Knowledge-based Neural Framework for Sexism Detection and Classification

Harika Abburi, Shradha Sehgal, Himanshu Maheshwari, and Vasudeva Varma

LTRC, IIIT-Hyderabad, India
harika.a@research.iiit.ac.in, shradha.sehgal@students.iiit.ac.in
himanshu.maheshwari@research.iiit.ac.in, vv@iiit.ac.in

Abstract. Sexism, a prejudice that causes enormous suffering, manifests in blatant as well as subtle ways. As sexist content towards women is increasingly spread on social networks, the automatic detection and categorization of these tweets/posts can help social scientists and policymakers in research, thereby combating sexism. In this paper, we explore the problem of detecting whether a Twitter/Gab post is sexist or not. We further discriminate the detected sexist post into one of the fine-grained sexism categories. We propose a neural model for this sexism detection and classification that can combine representations obtained using RoBERTa model and linguistic features such as Empath, Hurtlex, and Perspective API by involving recurrent components. We also leverage the unlabeled sexism data to infuse the domain-specific transformer model into our framework. Our proposed framework also features a knowledge module comprised of emoticon and hashtag representations to infuse the external knowledge-specific features into the learning process. Several proposed methods outperform various baselines across several standard metrics.

Keywords: Sexism detection · Sexism classification · Transformers · Knowledge-base.

1 Introduction

Sexism encompasses stereotypes, prejudice, or discrimination based on a person's sex or gender, most often women. It occurs in various subtle and overt forms, causing immense suffering to women and girls. Inequality and sexism against women prevalent in society are constantly being mirrored on the internet. Women are affected in many aspects of their lives, including domestic and parental responsibilities, work prospects, sexual appearance, life desires, by the most nuanced manifestations of sexism. So, the automatic detection and categorization of sexism into well-defined categories may help to analyze sexism to

IberLEF 2021, September 2021, Málaga, Spain.

improve sensitization programs and put in place other mechanisms to combat this oppression. It could also aid in the development, design, propagation of new equality policies and as well as to promote positive societal action.

The detection of sexism varies from and may supplement the classification of sexism. Sexism detection will be used to identify the sexist posts where instances of sexism are mixed with other posts unrelated to sexism on which to perform sexism classification. Some hate speech classification works [4, 25] detect sexism as a type of hate; however, it does not conduct sexism classification. The prior work on classifying sexism [3, 13, 11, 25] has identified explicit hatred or violence towards women and classified the given tweet into two to five categories. In this paper, we examine the problem of detecting and classifying sexism in a broad context, from overt misogyny to other indirect phrases that include latent sexist behaviours. More specifically, we attempt to solve two classification problems using tweets and gab posts in Spanish and English. Firstly, we assign a binary label to a tweet indicating whether it is sexist or not. Secondly, if the tweet is classified as sexist, we classify it further into fine-grained sexism categories.

We develop a novel neural framework for sexism detection and classification that enables a flexible combination of text representations generated by the domain-specific transformer model with the linguistic and semantic feature representations through recurrent operations. Our model can be better equipped to capture the semantic aspects by using general-purpose transformer models like RoBERTa [17] since it is trained on text data that is much larger than the domain-specific labeled data we have. However, in order to incorporate domain-specific elements into our model, we further retrain the RoBERTa model using the unlabeled instances of sexism. The representations from this domain-specific model complement the representations built from linguistic and semantic features such as Empath [10], Hurtlex [5] and Google's Perspective API https://www.perspectiveapi.com/ as a function of end-to-end trainable neural network parameters. Further, to fully comprehend the style of text, we infuse external knowledge information into our framework by leveraging the pragmatics of emojis, smileys, and the specific context of the hashtags as additional context representations. Our experimentation has shown that multiple instances of the proposed framework outperforming several diverse baselines on established metrics.

Our key contributions are summarized below.

- We propose a neural framework that can combine post representations built from different linguistic features with those created using the transformer model through learnable model parameters.
- Our proposed framework is also aided by external knowledge information by leveraging the hashtags and emojis present in the Tweets or Gab posts.
- The proposed methods outperform numerous baselines across established metrics.

