Journal of Data Intelligence, Vol. 2, No. 4 (2021) 418–433

TOWARDS ATTENTION-BASED CONTEXT-BOOSTED CYBERBULLYING DETECTION IN SOCIAL MEDIA

NABI REZVANI
Computing Department, Macquarie University
Sydney, Australia
nabiallah.rezvani@hdr.mq.edu.au

AMIN BEHESHTI
Computing Department, Macquarie University
Sydney, Australia
amin.beheshti@mq.edu.au

Cyberbullying detection is a rising research topic due to its paramount impact on social media users, especially youngsters and adolescents. While there has been an enormous amount of progress in utilising efficient machine learning and NLP techniques for tackling this task, recent methods have not fully addressed contextualizing the textual content to the highest possible extent. The textual content of social media posts and comments is normally long, noisy and mixed with lots of irrelevant tokens and characters, and therefore utilizing an attention-based approach that can focus on more relevant parts of the text can be quite pertinent. Moreover, social media information is normally multi-modal in nature and may contain various metadata and contextual information that can contribute to enhancing the Cyberbullying prediction system. In this research, we propose a novel machine learning method that, (i) fine tunes a variant of BERT, a deep attention-based language model, which is capable of detecting patterns in long and noisy bodies of text; (ii) extracts contextual information from multiple sources including meta-data information, images and even external knowledge sources and uses these features to complement the learner model; and (iii) efficiently combines textual and contextual features using boosting and a wide-and-deep architecture. We compare our proposed method with state-of-the-art methods and highlight how our approach significantly out-performing the quality of results compared to those methods in most cases.

Keywords: Cyberbullying, contextualization, deep learning, attention-based models

1. Introduction

Cyberbullying has been an ongoing and ever growing problem throughout the course of last few years, therefore automatic detection of it has been an appealing topic for researchers as well. Cyberbullying or hate speech is the act of posting hateful content on social media which is found abusive due to its repetitive and, abusive nature [35]. Many research works have addressed this task, most of which have focused on the textual content only. These works have utilized Natural Language Processing (NLP) [35, 8] best practices to analyze, summarize, and detect occurrences of abusive content in social media. Starting from very primitive approaches (such as Bag of Words [43] analysis or TF-IDF [45] for textual feature extraction) to smarter methods like character n-gram and token n-gram [19]. These methods that address only textual features on the surface are shown to yield fairly good results, yet they have left huge room for improvement.

Some researches have taken a more complicated approach by looking at higher-level semantic features. There are a number of interesting works that have adopted a more conceptual approach; some works have employed word generalization techniques to tackle the inherent sparsity of textual bodies via applying clustering methods [42] or word embedding techniques [29]. Besides, there are other works that have leveraged sentiment analysis [21, 16], utilized lexical resources, by making use of profane terms list for instance [9, 31, 10], or have employed linguistic features such as modeling word and sentence relationships [9, 21]. Another line of work utilized sequence based neural network models [6, 5] like long short-term memory (LSTM) and Recurrent Neural Networks (RNNs) for text classification [47] which have either trained these learner models from scratch or have fine-tuned existing pre-trained models using transfer learning. There have even been works that have considered using attention-based models for the task [40, 1]. [30, 32] have applied BERT on this domain and have reported promising results. Although they have not included any contextual features beyond the textual corpus.

While all these methods have been able to improve the accuracy of the Cyberbullying task to some extent, we believe that they have not fully utilized the insightful information beyond textual data, what we call context in the scope of this paper. The first category of context, is internal context which includes any non-textual information that could be extracted from the dataset itself, e.g., user's liked posts, followers, or multimedia content linked to the social media post, e.g., images and videos [20]. User metadata features are leveraged along with textual features in some research works [2]. Utilizing visual features in images and videos has also been addressed in some researches and is shown to have yielded impressive results [25, 44, 26]. Even combining textual and visual features have been carried out in a few related works and is shown to improve the accuracy of the Cyberbullying task [25, 44, 26, 22]. There is also another insightful aspect of context reflected in the external information sources such as social media trends and hot news topics which we call external context. Exploiting the external context has been performed by a limited number of research works, mostly by building external rule engines and Knowledge Bases (KBs) [16, 4, 19, 3].

In our previous work [34], we have put the first step towards combining textual and contextual information. Although results were quite promising, the language model was not capable of handling long text instances, also the combination model was a bit simplistic. In this paper, we extend our previous work and present an attention-based context-boosted cyberbullying detection approach to alleviate our previous work's shortcomings. We believe that all these researches lack a well defined and efficient approach for feature engineering that is capable of combining textual and contextual features in order to maximize the utilization of all the insight dimensions for performing the Cyberbullying detection task in the most efficient way possible. In this paper, we propose a method that is capable of not only utilizing a state of the art text classification technique, but also enriches textual features by adding contextual features including features extracted from images, social network metadata and even external knowledge bases. The unique contributions of this paper are:

- Proposing an attention-based method for text analysis so that it can maximize context utilization within the textual corpus of our datasets.

