Not All Swear Words Are Used Equal: Attention over Word n-grams for Abusive Language Identification

Horacio Jesús Jarquín-Vásquez(✉), Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
{horacio.jarquin,mmontesg,villasen}@inaoep.mx

Abstract. The increasing propagation of abusive language in social media is a major concern for supplier companies and governments because of its negative social impact. A large number of methods have been developed for its automatic identification, ranging from dictionary-based methods to sophisticated deep learning approaches. A common problem in all these methods is to distinguish the offensive use of swear words from their everyday and humorous usage. To tackle this particular issue we propose an attention-based neural network architecture that captures the word n-grams importance according to their context. The obtained results in four standard collections from Twitter and Facebook are encouraging, they outperform the F_1 scores from state-of-the-art methods and allow identifying a set of inherently offensive swear words, and others in which its interpretation depends on its context.

Keywords: Abusive language · Text classification · Attention mechanism · Social media

1 Introduction

The exponential growth of user interactions through social media has revolutionized the way we communicate and share information. Unfortunately, not all these interactions are constructive; it is common to see that users make use of abusive language to criticize others, disqualify their opinions, or win an argument. As a consequence, affected users may present some psychological damage, and even, in extreme cases, commit suicide [9]. This situation has stimulate the interest of social media companies and governments in the automatic identification of abusive language.

Abusive language is characterized by the presence of insults, teasing, criticism and intimidation. Mainly, it includes epithets directed at an individual's characteristic, which are personally offensive, degrading and insulting. Its identification in social media is not an easy task, the use of word filters and moderators is far from being a good and sustainable solution to the problem.

K. M. Figueroa Mora et al. (Eds.): MCPR 2020, LNCS 12088, pp. 282–292, 2020.
https://doi.org/10.1007/978-3-030-49076-8_27

One of the most important issues in the abusive language identification task is to distinguish between the use of swear words and vulgarities in offensive and non-offensive contexts. As an example consider the following two tweets using the word "fucking"[1]: `"@USER You're a fucking idiot"` and `"@USER I'm so fucking ready"`. They clearly show that the importance and interpretation of a word is highly context dependent, and accordingly they evidence one of the reasons why traditional bag-of-words methods and deep learning models tend to generate many false positives in their predictions.

Only few works related to abusive language identification have explored the importance of words in accordance to their context; particularly, the use of attention mechanisms has been the most used approach to handle this issue [5,14]. The idea behind attention is to provide the classification model with the ability to focus on a subset of inputs (or features), handling in this way the importance of words in their context. However, this importance has been only observed at a single word level. We hypothesize that not only the interpretation of swear words is highly context dependent, but also the meaning of certain word sequences, and, therefore, that extending the use of attention to word sequences will allow capturing distinctive patterns for the abusive language identification task. As shown in the previous examples, word n-grams such as `"fucking idiot"`, `"fucking ready"`, and even `"You're"` and `"I'm"`, are very important in discriminating offensive from non-offensive posts.

The main contribution of this work is the extraction of two groups of swear word expressions relevant for the task of abusive language identification, one consisting of inherently offensive word sequences, and another consisting of word sequences with context-dependent offensive interpretation. To extract these word patterns, we propose an attention-based deep neural network architecture that allows capturing the importance of word n-grams, and an approach to extract and visualize inherently and context-dependent offensive word sequences, through the attention weights of our proposed architecture.

2 Related Work

Several works have proposed different models and datasets for the task of automatic abusive language identification [6,9,16,18]. Among them, a great variety of features have been used to tackle this problem. Initial works used bag-of-words representations, considering word n-grams as well as character n-grams as features [4,13,16]. Aiming to improve the generalization of the classifiers, other works have also considered word embeddings as features [13,19]. More recently, some works have used sophisticated text representations by applying pre-trained ELMO and BERT models, and fine-tune their parameters to the abusive language identification task [10,12].

Regarding the classification stage, different approaches and techniques have also been proposed. These approaches could be divided in two categories; the first category relies on traditional classification algorithms such as Support Vector

[1] Taken from the Offensive Language Identification Dataset [18].

Machines, Naive Bayes, Logistic Regression and Random Forest [4,6,8,15,16], on the other hand, the second category includes deep learning based methods, which employ Convolutional Neural Networks (CNN) for word and character based feature extraction [2,7], Recurrent Neural Networks (RNN) for word and character dependency learning [2,5], and the combination of both for creating powerful structures that capture order information between the extracted features [19].