2 Related Work

In this section, we first look at work on hate speech detection and classification since some of it is related to our work in some ways, such as detecting sexist hate. We then explain previous research on sexism classification.

The detection of sexism is performed by some hate speech classification approaches that include sexism as a category of hate [25, 4, 7, 29]. [11] presents an approach for detecting sexism and misogyny from tweets. [6] build a data-driven model of cyberhate to identify disability, race, and sexual orientation using bag-of-words, dictionary, and text parser to extract typed dependencies. [25, 24] classified tweets as sexist, racist, or neither using character n-grams along with extra-linguistic features. Deep learning algorithms such as fastText, RNN, and CNN are investigated by [4] to classify the given tweet as racist, sexist, and neither. [30] proposed deep learning ensemble techniques on the existing datasets. [20] provide a hierarchical Conditional Variational Autoencoder model for fine-grained hate speech classification. [29] explored the word embeddings with a combination of GRU and CNN and skipped CNN to classify tweets as sexism, racism, both, and non-hate.

In [13], tweets are classified as benevolent, hostile, or non-sexist using biLSTM with attention, SVM, and fastText. Using features such as Part of Speech (POS) identifiers, n-grams, and text embedding, tweets described as misogynist, are categorized as stereotype and objectification, discredit, threats of violence, sexual harassment, and dominance, or derailing in [3]. [15] has investigated deep learning strategies for classifying tweets of sexual violence but has not directly focused on developing a comprehensive method to detect recollections of personal stories of abuse. [12] use ConceptNet and Wikidata to classify the tweets related to sexual harassment by text augmentation and text generation. The first dataset of sexist phrases and attitudes in Spanish on Twitter (MeTwo) is created [21] and investigates the possibility of using both conventional and novel deep learning models for automatically detecting various forms of sexist conduct.

In [27], a density matrix encoder inspired by quantum mechanics is used for the classification of personal stories of sexual harassment. [14] explores CNN, RNN, and a combination of them for categorizing personal experiences of sexual harassment into one or more of three classes. [18, 1] explores multi-label categorization of accounts reporting any kind(s) of sexism. They developed supervised and semi-supervised methods for classifying sexism at the fine-grained level using transformer models such as BERT [8]. While their study focuses on what an incident of sexism entails, where it happens, and who perpetrates it, but our work focuses on detection and categorization of sexism pertains to how it is stated. [2] propose a multi-task approach to perform multi-label sexism classification. They use sexism detection as one of the auxiliary tasks in a multi-task setup. In contrast, our work performs the sexism detection task explicitly. Our work also makes use of cutting-edge transformer models [23] that have been trained on vast amounts of data to produce stable and semantically rich embeddings which can be used for downstream tasks such as sexism detection and classification.

3 Dataset

In this section, we will explain both labeled and unlabeled datasets that are used for the sexism detection and classification tasks. We also provided the label distribution across the languages for both the tasks.

3.1 Labeled data

The dataset for the tasks were provided by the organizers of the EXIST shared task [22]. The dataset is composed of tweets and gab posts in two languages: English and Spanish. It consists of 6977 tweets for training and 3386 tweets for testing. Test set also has 492 gabs in English and 490 in Spanish from the uncensored Gab social network.

The first subtask is a binary classification, predicting whether the given text (tweet or gab post) is sexist (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour) or non-sexist. Table 1 shows the number of posts present in each class for both languages. Once a message has been classified as sexist, the second task aims to categorize the given text into five sexism categories. The number of posts in each class for both languages is shown in table 2.

Table 1. Label distribution across the languages for sexism detection task.

Category	English	Spanish
Sexist	2794	2864
Non-sexist	2850	2837

1. **Ideological and Inequality:** This class includes text that criticises the feminist movement, condemns gender discrimination, or portrays men as victims of gender-based oppression.

Table 2. Label distribution across the languages for sexism classification task.

Category	English	Spanish
Ideological and Inequality	719	768
Stereotyping and Dominance	628	645
Objectification	406	418
Sexual Violence	542	375
Misogyny and Non-sexual Violence	499	658

2. **Stereotyping and Dominance:** The text presents false views about women, such as that they are more desirable for certain positions or that they are inadequate for certain activities, or that men are somehow superior to women.
3. **Objectification:** The text portrays women as subjects separate from their integrity and personal aspects, or defines those physical characteristics that women must have in order to meet conventional gender norms.

