harassment of individuals with various genders, nations, races, cultures, and physical features (Fox et al., 2015; Litchfield et al., 2018). Online sexism is a rising issue nowadays. Sexism is a prevalent form of hate speech that is presently seen as a negative trend on social media. Sexism involves the unjust and discriminatory treatment of women or holding negative attitudes towards them solely because of their gender or because of their gender in combination with other aspects of their identity, such as their race, religion. (Badjatiya et al., 2017; Sharifirad and Matwin, 2019; Kirk et al., 2023).

Automatic approaches (Badjatiya et al., 2017; Davidson et al., 2017; Del Vigna12 et al., 2017; Fersini et al., 2018; Basile et al., 2019) are increasingly used to find and assess sexist content at scale, however, most of them just classify information into general, high-level categories without offering any further explanation. The ability to identify sexist content and to explain why it is sexist enhances the interpretability, trust, and comprehension of the choices made by automated systems, giving users and moderators more control. To address this scope, a shared task was introduced at SemEval-2023 by (Kirk et al., 2023) to facilitate the evolution of more precise and explainable English-language models for detecting sexism, including fine-grained classifications for sexist content. The task is hierarchically structured into three sub-tasks.

- Task A (binary sexism detection) involves predicting whether a given post is sexist or not sexist.
- Task B (category of sexism) requires the system to categorize sexist posts into one of four classes, which are threats, derogation, animosity, and prejudiced discussions, through a four-class classification.

The first three authors have equal contributions.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2091–2099

Text	Label		
	Task A	Task B	Task C
E#1: Women use the truth as currency, meaning its only given if there is some advantage to the woman.	sexist	derogation	descriptive attacks
E#2: Holding boys back is part of the female supremacy plan.	sexist	prejudiced discussions	supporting systemic discrimination against women as a group
E#3: go out to enjoy yourself. not to meet women.	not sexist	none	none

Table 1: Examples of sexist and not sexist texts with hierarchical category.

- Task C (fine-grained vector of sexism) involves classifying sexist posts into one of the 11 fine-grained vectors which are briefly outlined in subsection 4.1.

Table 1 represents some examples of sexist and not sexist texts with their hierarchical category from the SemEval-2023 EDOS task dataset (Kirk et al., 2023). Here, the first example (E#1) conveys that women are deceitful and manipulative as they use the truth only as a means of gaining advantage or power. This generalization is not only false but also disrespectful to the diversity of personalities among women. That is why, E#1 is a sexist statement belittling the credibility and integrity of women as a group. Therefore, the derogatory assertion implies descriptive attacks on women. The second instance (E#2) is a sexist proclamation characterizing girls and women as inherently superior to men by intentionally holding boys back which is based on unfounded assumptions and stereotypes. Also, the statement contributes to discrimination promoting an unfair portrayal of women. The third (E#3) instance encourages one to prioritize having fun and enjoying their leisure time instead of focusing on women referring to the example as not sexist.

In this study, we expound our approach to address the challenges posed by the EDOS task. The key contribution of this paper is our proposal to utilize the BERTweet (Nguyen et al., 2020a) model with the accumulation of features from stacked LSTM (Hochreiter and Schmidhuber, 1997) and a fully connected layer (FCL). This approach allows us to leverage the strengths of each component: BERTweet furnishes semantically meaningful text embeddings, LSTM captures long-term dependencies, and the FCL selects and encapsulates the most effective contextual features.

The subsequent sections of this manuscript are arranged in the following manner: Section 2 provides an overview of related work, whereas Section 3 elucidates our proposed framework. In Section 4, we explicate our experimental settings and present a performance analysis of our model. Finally, we draw our paper to a close in Section 5.

2 Related Work

Sexism is defined as any violence or hate speech directed against women that are motivated by their gender alone, or by their gender in combination with one or more other identifying characteristics (Swim et al., 2004). The last ten years have witnessed a surge in the amount of sexist and derogatory content directed toward women on social media. Disclosure of sexist speech has negative effects on the lives of women and restricts their right to give free opinions. Prior research has concentrated on identifying bias against or violence against women (Samory et al., 2021). The majority of relevant NLP research is particularly concerned with detecting abusive language. Even though overt sexism appears to be simple to identify, it is hidden nuances and varied presentations are not. In recent days, automated sexism detection techniques have been the subject of research (Kirk et al., 2023).

