World Wide Web
https://doi.org/10.1007/s11280-021-00920-4

A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media

Akshi Kumar[1] · **Nitin Sachdeva**[1]

Received: 31 March 2021 / Revised: 20 June 2021 / Accepted: 30 June 2021

Abstract
As a constructive mode of information sharing, collaboration and communication, social media platforms offer users with limitless opportunities. The same hypermedia can be transposed into a synthetic and toxic milieu that provides an anonymous, destructive pedestal for online bullying and harassment. Automatic cyberbullying detection on social media using synthetic or real-world datasets is one of a proverbial natural language processing problem. Analyzing a given text requires capturing the existent semantics, syntactic and spatial relationships. Learning representative features automatically using deep learning models efficiently captures the contextual semantics and word order arrangement to build robust and superlative predictive models. This work puts forward a hybrid model, Bi-GRU-Attention-CapsNet (Bi-GAC), that benefits by learning sequential semantic representations and spatial location information using a Bi-GRU with self-attention followed by CapsNet for cyberbullying detection in the textual content of social media. The proposed Bi-GAC model is evaluated for performance using F1-score and ROC-AUC curve as metrics. The results show a superior performance to the existing techniques on the benchmark Formspring.me and MySpace datasets. In comparison to the conventional models, an improvement of nearly 9% and 3% in F-score is observed for MySpace and Formspring.me dataset respectively.

Keywords CapsNet · Bi-GRU · Cyberbullying · Deep learning · Social media

This article belongs to the Topical Collection: *Special Issue on Synthetic Media on the Web*
Guest Editors: Huimin Lu, Xing Xu, Jože Guna, and Gautam Srivastava

✉ Akshi Kumar
akshikumar@dce.ac.in

Nitin Sachdeva
nits.usit@gmail.com

[1] Department of Computer Science & Engineering, Delhi Technological University, Delhi, India

Published online: 22 July 2021

1 Introduction

Social media platforms like Facebook, Instagram, Twitter and more are giving people an opportunity to connect with each other across distances. Users have become individual information sources and clearly a part of a limitless sharing community that has broken the chains of the ancient information mass media order. The social media dynamics with pervasive internet access, no-restriction posting, hidden agendas and longitudinal-latitudinal reach has appended a polarized, radicalized, toxic and an abusive facet to it. Social media deception poses a serious threat to people and eventually society as disinformation, misinformation and malinformation contaminate the information streams [3, 34]. Misuse of social media to abuse and inflict harm taints this collaborative-communicative Web as an amplifying tool for negative narratives or events. Social media platforms now offer a completely new arena where bullies are assured anonymity and ready access to their targets incessantly with minimal retaliation. Cyberbullying [4] is a form of manipulation, belittlement, and targeted abuse using mean-spirited messages and negative electronic postings. It can be as straightforward as sending mean, hurtful, rude texts or instant messages to as devious as spreading secrets or rumours about people online. Though bullying in electronic form can have multiple-dimensions, such as exclusion, harassment, outing, trickery, cyberstalking, dissing, fraping, masquerading, trolling and flaming [16, 31], the obvious intention to hurt and harm is common. This inveterate nuisance creates mental, emotional and physical risks for the bullied. The targets (victims of cyberbullying) feel overwhelmed, powerless, vulnerable, unsafe, worthless, humiliated, isolated, depressed, embarrassed, vengeful and at times suicidal. An accurate detection can facilitate timely intervention by alarming the moderators to take countermeasures. But content moderation practices on these platforms by human moderators is often inconsistent and done in a non-transparent manner. It also suffers from biasing and may apprehend freedom of expression online. Moreover, spotting bullying instances explicitly as well as spotting victims is tricky too. Blocking and reporting might augment the reality if the bully is within the same professional or personal community, for example, a classmate. Sadly, the scale and impact of cyberbullying can be seen across social media platforms even though its awareness is at an all-time high. Figure 1 shows how social media platforms are hotbeds of cyberbullying activities for young people.

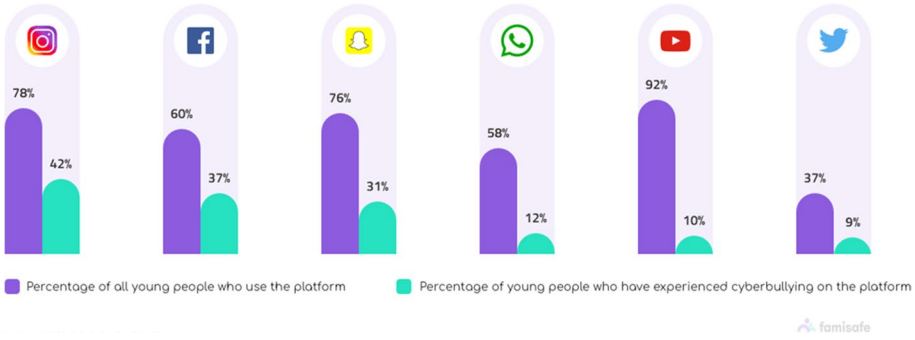

Figure 1 Statistics on share of social media platforms where cyberbullying occurs. https://www.ditchthelabel.org/