4. **Sexual Violence:** Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.
5. **Misogyny and Non-Sexual Violence:** The text expresses hatred and violence towards women.

In this work, we converted all Spanish posts to English using the Microsoft Translator API and use the entire data to perform the experiments.

3.2 Unlabeled data

We also investigate unlabeled data in order to construct the domain-specific transformer model. We make use of both unlabelled (crawled) and labeled data created by [18]. They crawled over 90,000 unlabeled examples of sexism from the Everyday Sexism Project, which includes hundreds of thousands of sexism accounts from witnesses and survivors. Along with these unlabeled instances, we also use the 13023 labelled data after removing the labels. We used a total of 1,03,023 unlabeled posts in our work.

4 Proposed Sexism Detection and Classification Approach

In this section, we detail our approaches for carrying out the detection and classification of sexist posts. We begin the section with the description of the features explored. Next, we discuss the proposed neural architecture, which enables combining different post representations by infusing external knowledge information into the learning process.

4.1 Features Explored

We investigated linguistic and semantic features to see if they could aid in detecting and classifying sexist posts.

Perspective API (*Pe*): Google's Perspective API is an API that uses machine learning algorithms to predict the perceived effect of a text by analysing different emotional concepts. The API provides scores in real numbers between 0 and 1. The API provides the following attributes for English text: toxicity, severe_toxicity, identity_attack, insult, profanity, threat, sexual_explicit, obscene, and flirtation. For each sentence in a post, we created 9-dimensional feature vector.

HurtLex (*Hu*): HurtLex is a lexicon of aggressive, offensive, and hateful words/phrases. These hateful words are divided into 17 categories. It also has a 2-level structure named conservative and inclusive. Conservative lemmas are obtained by translating offensive senses of the words in the original lexicon. Inclusive lemmas are obtained by translating all the potentially relevant senses of the words in the original lexicon.

We consider nine categories that are relevant to our problem; negative stereotypes ethnic slurs (PS), professions and occupations (PA), physical disabilities and diversity (DDF), cognitive disabilities and diversity (DDP), female genitalia (ASF), words related to prostitution (PR), words related to homosexuality (OM), with potential negative connotations (QAS), and derogatory words (CDS). We consider both conservative and inclusive levels results in an 18-dimensional feature vector. For each sentence in a post, we check if the lexicon's phrases/words present in these categories and create a feature vector with a frequency of each category.

Empath (*Em*): Empath is a tool for analyzing text across lexical categories (similar to LIWC) and can also generate new lexical categories. Empath draws connotations between words and phrases by deep learning across more than 1.8 billion words of modern fiction.

We filtered out the built-in human-validated categories relevant to our task. The categories chosen were sexism, violence, money, valuable, domestic_work, hate, aggression, anticipation, crime, weakness, horror, swearing_terms, kill, sexual, cooking, exasperation, body, ridicule, disgust, anger, and rage result in a 21 dimension feature vector.