Finally, it is important to mention that some recent works in abusive language identification have considered deep learning methods with attention mechanisms. One of the first works introducing the attention into the task employed self-attention models to detect abuse in portal news and Wikipedia [14]. Subsequently, [5] showed that contextual attention improved the results of self-attention in this task. Contextual attention was first introduced by [17] with the use of a hierarchical contextual attention neural network, based on a Gated Recurrent Unit (GRU) architecture, and used for document classification. Motivated by the results from [5,17], in this paper we extend the use of contextual attention by proposing an attention-based deep neural network architecture that attempt to capture the word n-grams importance, and also by presenting an approach that measures and plots the relevance of word sequences in accordance to their context.

3 Proposed Method for Abusive Language Identification

Figure 1 shows the general architecture of the proposed attention-based deep neural network for abusive language identification. This architecture consists of the following four major stages.

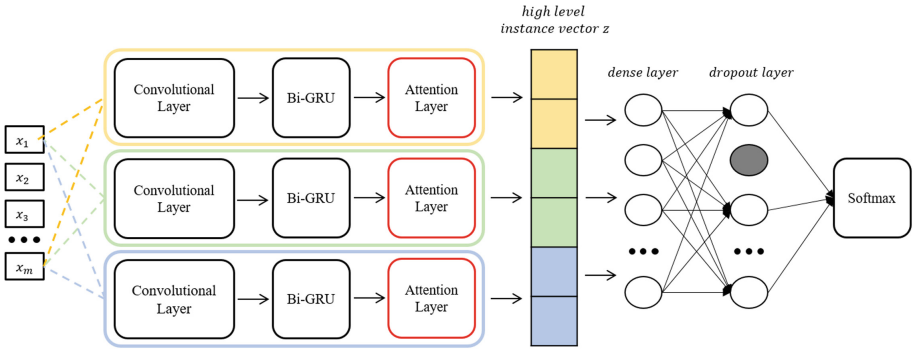

Fig. 1. Attention-based deep neural network architecture.

First stage: it corresponds to the input layer, which receives a sequence of m d-dimensional word vectors x_i; in other words, an input matrix of size $d \times m$.

Second stage: it is conformed by the convolutional layers, the bidirectional GRU layers, and the attention layers. We use an arrangement of them for each one of the considered channels, with the purpose of independently computing the weights of n-grams of different lengths. In the figure, the yellow, green, and blue rectangles correspond to the unigrams, bigrams and trigrams channels respectively.

In particular, the convolutional layer is used to extract different features of word n-grams from the input sequence X. The output of the convolutional layer passes to the bidirectional GRU layer to accomplish the sequence encoding. This layer captures word n-grams annotations by summarizing information from both directions. To get a word n-gram annotation h_i (Eq. 1), the forward and backward hidden states of the bidirectional GRU are concatenated, this summarizes the information of the whole sequence centered around the word n-gram annotation.

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \tag{1}$$

Since not all word n-grams contributes equally for the meaning and representation of an instance, we used the attention layer to extract the importance of each word n-gram, and combine them with its corresponding word n-gram annotation h_i, forming a new instance vector v. Below are listed the attention mechanism equations.

$$u_i = tanh(W_h.h_i + b_h) \tag{2}$$

$$\alpha_i = exp(u_i^T u_h) / \sum_j exp(u_j^T u_h) \tag{3}$$

$$v = \sum_j \alpha_j h_j \tag{4}$$

To obtained the instance vector v (refer to Formula 4), we first feed each word n-gram annotation h_i to a one Multi-Layer Perceptron (MLP) layer, getting this way a hidden representation u_i (Formula 2) of h_i. Later, we measure the word n-gram importance as the similarity of u_i with the word n-gram level context vector u_h and get a normalized importance weight α_i (Formula 3) through a softmax function. After that, the instance vector v (Formula 4) is computed as a weighted sum of the word n-gram annotation h_i and its importance α_i; the word n-gram context vector u_h is randomly initialized and jointly learned during the training process.

Third stage: it performs the concatenation of the output vector of each channel (instance vector v), forming a new vector z that contains a high level representation of the different word n-grams; this vector is used as input for the instance classification.

Fourth stage: it includes the classification layers; three layers handle the final classification, a dense layer, followed by a dropout layer and a fully-connected softmax layer to obtain the class probabilities and get the final classification.