Scholars have focused on confrontational and overt sexism, ignoring subtle or implicit sexism. Several datasets were created to research sexism, which includes the display of hatred or hostility toward women (Sharifirad et al., 2019). Other datasets concentrated on the many categories of sexism (Parikh et al., 2019) as well as other media that depict the various types of sexism (Fersini et al., 2019). The automatic misogyny detection of the IberEval competition in the Twitter section focused on the automatic recognition of sexism (Fersini et al., 2018). The dataset for SemEval-2019 Task 5 was likewise constructed using the data from this contest (Basile et al., 2019). The many strategies

that were put out for the competition were primarily focused on supervised machine learning on various textual features, such as n-grams (Frenda et al., 2018), sentiment and syntactic data, and user-based contents, such as the number of repeated tweets, etc. On the other hand, numerous lexical resources (Pamungkas et al., 2018), including swear words and sexist insults, are used, along with deep learning techniques (Rodríguez-Sánchez et al., 2020) like recurrent neural networks and word embedding features, to obtain the final labels.

The detection of hate speech is connected to recent research on the labeling of sexism. For identifying hate speech, there are numerous methods available (Waseem, 2016). Methods encompassed bag-of-words techniques, machine learning-based classifiers (Davidson et al., 2017), standard machine learning-based approaches including support vector machines, logistic regression, and decision trees (Del Vigna12 et al., 2017), as well as sentiment and lexical-based features. Over the years, deliberation has been drawn to the use of transfer learning-based techniques and neural models for the detection of hate speech. Typically, these models used deep learning strategies like CNN and LSTM networks (Badjatiya et al., 2017), which performed remarkably well on the diverse tasks of natural language processing.

Briefly said, we observed that the majority of researchers examined the traditional strategies in their suggested methodologies, which included a collection of different preprocessing techniques, manually created features, and statistical classifiers where transformer-based models including RoBERTa, BERT, ALBERT, etc. play a vital role as models built on transformers are effective at understanding sentence context. However, it capitulates some specific context learning whereas in our proposed approach we amalgamated transformer document embeddings with stacked LSTM and an FCL layer. Finally, this merged form is exploited to an FCL output layer to obtain the final prevision labels. Transformer document embeddings with staked LSTM extract diverse context of a single sentence that helps us to procure a competitive performance in all the subtasks of SemEval-2023 Task 10.

3 Proposed Method

Our proposed framework is focused on distinguishing interpretable instances of sexism from informal user-generated text. Figure 1 illustrates a graphical depiction of our proposed framework. We extract document embedding features for a given text incorporating the transformer model BERTweet. We leverage the FLAIR (Akbik et al., 2019) to generate the document embeddings. Extracted document embeddings vector is passed through two distinct segments. The first one is stacked LSTM and the other is an FCL. Later, the features from the stacked LSTM and the FCL are accumulated through concatenation to generate the final text representation. The final accumulated features combination is fed to a fully-connected output layer to obtain predicted labels.

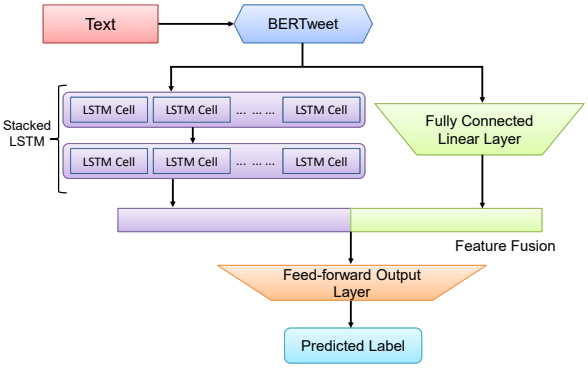

Figure 1: Proposed framework.