4.2 Proposed Architecture

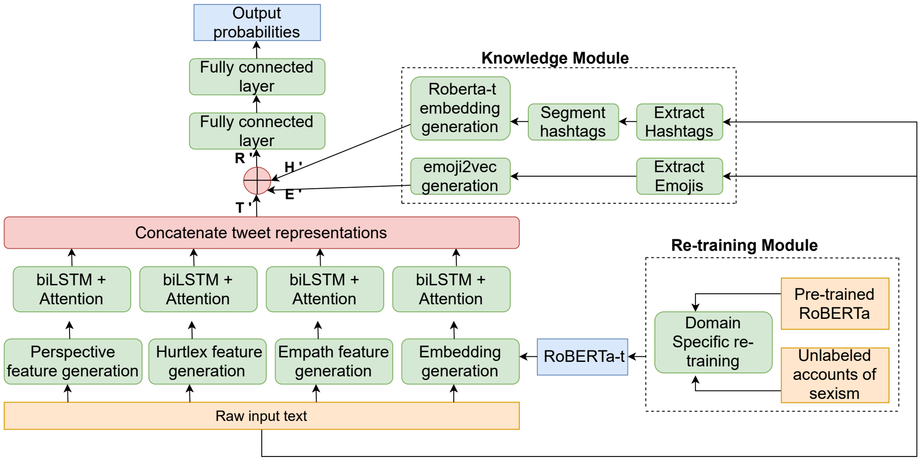

Fig. 1. Proposed knowledge-based neural architecture

Figure 1 depicts our proposed architecture. Each tweet or gab post (raw input text) is passed to the Perspective API, Empath tool, Hurtlex lexicon, and transformer model to generate the sentence representations. We employ a domain-adapted RoBERTa variant named *RoBERTa_t* for generating more effective sentence representations than those produced by off-the-shelf RoBERTa models. *RoBERTa_t* is created by further training a pre-trained RoBERTa model in an unsupervised manner using unlabelled accounts of sexism. We incorporate *RoBERTa_t* into our end-to-end training and the weights are updated during the training. Further, these sentence representations are passed to the bidirectional LSTM and an associated attention mechanism [28] to generate the entire text representation. For each sentence in a text, the biLSTM layer produces an h-dimensional output length. These output lengths are aggregated into a vector representation by the attention layer. Overall, this results in four different text representations, which are then concatenated to generate the final post representation T'.

To improve the efficiency of our model, we add external information to the current architecture, which includes knowledge of the pragmatics of emojis and smileys, as well as the context in which those hashtags are used. The motivation behind this is that the model may not capture the characteristics and true meaning of emojis and hashtags present in the text. We utilize emoji2vec [9] to obtain a semantic vector representing the particular emoji. The hashtags are segment into meaningful tokens using the ekphrasis segmenter https://github.com/cbaziotis/ekphrasis. We generate the segmented hashtag embeddings using the *RoBERTa_t* such that the text representations and hashtag embeddings are grounded in the same latent space. For each input text, to obtain the centralized emoji representation E' and hashtag representation H', we average the vector representations of all the individual emojis and the segmented hashtags. Then we concatenate all three representations T', E', and H' to get the final tweet representation R'. This R' is passed to the fully connected layer. Finally, a fully connected layer with nonlinearity generates the output probabilities.

5 Experiments

This section provides the experimental evaluation of the proposed methods present analysis. Our code, as well as all hyper-parameter values, are available at GitHub `https://github.com/Harikavuppala1a/EXIST_shared-task`.

For sexism detection we reported results on F score (F) and Accuracy (Acc) and for sexism classification results are reported on F weighted (F_w), F macro (F_{macro}), and Accuracy (Acc).

5.1 Baselines

Random

For each test sample, labels are selected randomly as per their normalized frequencies in the training data.

Traditional Machine Learning (TML)

We report the performance using Support Vector Machine (SVM), Logistic

Regression (LR), Gradient Boosted Trees (GBT), and Random Forests (RF), each applied on three feature sets, namely the word n-grams, character n-grams, and the average of the ELMo vectors [19] for a post's words (ELMO).

Deep Learning (DL)

- biLSTM: The word embeddings correspond to each post are fed through a bidirectional LSTM.
- biLSTM-Attention: The biLSTM-Attention is similar to biLSTM, but with the attention scheme from [28].
- C-biLSTM: This architecture is somewhat similar to approach [14]. After the convolution operation has been carried out on the word vectors of each post, the filter dimensions are stacked to create a series of window vectors that are then transmitted through biLSTM.
- CNN-Kim: Word vectors of a post are passed through convolutional and max-over-time pooling layers similar to [16].
- BERT and RoBERTa : Sentence embeddings are generated using BERT via bert-as-service [26] and RoBERTa and passed through a biLSTM with attention separately.