4 Experimental Settings

This section presets the experimental settings. First, it introduces the four used datasets, which correspond to Twitter and Facebook collections. Then, with the purpose of facilitating the replicability of our results, it shows the implementation details of the proposed attention-based deep neural network.

4.1 Datasets for Abusive Language Identification

Abusive language can be of different types, depending to the target and severity of the insults. Accordingly, different collections have used for its study. Below we present a brief description of the four datasets we used in our experiments. From now on we will refer to them as DS1, DS2, DS3, and DS4.

DS1 [16] and DS2 [6] were some of the first large-scale datasets for abusive tweet detection; DS1 focuses on the identification of racist and sexist tweets, whereas DS2 focuses on identifying tweets with abusive language and hate speech. On the other hand, DS3 [18] and DS4 [9] were used in the SemEval-2019 Task 6, and in the First Workshop on Trolling, Aggression and Cyberbullying respectively. DS3 focuses on identifying offensive tweets, whereas DS4 focuses on identifying Overtly Aggressive (OAG) and Covertly Aggressive (CAG) Facebook posts and comments.

Table 1 resumes information about the classes distribution of the four collections. It is important to notice their high imbalance, which indeed correspond to their real-life occurrence.

Table 1. The classes distribution of the four used datasets.

Dataset	Classes distribution			Total
DS1	Racist	Sexist	Neither	16,914
	1,972	3,383	11,559	
DS2	Offensive	Hate	None	25,112
	19,326	1,428	4,288	
DS3	Offensive	Non-offensive		13,240
	4,400	8,840		
DS4	OAG	CAG	NAG	12,000
	2,708	4,240	5,052	

4.2 Implementation Details

Different text preprocessing operations were applied: user mentions and links were replaced by default words; hashtags were segmented by words in order to enrich vocabulary (e.g. #BuildTheWall - build the wall); all emojis were converted into words; stop words were removed, except personal pronouns, which

have been recognized as useful for this task; all text were lowercased and non-alphabetical characters were removed. On the other hand, for word representation we used pre-trained fastText embeddings [3], trained with subword information on Common Crawl.

Table 2. Proposed attention-based deep neural network hyperparameters.

Layer	Settings			
Input	Sequence length		75	
Embedding	Word dimensions		300	
Convolutional	Kernel sizes	{1, 2, 3}	Filters	{256, 256, 256}
Bi-GRU	Units	{75, 74, 73}	Dropout rate	20%
Attention	Neurons		{75, 74, 73}	
Concatenation	Vector size		222	
Dense	Neurons	128	Activation function	relu
Dropout	Rate		20%	
Dense	Neurons	# Classes	Activation function	softmax

Table 2 presents the hyperparameter settings of our proposed NN architecture. The network was trained with a learning rate of 0.001, using Adam optimizer, and a total of 10 epochs. In order to compare the robustness of our proposal, we consider two baseline architectures: a simple GRU network, which receives words as input but does not use word n-grams nor attention, and a second architecture that employs the same GRU network but including one attention layer. These two baselines architectures and our proposed architecture are referred in the experiments as GRU, GRU+ATT, and CNN-GRU+ATT respectively. It is important to mention that both baseline architectures used the same hyperparameter settings, skipping the attention, concatenation, and convolutional layers, respectively.

5 Results

This section is organized in two subsections. Section 5.1 presents the quantitative results of the experiment; it compares the proposed architecture and baseline approaches with state-of-the-art results. Section 5.2 describes the analysis of the results using the attention word sequences visualization, and presents as qualitative results a list of inherently and context-dependent offensive word sequences.

5.1 Effectiveness of the Proposed Architecture

Table 3 shows the results of the proposed NN architecture (CNN-GRU+ATT) as well as two baselines results obtained by the GRU and GRU+ATT simplified

architectures. For sake of comparison, we used two different evaluation measures commonly used in the abusive language identification task; for DS1, DS2, and DS3 the macro-average F_1 score, and for DS4 the weighted macro-average F_1 score. The results indicate that the use of the *contextual attention* outperformed the base GRU network (column 3 vs column 2) by at least a margin of 3%. In addition, the use of *attention over word n-grams* outperformed the use of word attention (columns 4 vs columns 3) by at least a margin of 2%. We compared GRU+ATT vs GRU and CNN-GRU+ATT vs GRU+ATT with McNemar's statistical test and Student's t-test, obtaining statistically significant values with $p \leq 0.05$ and $p \leq 0.01$ respectively.