3.1 Transformer Document Embeddings

Transformer document embeddings (Akbik et al., 2019) offer an embed for the complete text where we can obtain embeddings directly from a pre-trained transformer for an entire sentence. We can amalgamate various state-of-the-art transformers in the transformer document embeddings model. The main advantage of exploiting a transformer document is to capture the whole sentence context instead of a single word. Moreover, it extracts the semantic similarity between the documents. We utilize a BERTweet embeddings model in the FLAIR architecture to build document embeddings for the particular provided text. To extract the whole sentence context more effectually, we exploit the fragmented transformer layers of this transformer embeddings model.

3.1.1 BERTweet Embeddings

The massive-scale pre-adapted language model BERTweet was learned using English Tweets (Nguyen et al., 2020a). The RoBERTa pre-training technique is used to train it, and it has the same architecture as BERT-base. The

Category	Train	Dev	Test
Task A: binary sexism detection			
Sexist	3398	486	970
Not sexist	10602	1514	3030
Total	14000	2000	4000
Task B: category of sexism			
animosity	1165	167	333
threats, plans to harm and incitement	310	44	89
prejudiced discussions	333	48	94
derogation	1590	227	454
Total	3398	486	970
Task C: fine-grained vector of sexism			
descriptive attacks	717	102	205
threats of harm	56	8	16
incitement and encouragement of harm	254	36	73
aggressive and emotive attacks	673	96	192
casual use of gendered slurs, profanities, and insults	637	91	182
dehumanising attacks & overt sexual objectification	200	29	57
immutable gender differences and gender stereotypes	417	60	119
backhanded gendered compliments	64	9	18
supporting mistreatment of individual women	75	11	21
supporting systemic discrimination against women as a group	258	37	73
condescending explanations or unwelcome advice	47	7	14
Total	3398	486	970

Table 2: The statistics of SemEval-2023 task 10 dataset.

results of experiments demonstrate that BERTweet outperforms robust baselines of RoBERTa-base and XLM-R-base, achieving performance results that are superior to those of the prior state-of-the-art methodologies. The contextual aspects for the entire text are extracted using "vinai/bertweet-base (Nguyen et al., 2020b)." BERTweet embeddings consume both semantic and syntactic contexts for a single sentence. Besides, it can handle various informal language including slang, abbreviations, and misspellings to produce more accurate embeddings for tweets.

3.2 Stacked LSTM

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a type of recurrent neural network (RNN) that can process sequential data containing long-term dependencies. The main benefit of LSTM is that it can selectively retain and forget information over time. Multiple layers of LSTM are combined to form stacked LSTM. In stacked LSTM architecture, the subsequent LSTM layer receives the output of the previous layer as an input, and each layer can have a different number of LSTM cells (Wang et al., 2018). The equation of the stacked LSTM with two layers can be defined as follow:

$$h_1 = LSTM_1(x_1, h_0)$$
$$h_2 = LSTM_2(h_1, h_0)$$

Here x_1 is the input sequence, h_0 is the initial hidden state, h_1 is the output of the top layer, which is fed into the subsequent layer, h_2. Stacked LSTM enhances the capacity of the network (Staudemeyer and Morris, 2019). This allows the network to learn more intricate hidden patterns and deep-level features from the input data. These characteristics of stacked LSTM improved our model to capture the differences between types of sexist content.

3.3 Classification Module

We utilized the feed-forward FCL layer as the last layer of our model. After concatenating the feature vectors of stacked LSTM and FCL, we obtained a

fused feature vector. Then the feed-forward FCL layer processes the fused feature vector and passes the output feature vector to the softmax layer. The equation for the feed-forward linear layer can be defined as follows:

$$K = Z * [M; N] + P$$

where K is the output feature vector of the feed-forward layer, M stands for m-dimensional features from the stacked LSTM layer, N stands for n-dimensional features from the FCL layer, ';' stands for concatenation of two feature vectors, Z and P are the input weight and bias of that layer. Softmax activation function provides the normalized probabilities with output classes as follows:

$$softmax(K_i) = \frac{e^{K_i}}{\sum_{i=1}^{n} e^{K_i}}$$

Here, K_i is the i-th element of feature vector K and n is the length of the vector. The final predicted label is the class with the highest probability score.