Table 3. Results with traditional machine learning: (1) Sexism detection (2) Sexism classification

Classifier	Features	Sexism Detection		Sexism Classification		
		F	**Acc**	$\mathbf{F_W}$	$\mathbf{F_{macro}}$	**Acc**
SVM	Word n-grams	0.691	0.689	0.568	0.481	0.572
	Character n-grams	0.697	0.698	**0.574**	**0.485**	0.578
	ELMo average	0.662	0.651	0.474	0.372	0.472
LR	Word n-grams	0.703	0.707	0.507	0.380	0.574
	Character n-grams	0.686	0.701	0.511	0.377	0.580
	ELMo average	0.672	0.669	0.520	0.413	0.537
GBT	Word n-grams	0.676	0.695	0.560	0.454	0.602
	Character n-grams	0.687	0.703	0.540	0.422	**0.594**
	ELMo average	0.686	0.689	0.482	0.343	0.548
RF	Word n-grams	**0.717**	**0.712**	0.557	0.457	0.591
	Character n-grams	0.680	0.696	0.520	0.397	0.582
	ELMo average	0.644	0.667	0.356	0.164	0.499

5.2 Results

We set aside 15% from original labeled data for validation. During the testing phase, the validation set was merged with the training set. For all the methods, the mean of the results obtained over three runs is given for each metric. Most proposed methods outperform all baselines across all metrics.

Table 3 shows results produced using four traditional ML methods (SVM, LR, GBT, and RF) across three different feature sets (word n-grams, character

Table 4. Results for Sexism detection (baselines use ELMo embeddings)

	Approach		**F**	**Acc**
Baselines	Random		0.502	0.498
	biLSTM		0.691	0.698
	biLSTM-Attention		0.731	0.716
	CNN-Kim		0.740	0.713
	C-biLSTM		0.738	0.720
	BERT		0.705	0.697
	RoBERTa		0.725	0.712
Proposed methods	***biL-att* applied on**	**External knowledge**		
	RoBERTa_t		0.752	0.749
	RoBERTa_t	Hashtag	0.761	0755
	RoBERTa_t	Emoji, Hashtag	0.756	0.754
	RoBERTa_t, Em	Emoji	0.765	0.759
	RoBERTa_t, Em	Hashtag	0.761	0.758
	RoBERTa_t, Hu	Emoji	0.768	0.759
	RoBERTa_t, Pe		0.765	0.756
	RoBERTa_t, Pe, Hu		0.760	0.757
	RoBERTa_t, Pe, Hu	Emoji, Hashtag	0.763	0.754
	RoBERTa_t, Em, Pe		0.765	0.756
	RoBERTa_t, Em, Pe	Emoji, Hashtag	0.761	0.756
	RoBERTa_t, Pe, Em, Hu		0.763	0.758
	RoBERTa_t, Pe, Em, Hu	Emoji	0.764	0.757
	RoBERTa_t, Hu, Em	Hashtag	0.759	0.755
	RoBERTa_t, Hu, Em		**0.766**	**0.760**

n-grams, and ELMo average) for both the tasks. For SVM and LR, we apply class imbalance correction across both tasks. Among these combinations, RF with word n-grams emerges as the top sexism detection method. SVM with character n-grams produces the best F scores for sexism classification.

Table 4 and Table 5 provides sexism detection and sexism classification results for random, deep learning and various combinations of proposed framework. For proposed methods, the sub-columns in each row specify which neural method or linguistic features are used to generate post representations and the knowledge information employed. We note that the results are reported for a subset of the possible outcomes of the proposed neural framework.

In Table 4, for sexism detection task, the random method performs poorly as expected. The best deep learning baseline is C-biLSTM based on Acc, and it outperforms its traditional ML counterpart. Our best method involves bil-att processing on *RoBERTa_t, Hu, and Em* separately to generate the final post representations. When external information such as Emoji and Hashtag representations are concatenated with this best method, the performance is marginally reduced. In Table 5, also the random method performs poorly as expected, reflecting the challenging nature of the sexism classification. The best deep learning baseline is biLSTM-Attention based on F_{macro}, and it outperforms its traditional ML counterpart. Among the various combinations of proposed methods, our best

Table 5. Results for Sexism classification (baselines use ELMo embeddings)