- Engineering contextual features by performing metadata analysis and extracting pat-

terns structural and behavioral information beyond text.

- Proposing a wide-and-deep architecture that allow combining textual and contextual information for performing Cyberbullying detection more efficiently.

The rest of the paper is organized as follows: we discuss the related works in Section 2 and follow that by a formal problem statement section. The proposed methodology and our experiments are presented in Sections 4 and Section 5 before we conclude the paper in Section 6.

2. Background and Related Work

2.1. *Conventional Text Analysis Methods*

There is a large number of research works around abusive behavior detection which mainly focus on the text content posted by users (and also text reactions to those posts) and analyze the text content using NLP methods [27, 11]. A comprehensive survey has covered almost all aspects of abusive behavior detection in social networks [35]. For more details on text-focused methods please refer to Section 2.5.

2.2. *Multi-modal Approaches*

Apart from analyzing the textual content of social media content, researchers have devoted efforts to take into account complementary and meta-information around the textual body. Some research has shown attention towards analyzing meta-features that could be extracted from the social content, such as user connection graphs [23], likes, history of activities, and more [2]. There are a number of research works that are mainly concentrated on images and visual features of the social content units. While some have used crowdsourcing techniques for image labeling [25], some works have taken advantage of image recognition schemes such as Support-Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) for classification and prediction [44, 26, 22].

There are quite a few researches conducted around combining textual, metadata, and visual information of content units [46]. Some methods have even combined textual, visual, and audio features of social media content for inappropriate content detection [36]. Some works have proposed a well defined mathematical foundation for combining features from different modalities. [13] have proposed a network representation learning approach along with a feature clustering method. Although quite novel, this method is not intuitive and can be hard to extend as well.

2.3. *Deep Learning and Attention-based Methods*

Deep learning models, due to their impressive results among different use cases and domains, have drawn researchers' attention and Cyberbbullying has not been an exception.

2.3.1. *Conventional and Memory-based Models*

[1] have experimented with a few deep architectures for the Cyberbullying detection task. Not only they have tried more conventional deep architectures like CNNs, but they also have

performed experiments on memory-based models like LSTM and bidirectional-LSTM. They have mainly followed a transfer learning approach to fine-tune existing language models for their specific classification task.

2.3.2. *Attention-based Models*

Attention-based methods are more modern approaches towards text analysis and natural language processing. These models employ the attention approach to enhance the ability of the model to detect longer dependencies in the inherent sequential structure of text [40]. [1] have experimented with attention-based models for Cyberbullying detection and compared it to conventional sequence models. [30, 32] have applied BERT, a very modern attention-based model [15], to the Cyberbullying domain.

2.3.3. *Hybrid Architectures*

Although deep learning-based methods and especially attention-based methods have shown very promising results in the context of Cyberbullying detection, but researchers have rarely combined results of textual deep models with non-textual ones. [22] have combined results of two deep models trained on text and images for hate speech detection. In our previous work we have combined hand crafted contextual features extracted from multiple sources, with a fine-tuned LSTM model to build a multi-modal classifier[34]. This paper is an extension to that work and aims to employ a BERT language model which is capable of handling longer text sequences due to its attention-based architecture. There are few works that have addressed combining textual features of an attention-based method (e.g. BERT) with non-textual features. [41] have used used novel architectures like Wide and Deep Networks [12] to combine these features together. This work inspires our proposed architecture which will be discussed in the following sections.

2.4. *Datasets And Data Gathering*

There are many different techniques that have been used for preparing datasets that address the task of Cyberbullying. Cyberbullying and hate speech detection datasets are mostly focused on text [19, 31], but there are limited number of related works that have included metadata and multimedia features as well [25, 22]. A very common approach among most of the research works around Cyberbullying data gathering, is employing crowdsourcing for labeling instances in the training and test set [25, 19]. In general, the process of gathering data for the task of Cyberbullying detection and then properly labeling them for experimental usages is quite cumbersome and requires a lot of effort [17].

2.5. *Natural Language-based Techniques for Cyberbullying Detection*

Although only focused on textual content, these works, have looked at the problem from different perspectives and by analyzing a variety of features. In the following, We will briefly touch on different NLP based techniques used for the Cyberbullying detection task.

Basic Features. Although classic approaches such as Bag of Words analysis [43] or TF-IDF [45] have been around for quite a long time, they are proven to yield fairly good results, as just baseline methods. There are also slightly Smarter methods like character n-gram and

token n-gram [19], which are proven to improve those frequency-based methods.

Generalization Techniques. Moving on from the basic features, there are a number of interesting works that have utilized word generalization techniques to tackle the inherent sparsity of textual bodies, these mainly fall into two categories of clustering methods [42] and word embedding methods [29].

Sentiment Analysis. Negative sentiment and hate speech are proven to be correlated [35]. Researchers that have adopted sentiment analysis, have either performed a multi-step process to first featurize and then classify textual content [16] or have directly performed classification based on a "sentiment as a feature" approach [21].