4 Experiments and Evaluations

4.1 Dataset Description

We used the benchmark dataset from SemEval-2023 Task 10 to evaluate how well our suggested solution performed. We employ the SemEval-2023 Task 10 published dataset (Kirk et al., 2023) for all of the subtasks, including Task A, Task B, and Task C. To create a diversified dataset, organizers gathered a substantial amount of corpora from Gab (Jasser et al., 2021) and Reddit (Ribeiro et al., 2021). Each entry was given three annotations, with the women annotators receiving top attention for reducing implicit bias. Organizers used 20,000 cleaned entries from the Reddit and Gab datasets for labeling. They later divided the chosen dataset into 70:10:20 segments as the train:development:test set.

The first subtask is to distinguish between sexist and non-sexist information; sexism is defined as an abusive response against women based on their gender and other similar characteristics. In total, there are 14000 train data, 10602 of which are classified as 'sexist', 3398 of which are classified as 'not sexist', 2000 development data, 1514 of which are classified as 'sexist' 486 of which are classified as 'not sexist' and 4000 test data, 3030 of which are classified as 'sexist', 970 of which are classified as 'not sexist'.

To broaden the scope of the categories of sexist data, the second subtask is identifying the four conceptual and analytical categories of sexist content, including

1) *Threats, plans to harm, and incitement:* This type of sexist material express the purpose to hurt women by encouraging others to do the same. The dataset for Task B contains 3398 train records, 486 development records, and 970 test records of which 310, 44, and 89 comprise of 'threats, plans to harm and incitement' subsequently.
2) *Derogation:* Disrespectful language that demeans and insults women. The train, development, and test set comprises 1590, 227, and 454 data of 'derogation' respectively.
3) *Animosity:* Implicit language that subtly reflects stereotypes, including sexism. In train set 1165, development set 167, and test set 333 are listed with 'animosity'.
4) *Prejudiced Discussion:* Contradictory rhetoric that rejects the presence of discrimination and rationalizes sexist behavior. The train, development, and test set covers 333, 48, and 94 data of 'prejudiced discussion' correspondingly.

Parameters & Embeddings	Settings		
	Task A	Task B	Task C
Learning Rate	2e-5	2e-5	2e-5
Max Epoch	8	8	18
Batch Size	4	4	4
Anneal Factor	0.5	0.5	0.5
Patience	3	3	3
Transformer document embeddings	"vinai/bertweet-base", Layers = "-1,-2,-3,-4", Layer Mean = True		
LSTM	Input Size = 768, Hidden Size = 1024, Number of Layers = 2, Bidirectional = False		
Linear Architecture	Input Size = 768, Output Size = 512		

Table 3: Parameters and Embeddings settings.

The third subtask is to break down the sexist content categories into 11 fine-grained sexism vectors to improve the categories of sexist data.

1.1) *Threats of harm:* State the intention to hurt a woman. Task C dataset consists of 3398 train data, 486 development data, and 970 test

data with 56, 8, and 16 instances of 'threats of harm' subsequently.

1.2) *Incitement and encouragement of harm:* Encourage an audience to mistreat a woman. The train, development, and test set comprises 254, 36, and 73 data of 'incitement and encouragement of harm' respectively.

2.1) *Descriptive attacks:* Negatively disparages women. The train, development, and test set encompasses 717, 102, and 205 examples of 'descriptive attacks' correspondingly.

2.2) *Aggressive and emotive attacks:* Demonstrate a strong bias against women. This level accommodates 673 instances of the train, 96 instances of development, and 192 instances of test set consequently.

2.3) *Dehumanising attacks and overt sexual objectification:* Comparing women to non-human creatures. 200 train data, 29 development data, and 57 test data consist of 'dehumanising attacks and overt sexual objectification' proportionately.

3.1) *Casual use of gendered slurs, profanities, and insults:* Utilize gender-based insults and profanities that unintentionally damage women. 637, 91, and 182 examples of train, development, and test set encompass with 'casual use of gendered slurs, profanities, and insults'.