	Approach		$\mathbf{F_w}$	$\mathbf{F_{macro}}$	**Acc**
Baselines	Random		0.292	0.165	0.299
	biLSTM		0.561	0.481	0.558
	biLSTM-Attention		0.571	0.505	0.564
	CNN-Kim		0.564	0.470	0.588
	C-biLSTM		0.582	0.500	0.587
	BERT		0.513	0.449	0.499
	RoBERTa		0.536	0.462	0.527
Proposed methods	***biL-att* applied on**	**External knowledge**			
	RoBERTa_t		0.626	0.544	0.629
	RoBERTa_t	Emoji	0.632	0.551	0.636
	RoBERTa_t	Hashtag	0.630	0.548	0.634
	RoBERTa_t	Emoji, Hashtag	0.632	0.549	0.635
	RoBERTa_t, Em	Emoji	0.633	0.552	0.637
	RoBERTa_t, Em	Emoji, Hashtag	0.633	0.550	0.635
	RoBERTa_t, Pe	Hashtag	0.633	0.551	0.639
	RoBERTa_t, Hu	Emoji	0.631	0.549	0.637
	RoBERTa_t, Em, Pe	Hashtag	0.633	0.554	0.634
	RoBERTa_t, Em, Pe	Emoji, Hashtag	0.632	0.550	0.635
	RoBERTa_t, Hu, Em	Emoji, Hashtag	0.628	0.547	0.630
	RoBERTa_t, Pe, Em, Hu	Hashtag	0.630	0.548	0.633
	RoBERTa_t, Pe, Em, Hu	Emoji, Hashtag	0.628	0.545	0.629
	RoBERTa_t, Pe, Hu		0.629	0.545	0.634
	RoBERTa_t, Pe, Hu	Emoji	0.633	0.551	0.637
	RoBERTa_t, Pe, Hu	Hashtag	**0.635**	**0.555**	**0.638**

method uses the biL-att processing on *RoBERTa_t, Pe, and Em* features along with the external hashtag context vector.

Overall, we observed that linguistic features, when combined with the domain-specific transformer model involving recurrent components, help to improve the detection and classification performance to some extent. It has also been noted that in most cases, the external knowledge information (Emoji and Hashtag features) helps to improve the performance even further. Furthermore, several variants of the proposed framework outperform all baselines across all the metrics.

Figure 2 compares the class-wise performance of our best method (*Roberta_t, Pe, Hu, Hashtag*) with that of the best baseline (biLSTM-Attention) for sexism classification. For each class, the average of the F scores over three runs are shown for both methods. For all the classes, the F score of the proposed method outperforms the baseline F score.

We analyze the impact of our proposed sexism detection and classification methods on different social networks, including Twitter ("with consent control") and Gab.com ("without content control"). Table 6 shows the total number of posts present in the test set for each network and the number of correctly predicted posts by baselines and proposed methods for both tasks. Our proposed

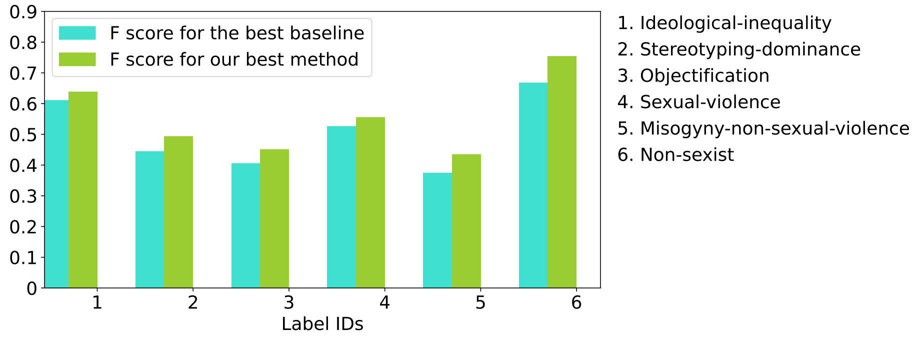

Fig. 2. Class-wise sexism classification F-scores for the best performing baseline (biLSTM-Attention) and our best method (*Roberta_t, Pe, Hu, Hashtag*)

approaches perform effectively on both social media posts compared to the baselines with a reasonable margin for both tasks.