Lexical Resources. Based on many research works, there is a clear association between a specific list of words and the existence of profane content in textual bodies [35]. This leads to the idea of exploiting lexical resources for reinforcing Cyberbullying detection. The idea is using a dictionary of related words [9, 31] which might be adjusted based on different contexts [10]. Lexical resources are normally used as complementary methods for other analysis and detection techniques.

Linguistic Features The idea behind these sets of works is modeling the deeper and higher-level semantic relationships between words in sentences. This can be achieved by manually building dependency rule engines [21] or using statistical approaches [9]. There are also two main categories of works that have either looked at the relationships from a generic perspective or have employed relationships tailored to the problem.

3. Problem Statement

Given $C = \{S_1, S_2, ..., S_N\}$ the corpus of social media sessions, each session instance (also referred to as a post) has an image, a main image description and a number of comments. The comments for each session instance are denoted as $C_i = \{C_i^1, ..., C_i^{c_i}\}$ where c_i is the number of comments for the ith session instance. Other than the session instance image, main description and comments, each session has some contextual features as well such as number of likes, number of shares, owner user, number of follower and followees of the owner. Moreover, each post and comment have a timestmap. There is a label assigned to each post in the training set indicating if that session instance is a Cyberbullying case or not, e.g. $L = \{L_1, ..., L_N\}$ where $L_i = \{0, 1\}$. Eventually the task is to predict the label for the test set based on all the features that are available in the dataset.

4. Proposed Method

As discussed earlier, our proposed method aims to maximize context utilization by firstly employing an attention-based language model, and secondly leveraging non-textual information available from other modalities. We eventually propose a novel architecture for combining textual and contextual features.

4.1. *BERT Attention-based Model*

BERT is a modern sequence based deep learning model that has been able to outperform its predecessors significantly, due to its attention-based nature [15, 40]. As part of our proposed method, we perform fine-tuning on a BERT-based language model. Leveraging a pre-trained

language model which has is already trained with many language patterns, will allow us to better detect complicated instances of Cyberbullying. Besides, BERT is an attention-based model and will be able to detect patters that occur in a long sequence of text, which is normally the case in a relatively long thread of text in social media.

4.2. *Leveraging context from Non-textual Sources*

As discussed earlier in the related works section, one of the major shortcomings of the previous research around Cyberbullying detection is merely focusing on the textual content. Although many research works have attempted to maximize context utilization within the textual corpus, they have rarely taken into account invaluable context derived from sources beyond the textual corpus. We propose the usage of three very important knowledge sources that could contribute towards further contextualization.

4.2.1. *Metadata Features*

Firstly we extract features from the metadata that exist around the social content. We believe that there is a lot of insight that could be attained from the structural and behavioral information of the users. To be more specific, we extract the following features from the metadata that we have available:

- *Number of followers*: The total number of people that have followed the owner of the post.
- *Number of followees*: The total number of people that are followed by the owner of the post.
- *Number of likes*: The number of posts a user has liked in total.
- *Popular categories*: Main categories of posts that a user has reacted to. We will extract three features that are indicative of the three most popular categories for a user. That is done by first clustering all posts and then labeling them with the cluster label.
- *Average reactions*: Average number of times a user has reacted to a certain post.
- *Average replies*: Average number of times a user has replied to a certain user, within a post thread.
- *Frequent mentions*: Maximum number of times a user posting or commenting has mentioned another certain user. We have calculated the total mentions in the post and comments as a Cyberbullying case might have existed in the original post or in the comments posted by other users.

4.2.2. *Image Features*

There has been very little research dedicated to utilizing the visual features that could potentially be extracted from the image data that is posted along with text in social media. We propose the usage of visual features extracting labels from the images linked to the posts. We use the pre-trained ImageNet model [14] which is capable of labeling any given image with certain confidence levels. We pick the highest confidence labels for each image, based on a threshold, then build a vocabulary of image-extracted features. Each image then will yield a Boolean occurrence vector with labels in each element of the vector.

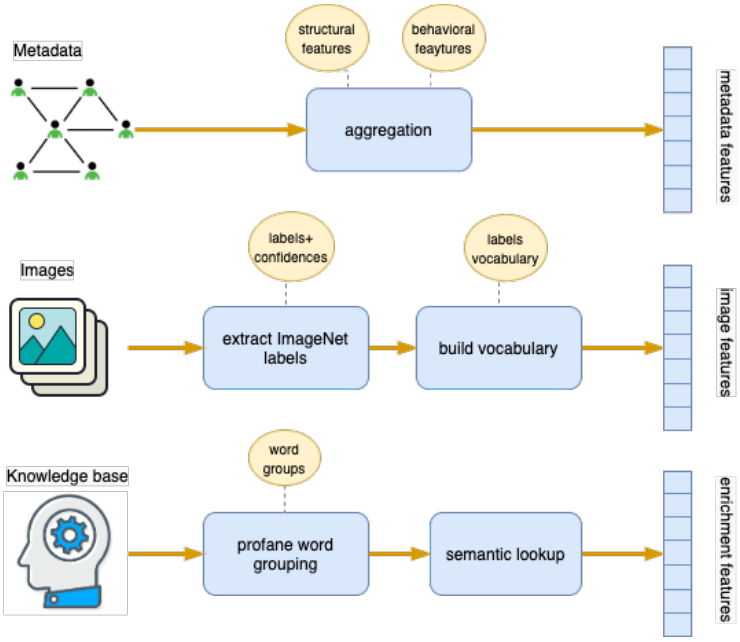


Fig. 1. Illustration of the feature extraction procedure for metadata, image and enrichment features.