Table 3 also compares the results from our proposed architecture (CNN-GRU+ATT) and state-of-the-art. It shows that the CNN-GRU+ATT neural network obtained better results in 3 out of 4 datasets, and, therefore, it allows concluding that the use of attention over word n-grams is useful for discriminating between offensive and non-offensive contexts. It is important to note that the result from [1] in DS4 only improved our results by margin of 1%. That work explores techniques of data augmentation and proposes a deep neural network trained on pseudo labeled examples. Despite of its better results, it lacks of interpretability, a key aspect of the current proposal.

Table 3. Comparison results from GRU, GRU+ATT, CNN-GRU+ATT and state-of-the-art methods in four datasets for abusive language identification (*for DS1, DS2 and DS3 the macro-average F_1 was used, and for DS4 the weighted macro-average F_1*)

Dataset	GRU	GRU+ATT	CNN-GRU+ATT	State-of-the-art
DS1	0.76 ± 0.0078	0.81 ± 0.0067	**0.83 ± 00.0066**	0.82 [19]
DS2	0.74 ± 0.0081	0.77 ± 0.0086	**0.79 ± 00.0072**	0.77 [11]
DS3	0.75 ± 0.0062	0.79 ± 0.0059	**0.84 ± 00.0083**	0.83 [10]
DS4	0.58 ± 0.0078	0.61 ± 0.0081	0.63 ± 0.0078	**0.64** [1]

5.2 Inherently and Context-Dependent Offensive Word Sequences

One of the major advantages of attention mechanisms is the interpretability of decisions. As part of this interpretability, we present the extraction and visualization of inherently and context-dependent offensive word sequences.

To extract the importance of the word sequences in accordance to their context, we limit their analysis to an instance level; to that end, the input text is evaluated by the model, and the normalized importance weights for all word n-grams (α_i) are computed. At this step, we process each channel (unigrams, bigrams and trigrams) separately. Then, for visualization, we associate a color intensity to each weight value, the greater the weight (α_i) of a word n-gram, the greater its color intensity. In the case of n-grams greater than one, some words may be contained in several sequences, causing a problem of visualization, to solve it we decide only displaying the word sequences with the greatest weights.

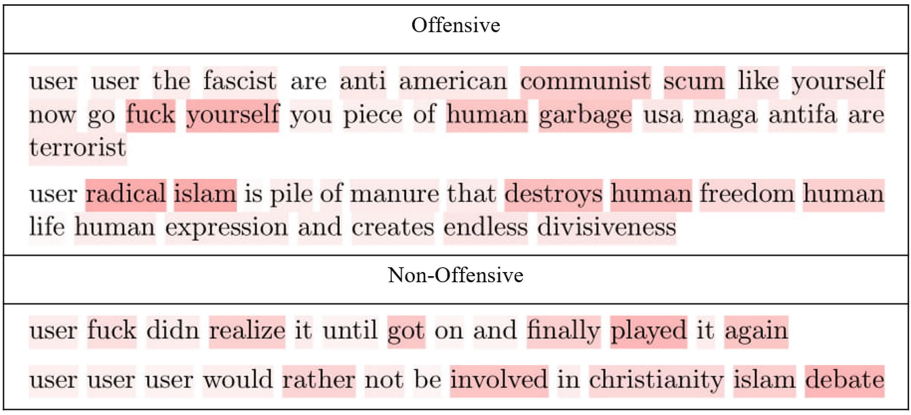

Fig. 2. Attention visualization for offensive and non-offensive texts from DS1 and DS2.

Figure 2 presents some examples of posts, containing words such as `fuck` and `islam`, which correspond to offensive and non-offensive posts. The produced visualization is able to show that the interpretation of words is context dependent, therefore, the presence of swear words such as fuck, or words commonly used in racist speech such as islam, not necessary indicates an aggression; it is the presence of word sequences like `radical islam` and `go fuck yourself` that provide a better way to explain offensive instances.

Table 4. Examples of inherently and context-dependent offensive word sequences.