Table 6. Analysis of different social media posts: (1) Sexism detection: Best baseline (C-biLSTM), Best proposed method (*Roberta_t, Hu, Em*) (2) Sexism classification: Best baseline (biLSTM-Attention), Best proposed method (*RoBERTa_t, Pe, Hu, Hashtag*)

Type of posts	**Posts in test set**	**Sexism Detection**		**Sexism Classification**	
		Best baseline	Best proposed method	Best baseline	Best proposed method
Tweets	3386	2437	2564	1878	2187
Gab posts	982	726	749	532	610

6 Conclusion

In this paper, we explored the sexism detection and fine-grained classification of tweets and Gab posts. We developed a knowledge-based neural framework that combines representations created using linguistic features and those obtained using the RoBERTa model trained in an end-to-end manner. We capitalized on unlabeled data to build the domain-specific transformer model. Our experiments show that the external knowledge representations fed into the neural framework aided in boosting the performance. All the variants of the proposed approach outperforms several deep learning and traditional machine learning baselines. Our analysis showed that our proposed approach is also effective at detecting and classifying the posts in Gab.com where the abusive content is not restricted. Directions for future work include developing approaches that conduct sexism classification more accurately as well as exploring the multilingual sexism detection and classification.

References

1. Abburi, H., Parikh, P., Chhaya, N., Varma, V.: Fine-grained multi-label sexism classification using semi-supervised learning. In: International Conference on Web Information Systems Engineering. pp. 531–547. Springer (2020)
2. Abburi, H., Parikh, P., Chhaya, N., Varma, V.: Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5810–5820. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
3. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In: International Conference on Applications of Natural Language to Information Systems. pp. 57–64. Springer (2018)
4. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. International World Wide Web Conferences Steering Committee (2017)
5. Bassignana, E., Basile, V., Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In: 5th Italian Conference on Computational Linguistics, CLiC-it 2018. vol. 2253, pp. 1–6. CEUR-WS (2018)
6. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Science **5**(1), 11 (2016)
7. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh international aaai conference on web and social media (2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., Riedel, S.: emoji2vec: Learning emoji representations from their description. arXiv preprint arXiv:1609.08359 (2016)
10. Fast, E., Chen, B., Bernstein, M.S.: Empath: Understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI conference on human factors in computing systems. pp. 4647–4657 (2016)
11. Frenda, S., Ghanem, B., Montes-y Gómez, M., Rosso, P.: Online hate speech against women: Automatic identification of misogyny and sexism on twitter. Journal of Intelligent & Fuzzy Systems **36**(5), 4743–4752 (2019)
12. Jafarpour, B., Matwin, S., et al.: Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 107–114 (2018)
13. Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of the second workshop on NLP and computational social science. pp. 7–16 (2017)
14. Karlekar, S., Bansal, M.: Safecity: Understanding diverse forms of sexual harassment personal stories. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2805–2811 (2018)
15. Khatua, A., Cambria, E., Khatua, A.: Sounds of silence breakers: Exploring sexual violence on twitter. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 397–400 (2018)

16. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751 (2014)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
18. Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., Varma, V.: Multi-label categorization of accounts of sexism using a neural framework. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1642–1652 (2019)
19. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of NAACL (2018)
20. Qian, J., ElSherief, M., Belding, E., Wang, W.Y.: Hierarchical cvae for fine-grained hate speech classification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3550–3559 (2018)
21. Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L.: Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access **8**, 219563–219576 (2020)
22. Rodríguez-Sánchez, F., de Albornoz, J.C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T.: Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural **67**(0) (2021)
23. Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems pp. 5998–6008 (2017)
24. Waseem, Z.: 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. pp. 138–142 (2016)
25. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
26. Xiao, H.: bert-as-service. `https://github.com/hanxiao/bert-as-service` (2018)
27. Yan, P., Li, L., Chen, W., Zeng, D.: Quantum-inspired density matrix encoder for sexual harassment personal stories classification. In: 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 218–220. IEEE (2019)
28. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
29. Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web pp. 1–21 (2018)
30. Zimmerman, S., Kruschwitz, U., Fox, C.: Improving hate speech detection with deep learning ensembles. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)