4.2.3. *Enrichment Features*

So far in our proposed method, we have discussed exploiting textual, metadata, and visual data for contextualization. We believe there is still one feature-set that is ignored by most researchers, while it has a prominent effect in contextualizing the data. We propose the usage of a knowledge base along with all the previously mentioned features. We use Google's standard profanity word list [18] which consists of 451 words marked as profane by Google. We perform a pre-processing on this list to yield a much smaller list of word groups. We then perform a simple lookup through our dataset and mark the appearance of those profane groups in each of the training samples. This eventually yields another Boolean occurrence vector which we call enrichment features. Figure 1 illustrates the feature extraction procedure.

4.3. *Feature Combination Architecture*

4.3.1. *Wide-and-deep architectures for combining features*

As the main feature combination method, we have also proposed a wide-and-deep neural network architecture for combining textual and contextual features [12]. The advantage of this approach is the join training capability that it offers which allows training parameters of both textual and contextual networks at the same time rather than merely combining their classification results. The wide and deep model consists of two components. The deep component (dense features), contains textual information attained from the BERT model. The wide component (sparse features), are the concatenation of contextual features and their

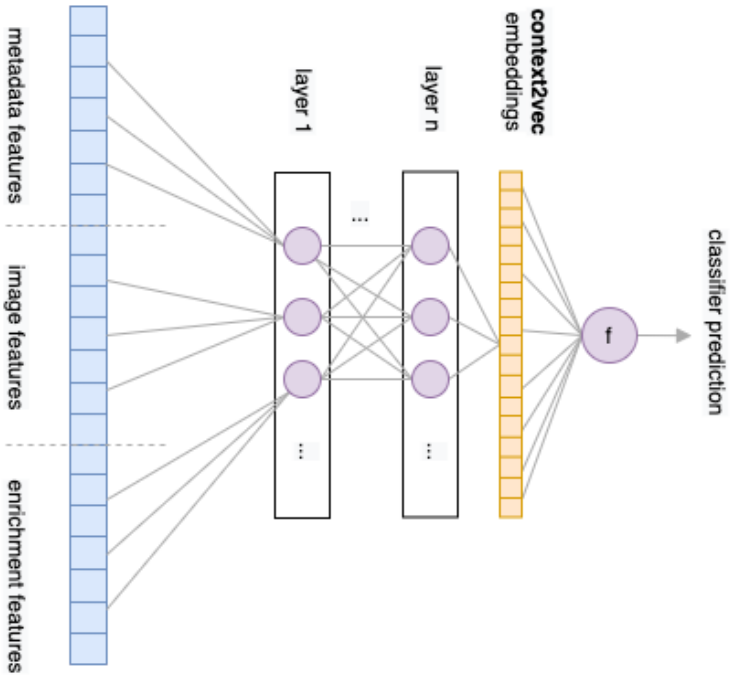


Fig. 2. The neural network model that builds the context2vec embeddings.

cross-product transformations, where cross-product transformations are calculated as below:

$$\phi_k(x) = \Pi_{i=1}^{d} x_i^{c_{ki}}, c_{ki} \in \{0, 1\}$$

Where ϕ_k is the k-th transformation calculated on contextual features. If the i-th feature is part of the k-th transformation, then c_{ki} is 1, otherwise it is 0. Assuming that all the features are boolean, the result of a transformation is 1 if all the features in it are 1 and 0 otherwise. As an example, if a transformation is $(numberOfLikes > 10 \wedge numberOfFollowers > 100)$, then the value of transformation is true if both those conditions are true.

Since our metadata features are not boolean (rather they are range features), we need to bucketize them to gain a number of boolean features. For image and enrichment features, we have boolean vectors that can be used for cross-product transformations. If $n_{context} = n(metaBuck) + len(imageVect) + len(enrichmentVect)$ is the total number of contextual features, we can have $(n_{context})!$ transformations out of which we randomly pick $m * n_{context}$ transformations (m is chosen heuristically). Finally, given that we are using a linear regression model for the wide component, the joint loss function that is being minimized, will be the following:

$$P(Y = 1|x_t, x_c) = \sigma(w_{wide}^T[x_c, \phi_{x_c}] + w_{deep}^T x_t + b)$$

Where w_{wide}^T are the weights for the wide component, $[x_c, \phi_{x_c}]$ are the concatenation of the contextual features and their cross-products, w_{deep}^T are the weights for the deep component and x_t is the textual features vector. $\sigma(.)$ is the sigmoid function and b is the bias.