3.2) *Immutable gender differences and gender stereotypes:* Highlighting the fundamental and innate distinctions between men and women. In train set 417, the development set 60, and in test set 119 data are 'immutable gender differences and gender stereotypes'.

3.3) *Backhanded gendered compliments:* Compliment women to minimize their inadequacy. 'backhanded gendered compliments' includes 64 instances of the train set, 9 instances of the development set, and 18 instances of the test set appropriately.

3.4) *Condescending explanations or unwelcome advice:* Offer patronizing advice to women while feigning to be more knowledgeable. The train, development, and test set comprises 47, 7, and 14 data of 'condescending explanations or unwelcome advice' properly.

4.1) *Supporting mistreatment of individual women:* Show support for the mistreatment of specific women. This level contains 75, 11, and 21 examples of all three evaluating datasets correspondingly.

4.2) *Supporting systemic discrimination against women as a group:* Exhibits support for the systematic discrimination of women as a group. In train set 258, development set 37, and test set 73 data incorporates 'supporting systemic discrimination against women as a group'.

Team (Rank)	**F1 score**
Task A: binary sexism detection	
CSECU-DSG(51th)	**0.8218**
Result of other competitors	
PingAnLifeInsurance(1st)	0.8746
stce(2nd)	0.8740
danch22(79th)	0.7184
NLP-CogSci(80th)	0.6325
Task B: category of sexism	
CSECU-DSG(40th)	**0.5986**
Result of other competitors	
JUAGE(1st)	0.7326
stce(3rd)	0.7203
OPEN SESAME(54th)	0.5591
PadmaDS(62th)	0.4782
Task C: fine-grained vector of sexism	
CSECU-DSG(28th)	**0.4419**
Result of other competitors	
PALI(1st)	0.5606
stce(2nd)	0.5487
SINAI(29th)	0.4376
PadmaDS(51th)	0.2866

Table 4: Comparative results (macro F1 score) with other participants on EDOS task. The result of our system is highlighted in boldface.

According to the dataset distribution, in Task A, 'sexist' content makes up about 76% of the entire dataset. In Task B, the 'derogation' field of the distributed dataset receives the biggest portion of 47% of the total. Consequently, in Task C, 'Descriptive attacks' absorb at a maximum level of 21% in the dataset distribution. We exploited the train set to learn the suggested technique for SemEval-2023 Task 10 and the development set to utilize for hyperparameter adjustment. Eventually, we employed the test set to appraise our system. Table 2 shows the distribution of the utilized dataset.

4.2 Model Configuration

For system training and parameter designs, we exploited the GPU of the Google Colab. In BERTweet embeddings, we exploit "vinai/bertweet-

Method	Task A			Task B			Task C		
	F1	P	R	F1	P	R	F1	P	R
BERTweet	0.8175	0.8127	0.8228	**0.6000**	0.6000	**0.6009**	0.4265	0.4348	**0.4298**
BERTweet + LSTM	0.8049	0.8050	0.8048	0.4709	0.4773	0.4791	0.3706	0.4130	0.3665
BERTweet + Stacked LSTM	0.7857	0.7890	0.7826	0.2881	0.2278	0.3929	0.1154	0.0970	0.1560
BERTweet + LSTM + FCL	0.8143	0.8154	0.8132	0.5822	0.5847	0.5832	0.4330	0.4567	0.4185
Proposed Method: BERTweet + Stacked LSTM + FCL	**0.8218**	**0.8199**	**0.8237**	0.5986	**0.6001**	0.5989	**0.4419**	**0.4674**	0.4278

Table 5: Individual component analysis on the test set of Task A, B, and C. The best results are highlighted in boldface.

base" (Nguyen et al., 2020b) and averaged its final four layers in embedding settings. A fully linked linear architecture and two stacked layers of the LSTM module were applied to a 768-dimensional vector from BERTweet, respectively, to produce 1024- and 512-dimensional feature vectors. The resultant vector of the stacked LSTM and the FCL layer are combined to create a 1536-dimensional fusion vector. The output of a 1536-dimensional concatenated vector is procured by a further feed-forward FCL output layer.