	Unigrams	Bigrams	Trigrams
DS1 inherently offensive	dick, bitch, nigga, dumb, idiot	fuck you, woman cant, fuck off, you idiot, stupid woman	radical islam on, sexist but fuck, fuck off my, sluts and cunt, fuck her like
DS1 context dependent	islam, cooking, fuck, black, sucks	they deserved, like islam, woman prefer, islam is, even woman	user islam is, a woman wants, the jews that, say about woman, the muslim immigrants
DS3 inherently offensive	asshole, idiot, nigga, nigger, bitch	evil nazi, fuck you, the nigga, doggie girlfriend, stupid black	go fuck yourself, an ugly black, shut the fuck, buying pussy bitch, she is shit
DS3 context dependent	white, fool, black, fuck, fucking	immigrant children, user fuck, who fucked, this shit, all woman	what the fuck, user oh shit, people like you, blaming woman not, person of color

With the intention of moving one step further in the understanding of aggressive speech, we extracted a set of inherently and context-dependent offensive expressions based on their computed weights. Basically, we define a word n-gram as inherently offensive if it shows high attention values with a small standard deviation, that is, if its presence was always important for the network to discriminate offensive from non-offensive messages. In contrast, context-dependent offensive n-grams are those with the greatest standard deviations, suggesting that their occurrences not always were important for the network decisions. Considering these criteria, we extracted 20 word sequences for each channel and dataset. Table 4 presents five examples of each type from the DS1 (racism and sexism) and DS3 (general offenses) collections. In spite of their clear differences due to the type of abusive language, they show interesting coincidences. For example, swear words related to low intelligence or sex tend to be inherently offensive (e.g., `idiot`, `dump`, `bitch` and `dick`), and, on the other hand, colloquially words and expressions such as `fucking`, `black`, `what the fuck`, and `this shit` can be used in both contexts.

6 Conclusions and Future Work

One of the main problems in abusive language identification is to distinguish between the use of swear words and vulgarities in offensive and non-offensive contexts. To tackle this issue we proposed an attention-based neural network architecture that captures the importance of word n-grams according to their context. Through the use of this architecture, we were able to extract and visualize inherently and context-dependent offensive word sequences. The results obtained in four collections, considering different kinds of aggressive speech, were encouraging, they improved state-of-the-art results in 3 out of 4 datasets, and, therefore, allowed concluding that the use of attention over word n-grams is useful for discriminating between offensive and non-offensive contexts.

As future work we plan to explore the combination of general and specific domain word vectors, with the intention of obtaining a higher quality input text representation. In addition, we consider the use of the inherently and context-dependent offensive word sequences as search keywords, to bootstrap a new abusive language dataset. Finally, we consider the application of the proposed architecture in other tasks where the interpretation of word sequences is highly context dependent such as the detection of deception or the detection of depressed social media users.

Acknowledgements. We thank CONACyT-Mexico for partially supporting this work under project grant CB-2015-01-257383 and scholarship 925996.

References

1. Aroyehun, T., Gelbukh, A.: Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying TRAC - 1 (2018)

2. 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 (2017)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
4. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci.EPJ Data Sci. **5**(1), 11 (2016). https://doi.org/10.1140/epjds/s13688-016-0072-6
5. Chakrabarty, T., Gupta, K., Muresan, S.: Pay "Attention" to your context when classifying abusive language. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 70–79 (2019)
6. 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)
7. Gambäck, B., Sikdar, U.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90. Association for Computational Linguistics (2017)
8. Gaydhani, A., Doma, V., Kendre, S., Bhagwat, L.: Detecting hate speech and offensive language on Twitter using machine learning: an N-gram and TFIDF based approach. In: IEEE International Advance Computing Conference (2018)
9. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of Hindi-English code-mixed data. In: Nicoletta Calzolari Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association ELRA (2018)
10. Liu, P., Li, W., Zou, L.: NULI at SemEval-2019 Task 6: transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of the 13th International Workshop on Semantic Evaluation SemEval (2019)
11. MacAvaney, S., Yao, H., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS One **14**, e0221152 (2019). https://doi.org/10.1371/0221152
12. Nikolov, A., Radivchev, V.: Nikolov-Radivchev at SemEval-2019 task 6: offensive tweet classification with BERT and ensembles. In: Proceedings of the 13th International Workshop on Semantic Evaluation SemEval (2019)
13. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, pp. 145–153 (2016)
14. Pavlopoulos J., Malakasiotis, P., Androutsopoulos, I.: Deeper attention to abusive user content moderation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017)
15. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: International Workshop on Natural Language Processing for Social Media, pp. 1–10 (2017)
16. 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)
17. 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)

18. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 task 6: identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the International Workshop on Semantic Evaluation SemEval (2019)
19. Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.): ESWC 2018. LNCS, vol. 10843. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4