Figure 3 illustrates the our proposed wide-and-deep architecture.

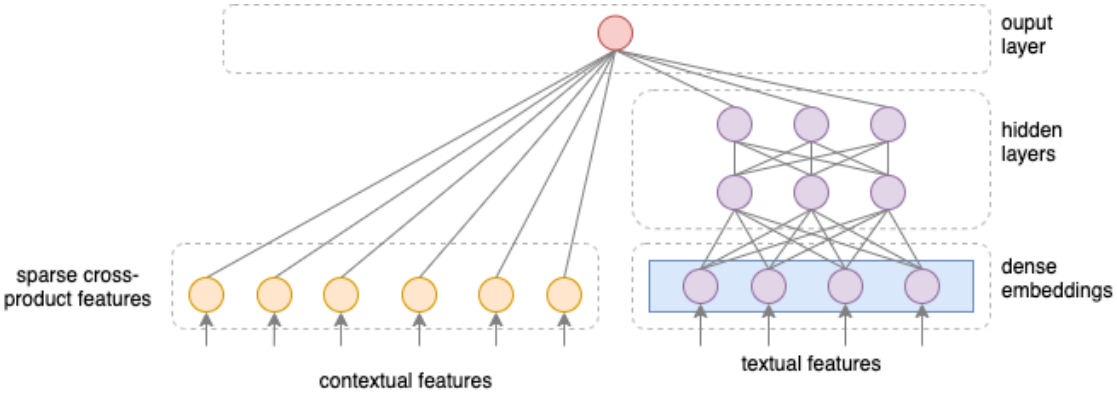


Fig. 3. Wide-and-deep architecture for combining textual and contextual features

4.3.2. *Boosting architecture for combining features*

The second approach is based upon an ensemble learning model, where we train two independent classifiers, then combine their results using a model averaging scheme. In this research we use a boosting model with soft voting [33] to combine the result of our textul and contextual classifiers. The following shows the equation for choosing the class label.

$$s_i = \sum_{j \in C} w_i \cdot p_{ij}$$

Where s_i is the score of label i, C is the set of all the labels, w_j is the weight of classifier j and p_{ij} is the probability generated by classifier j for class i. Eventually, s_i will yield the best class label i.

Figure 4 illustrates the boosting architecture for combining the two classifiers.

5. Experiments

In this section, we present the result of our experiments and the comparison of our proposed method implementation with state of the art methods.

5.1. *Dataset*

In order to run our experiments, we had to find datasets that contain social network posts, with text, metadata, and images which were labeled and prepared for the Cyberbullying task. We have chosen two main datasets to experiment with.

5.1.1. *Instagram Dataset*

The Instagram dataset [24], contains 2188 posts from Instagram along with images and comments, with metadata information including likes and friendship graph. The dataset contains two labels, Cyberbullying and Cyber-aggression which within the scope of this paper, we

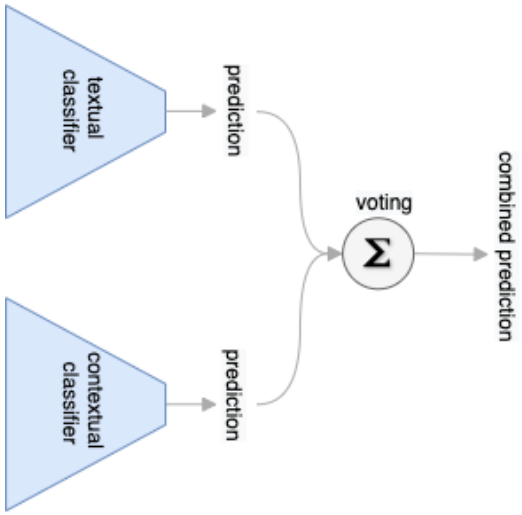


Fig. 4. Boosting architecture for combining textual and contextual features

focused on Cyberbullying only. Of all the samples in the dataset, 1540 were labeled as negative, and 678 were positive meaning that Cyberbullying case has been detected in them. The images were already tagged with labels extracted from a pre-trained ImageNet model. The textual content in the dataset also includes mentions and hashtags which we used to build the metadata features as explained earlier.

5.1.2. *Twitter Dataset*

The second dataset used for our experiments is a set of 7321 Tweets from Twitter, curated, and labeled for the cyberbullying detection task [37]. The dataset only contains the Tweet ID, along with a few labels. Using the Twitter API [39] we downloaded Tweet texts, comments, friendship graphs, likes, and re-Tweets. The Tweet text and its comments also included hashtags and mentions that we extracted as part of our pre-processing and used them for building metadata features. We realized that the mentions on this dataset were very sparse and decided not to include them in our metadata features for this dataset.