We explore various epochs, including 4,8,12,15, batch sizes, including 4,8, and learning rates, including 1e-5, 2e-5, 3e-5, and 4e-5 in hyper-parameter settings. Except for epoch rates, we achieved our competitive result at learning rates of 2e-5, batch sizes of 4, and anneal factors of 0.5. Epoch rates for Task A, Task B, and Task C are 8, 8, and 18, correspondingly. We present our system settings in Table 3.

4.3 Result

In this section, we present the performance of our system in the SemEval-2023 EDOS task. We have compared the result of our system with other participants' systems for three subtasks in Table 4. Here, the primary evaluation measure is the macro F1 score according to the organizers of the task.

From this result, it is depicted that our approached model obtained comparative performance in Task A, Task B, and Task C. Among these three tasks, our system achieved the highest score in Task A with an 82% F1 score. However, our system is ranked 51st in this Task, and it is trailing by 5.28% to the system in the first position, proposed by PingAnLifeInsurance. In Task B, it scored 18% less than the top team JUAGE with an F1 score of 0.5986 and ranked 40th. In Task C, it ranked 28th with an F1 score of 0.4419, that is 21.15% less than the top team PALI. Our proposed system scored comparatively less in Task C, the 11-class classification. Our system requires more data to obtain the characteristics of these 11 fine-grained vectors and achieve a better score in this Task.

4.4 Discussion

We carried out component analysis experiments on the test set to assess the significance of the individual components of our proposed system. Table 5 demonstrates the findings of this study in terms of precision (P), recall (R), and F1-score. Our proposed system surpasses other combinations by at least 0.43% and up to 3.61% in F1-score for Task A, and by at least 1.54% and up to 32.65% in F1-score for Task C. In the case of Task B, BERTweet shows slightly better performance, resulting in a 0.14% increase in F1-score compared to our proposed system, which can be deemed negligible. Even though BERTweet performs better and the aggregation of BERTweet and LSTM variants slightly degrades the performance of the system compared to BERTweet, our proposed approach still performs competitively and yields the most optimal outcomes for Task A and C, highlighting the effectiveness of combining these methods.

5 Conclusion

We presented our methodology for the SemEval-2023 EDOS task in this publication. We extracted transformer-based document embeddings and integrated them with stacked LSTM and an FCL layer. Our proposed approach obtained compar-

ative performance in all the subtasks. Transformer-based document embeddings can capture the intricate word relationships in a sentence, and stacked LSTM can extract the deep-level patterns of embeddings which helps our model to improve its performance. The inclusion of additional features or data sources could enhance the performance of our approach. We also plan to explore parameter tunings more meticulously, that can improve our system for multiclass classification.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: a study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in human behavior*, 52:436–442.

Simona Frenda, Ghanem Bilal, et al. 2018. Exploration of misogyny in spanish and english tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Greta Jasser, Jordan McSwiney, Ed Pertwee, and Savvas Zannettou. 2021. 'welcome to# gabfam': Far-right virtual community on gab. *New Media & Society*, page 14614448211024546.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Chelsea Litchfield, Emma Kavanagh, Jaquelyn Osborne, and Ian Jones. 2018. Social media and the politics of gender, race and identity: The case of serena williams. *European Journal for Sport and Society*, 15(2):154–170.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020a. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020b. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *CEUR Workshop Proceedings*, volume 2150, pages 234–241. CEUR-WS.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584.

Sima Sharifirad, Alon Jacovi, Israel Bar Ilan Univesity, and Stan Matwin. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *WNLP@ ACL*, pages 21–23.

Sima Sharifirad and Stan Matwin. 2019. When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in nlp. *arXiv preprint arXiv:1902.10584*.

Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding LSTM - a tutorial into long short-term memory recurrent neural networks. *CoRR*, abs/1909.09586.

Janet K Swim, Robyn Mallett, and Charles Stangor. 2004. Understanding subtle sexism: Detection and use of sexist language. *Sex roles*, 51:117–128.

Jin Wang, Bo Peng, and Xuejie Zhang. 2018. Using a stacked residual lstm model for sentiment intensity prediction. *Neurocomputing*, 322:93–101.

Zeerak Waseem. 2016. 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*, pages 138–142.