5.2. *Implementation Setup*

As mentioned earlier, we have used a pre-trained BERT language model as our attention-based text classifier model. We used Tensorflow[a]for transfer learning of the BERT model and also combining the two models using thew wide-and-deep model.

5.2.1. *Training the Textual Classifier*

We used base version of ALBERT [28], a faster and more memory efficient implementation of BERT, as the pre-trained language model. The pre-processing step is also carried out using a Tensorflow layer. We set the vocabulary size to 1024 (tokenization, lemmatization, removing

[a]https://protect-au.mimecast.com/s/zQc3Ck815RCWjGKVTQjWvq?domain=tensorflow.org

stop words and special characters is done as part of preprocessing). After preprocessing, an encoder and dropout layer are added. Finally, a dense layer is added for performing classification.

5.2.2. *Training the Contextual Classifier*

As mentioned before, we have prepared three sets of features as our contextual features. We created 10 metadata features, 20 image features, and 20 enrichment features which in total make up a vector of size 50. We set m=100 and randomly generate 5000 transformations (100 times the number of features) and generate cross-product features. We concatenate original and cross product features and run them through a dense neural network with one hidden layer. The resultant classifier was not as powerful as the BERT classifier but could reach the accuracy of roughly 0.65 which was not competitive as an independent classifier, rather it was used in our boosting set up along with the BERT model. Moreover, we used these contextual features in the joint-train setting of the wide-and-deep classifier.

5.2.3. *Combining Textual and Contextual Models*

Firstly, we combine our two classifiers first using a wide-and-deep combination model. We perform fine tuning of the ALBERT model in the first stage, then will joint-train both textual and contextual models using a wide-and-deep model. The joint-training, trains the wide regression model parameters along with the deep parameters of ALBERT. Secondly, our boosting combination technique combines the result of two independent classifiers using a weighted voting scheme. We chose 0.6 as the textual classifier weight and 0.4 as the contextual classifier weight as our textual classifier was demonstrating better results independently. We used the probability results of the last layer of both classifiers without thresholding to be able to perform soft voting.

5.3. ***Results***

We have compared our two proposed models with four baseline methods. The first two models both use a conventional Neural Network (NN) architecture as the learner model and the number of hidden layers and the number of nodes in each layer is optimized using hyperparameter tuning. The difference between the two models is in the features they use, where the first one uses TF-IDF features and the second uses Word2vec features. We have performed tokenization, lemmatization, stop-word, and punctuation mark removal during the pre-processing of these two models. Th third model is the contextualized LSTM methods that we proposed in our previous work[34]. The fourth model is the text-only ALBERT network undergone the first stage of training. It obviously only focuses on textual features and does take into account any contextual features. The last two models are the proposed models that combine textual and contextual features. We have calculated accuracy, precision, recall, and f-score for the five models that we have compared. Table 1 shows the result of those metrics for the Instagram dataset and Table 2 summarizes the results for the Twitter dataset.

It is clearly observable that both LSTM and BERT based models are significantly outperforming baseline methods on almost all the metrics. This is an expected behaviour as the memory-based capabilities in both models help in detecting patterns in long and noisy

Table 1. Instagram: Result of experiments on Instagram Dataset, comparisons of base NN model, ALBERT only model, and the proposed contextualized ALBERT model.

Method	Accuracy	Precision	Recall	F-score
NN with TF-IDF features	0.77	0.79	0.75	0.76
NN with Word2Vec features	0.78	0.79	0.76	0.77
LSTM + context2vec features (NN)	0.86	0.87	0.83	0.85
ALBERT (text only)	**0.86**	0.86	0.84	0.85
ALBERT + context2vec features (Boosting combiner)	0.86	0.85	0.84	0.84
ALBERT + context2vec features (Wide-and-deep combiner)	0.86	**0.88**	**0.85**	**0.87**

Table 2. Twitter: Result of experiments on Twitter Dataset, comparisons of base NN model, ALBERT only model, and the proposed contextualized ALBERT model.

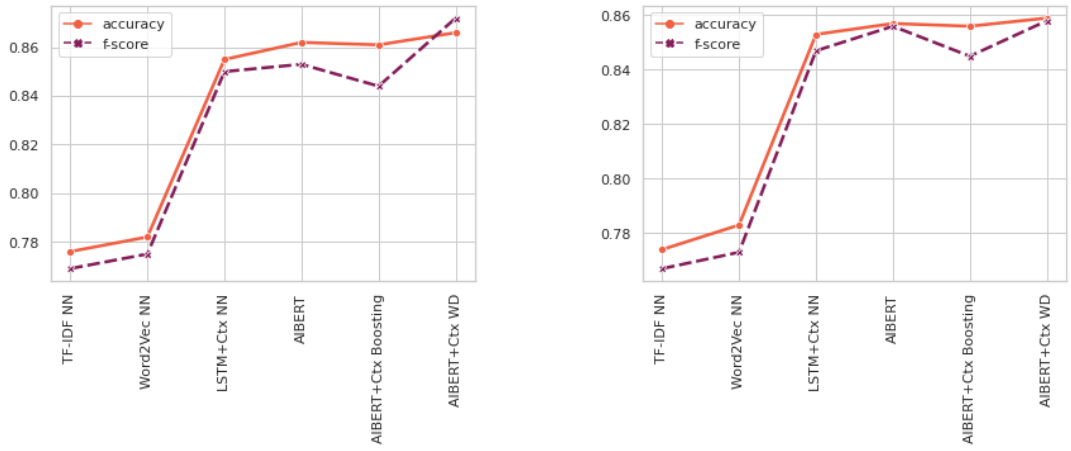
Method	Accuracy	Precision	Recall	F-score
NN with TF-IDF features	0.77	0.79	0.75	0.76
NN with Word2Vec features	0.78	0.79	0.76	0.77
LSTM + context2vec features (NN)	0.85	0.87	0.83	0.85
ALBERT (text only)	**0.85**	0.86	0.84	**0.85**
ALBERT + context2vec features (Boosting combiner)	0.85	0.84	0.85	0.84
ALBERT + context2vec features (Wide-and-deep combiner)	0.85	**0.86**	**0.85**	0.85

sequences. The ALBERT model also beats the LSTM model in majority of metrics for both datasets. This reinforces the hypothesis around an attention-based model being able to focus on parts of the textual corpus that are more relevant to the classification task. The precision of the ALBERT model though, has not been able to match the contextualized LSTM model which explains how much contextual features have been able to contribute to enhancing a weaker text-only classifier. Once contextual features are added to the ALBERT model, the classifier surpasses the LSTM model especially for the Instagram dataset. The wide-and-deep combiner has clearly contributed to boosting the text-only ALBERT model, while the boosting combiner has mostly performed equally and even worse in some cases.

ALBERT (text only) model has not been able to perform as well on the Twitter dataset as the textual content of the posts in it are shorter than the Instagram dataset in average, therefore ALBERT which mainly specializes in detecting context in long sequences, has not been able to perform that well given that limited training data has been available. Contextualization with the boosting combiner has been able to improve the recall but it has had negative impact on the precision and the f-score has remained the same. The Wide-and-deep combiner, on the other hand has been able to slightly improve the recall without affecting the precision and the f-score. This can be explained by not having strong contextual features in the Twitter dataset, so the overall accuracy (demonstrated in f-score) is not improved much while the wide-and deep architecture has been able to prevent the model degradation on precision as well. Figure 5 depicts the comparison of the above-mentioned metrics for baseline methods versus our proposed methods.

6. Conclusion and Future Work

Cyberbullying and its immense impact on social network users especially youngsters, is rapidly growing, therefore detection of Cyberbullying using modern NLP methods and machine learning has become a popular topic among researches. In this paper, we highlighted the importance of Cyberbullying, and did a literature review of how start of the art researches have



(a) (b)

Fig. 5. Accuracy and F-score comparison for different methods (a) Instagram dataset (b) Twitter dataset.

addressed this problem from different perspectives. Social network data is multimodal, noisy and complex in its nature. We hypothesized that most recent research works around Cyberbullying detection lacks a model that is capable of utilizing context within the textual corpus of the social media as well as the contextual information around it including social network metadata, images, etc. We presented a novel Cyberbbullying detection architecture that can model the social media data in an efficient way by factoring in, all the different aspects of the social media data. Our proposed method, employs an attention-based deep learning model which is able to detect patters in social media textual posts and comments even if the textual bodies are long, noisy and irrelevant in most parts. Moreover, we have proposed extracting contextual features from other modalities of the social media data including the social network graph, images and even external knowledge sources. We have eventually proposed methods for combining these multimodal features using modern machine learning architectures.

Our experimental results prove the effectiveness of our proposed method in comparison with baseline and state of the art methods. This reinforces this idea that enriching text by adding contextual information to it, can help in improving the accuracy of Cyberbullying detection task and any classification use case in general within the social media domain. As our future research plans we recommend focusing more on visual features by extracting more meaningful features from images or videos that are posted along with the textual data in social media posts. We plan to leverage storytelling approaches [38, 7] to facilitate interacting with visual features. Further to that, a deep learning model can be trained on the visual modalities rather than merely extracting labels from them, subsequently feeding them to the contextual feature set. The visual features can be combined with the textual and metadata features and the combination of all these latent variables can lead to building a more efficient Cyberbullying detection model.

Acknowledgments

We Acknowledge the AI-enabled Processes (AIP)[b]Research Centre for funding this research.

References

1. Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018.
2. Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.
3. Amin Beheshti, Boualem Benatallah, Quan Z. Sheng, and Francesco Schiliro. Intelligent knowledge lakes: The age of artificial intelligence and big data. In *Web Information Systems Engineering - WISE 2019 Workshop, Demo, and Tutorial, Hong Kong and Macau, China, January 19-22, 2020, Revised Selected Papers*, volume 1155 of *Communications in Computer and Information Science*, pages 24–34. Springer, 2019.
4. Amin Beheshti, Boualem Benatallah, Alireza Tabebordbar, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh, and Reza Nouri. Datasynapse: a social data curation foundry. *Distributed and Parallel Databases*, 37(3):351–384, 2019.
5. Amin Beheshti, Vahid Moraveji Hashemi, and Shahpar Yakhchi. Towards context-aware social behavioral analytics. In *MoMM 2019: The 17th International Conference on Advances in Mobile Computing & Multimedia, Munich, Germany, December 2-4, 2019*, pages 28–35. ACM, 2019.
6. Amin Beheshti, Vahid Moraveji Hashemi, Shahpar Yakhchi, Hamid Reza Motahari-Nezhad, Seyed Mohssen Ghafari, and Jian Yang. personality2vec: Enabling the analysis of behavioral disorders in social networks. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 825–828. ACM, 2020.
7. Amin Beheshti, Alireza Tabebordbar, and Boualem Benatallah. istory: Intelligent storytelling with social data. In *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 253–256. ACM / IW3C2, 2020.
8. Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, Srikumar Venugopal, Seung Hwan Ryu, Hamid Reza Motahari-Nezhad, and Wei Wang. A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing*, 99(4):313–349, 2017.
9. Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.
10. Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11, 2016.
11. Hao Chen, Susan Mckeever, and Sarah Jane Delany. Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems*, pages 187–205. Springer, 2017.
12. Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.
13. Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 339–347, 2019.
14. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,

[b]https://protect-au.mimecast.com/s/SfqkClx1OYUn84VJtqPJEp?domain=aip-research-center.github.io

pages 248–255. Ieee, 2009.
15. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
16. Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30, 2012.
17. Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity. *arXiv preprint arXiv:1910.11922*, 2019.
18. Robert Gabriel. *Google profanity word list*, 2004. https://github.com/RobertJGabriel/Google-profanity-words.
19. Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.
20. Seyed Mohssen Ghafari, Shahpar Yakhchi, Amin Beheshti, and Mehmet A. Orgun. Social context-aware trust prediction: Methods for identifying fake news. In Hakim Hacid, Wojciech Cellary, Hua Wang, Hye-Young Paik, and Rui Zhou, editors, *Web Information Systems Engineering - WISE 2018 - 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part I*, volume 11233 of *Lecture Notes in Computer Science*, pages 161–177. Springer, 2018.
21. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.
22. Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478, 2020.
23. Mohammad Hammoud, Dania Abed Rabbou, Reza Nouri, Seyed-Mehdi-Reza Beheshti, and Sherif Sakr. DREAM: distributed RDF engine with adaptive query planner and minimal communication. *Proc. VLDB Endow.*, 8(6):654–665, 2015.
24. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer, 2015.
25. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*, 2015.
26. Clifford Huang and Mikhail Sushkov. Instanet: Object classification applied to instagram image streams, 2016.
27. Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6, 2014.
28. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
29. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
30. Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2019.
31. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.
32. Sayanta Paul and Sriparna Saha. Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*, pages 1–8, 2020.

33. Gunnar Rätsch, Manfred KK Warmuth, and Karen A Glocer. Boosting algorithms for maximizing the soft margin. In *Advances in neural information processing systems*, pages 1585–1592, 2008.
34. Nabi Rezvani, Amin beheshti, and Alireza Tabebordbar. Linking textual and contextual features for intelligent cyberbullying detection in social media. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, pages 3–10, 2020.
35. Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.
36. Devin Soni and Vivek K Singh. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26, 2018.
37. Junming Sui. *Understanding and fighting bullying with machine learning*. PhD thesis, Ph. D. thesis, Ph. D. dissertation, The Univ. of Wisconsin-Madison, WI, USA, 2015.
38. Alireza Tabebordbar, Amin Beheshti, and Boualem Benatallah. Conceptmap: A conceptual approach for formulating user preferences in large information spaces. In *Web Information Systems Engineering - WISE 2019 - 20th International Conference, Hong Kong, China, November 26-30, 2019, Proceedings*, volume 11881 of *Lecture Notes in Computer Science*, pages 779–794. Springer, 2019.
39. Twitter. *Twitter trends API*. https://developer.twitter.com/en/docs/trends/trends-for-location/api-reference/get-trends-place.
40. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
41. Zhe Wang, Rundong Shi, Shijie Li, and Peng Yan. Gbdt and bert: a hybrid solution for recognizing citation intent. *Studies*, 55:12c2a39230188, 2020.
42. William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics, 2012.
43. Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.
44. Youshan Zhang, Jon-Patrick Allem, Jennifer Beth Unger, and Tess Boley Cruz. Automated identification of hookahs (waterpipes) on instagram: an application in feature extraction using convolutional neural network and support vector machine classification. *Journal of medical Internet research*, 20(11):e10513, 2018.
45. Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, pages 1–6, 2016.
46. Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958, 2016.
47. Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*, 2016.