With the volume and variety of user-generated content on complex social media platforms the challenges to detect cyberbullying in real-time have amplified. The influx of content makes it challenging to timely regulate online expression. Moreover, the anonymity and context-independence of expressions in online posts can be ambiguous or misleading. The automated cyberbullying detection has attracted growing interest over the past decade as it facilitates combating toxic online behaviour. A lot of research has been done on detecting cyberbullying in textual data using a myriad of features [31]. Many datasets have been made open-source to facilitate research enthusiasts. Efforts from an interdisciplinary research space of computational social science and computational linguistics are required to deal with this antagonistic online delinquency. As a classical problem in natural language processing (NLP), cyberbullying detection in real-time user generated content needs high-level semantic analysis. Most of the earlier attempts on cyberbullying detection rely on manual feature extraction methods [36]. Such methods are not only time-consuming and cumbersome, but often fail to correctly capture the meaning of the sentence. Few lexicon-based methods by maintaining a list of offensive, abusive and hateful words have also been used, but are quite limited in scope [10]. Recent research focuses on the application of deep learning models for various NLP tasks and has reported state-of-the-art results [37]. Basically, deep architectures are neural networks with multiple processing layers of neurons with each layer having a specific task [2]. Utilizing deep learning models trivializes the need of explicit feature extraction techniques as these models are highly skillful and fast in retrieval of essential features and patterns by themselves. With minimal human intervention these models report superior results than the conventional machine learning models. Various deep learning architectures have contributed significantly in computational analytics of text [37]. Pre-trained word embeddings like Word2Vec, GloVe, ELMo, fastText that represent text in vector forms and deep neural networks such as CNN, RNN, GRU, LSTM, CapsNet & hierarchical networks that automate the task of feature extraction demonstrate best practices for solving text classification problems [17, 22].

In this paper, we propose a Bi-GRU-Attention-CapsNet (Bi-GAC) model to detect bullying in online textual content. The model benefits by learning both sequential semantic representations and spatial location information in textual content and acquires better generalization and prediction capabilities for cyberbullying tasks. The improved text representation and feature learning offers a robust model which can avoid the vanishing gradient problem in comparison to baseline neural models. Capsule networks (CapsNet) [20, 40] have the competence to express the semantic meanings in a wider vector space using which it captures the input instantiation parameters, such as word order and their semantic representation. The feature encoding ability of capsules and dynamic routing allow aggregating the important information. Simultaneously, bi-directional gated recurrent unit (Bi-GRU) [8, 9] are known for their sequential modeling capabilities which learns relationships between words from both directions. GRUs are easy and fast to train on smaller sequence data like social media posts, making them apposite for various text classification tasks on social media. Further, self-attention mechanism [8] helps to model dependencies between different parts of the sequence and captures important information using the mutual input interaction. Self-attention based neural networks basically extract neural features for the input sentence [32]. Thus apprehending the dexterity of Bi-GRU, self-attention and CapsNet, we put forward a hybrid model, Bi-GAC to classify online posts into bullying categories. The pre-trained ELMo word embedding [28] is used to create the input embedding matrix. Bi-GAC uses the fully connected output layer with sigmoid activation to finally classify the positive as positive (+ 1) of bullying or negative (-1) of bullying.

The model is validated on two benchmark datasets, Formspring.me [26] and MySpace [5] and compared with state-of-the-art model using the F1-score as performance measure. An ablation study is also done to ratify the results. Thus, the primary work undertaken in this research includes:

- Building a Bi-GRU-Attention-CapsNet (Bi-GAC) model for bully content classification
- Validating improved classifier performance in small sequences like social media posts by capturing semantic information, context and dependencies
- Evaluation of Bi-GAC on two benchmark datasets

The paper is structured as follows: Sect. 2 briefs about the related studies from pertinent literature followed by the details of the proposed Bi-GAC model in Sect. 3. Section 4 discusses the results whereas the final Sect. 5 presents the conclusion and future work.

2 Related studies

Social misbehavior on online portals has been studied across various dimensions such as hate-speech, aggression detection, comment toxicity and detecting cyberbullying [31]. Earlier works on cyberbullying are primarily available to understand the socio-cognitive aspects and behavioural intentions of bullies and victim experiences with the help of questionnaires and online surveys [14]. Studies are available on cause and effects of cyberbullying on teenagers, adolescents, school, college, ethnic and minority groups [13, 35]. Two surveys in 2019 were reported on automatic cyberbullying detection. Rosa et al. [29] conducted an in-depth analysis of 22 studies on automatic cyberbullying. Kumar and Sachdeva [16] did a meta-analysis on application of soft computing techniques for cyberbullying detection on social multimedia. Studies reveal most of the work on cyberbullying detection has been done using hand-crafted features with machine learning. Recent works using deep learning models have reported superlative results [7]. Agrawal et al. [1] implemented CNN, LSTM and its variants using attention using Twitter, Formspring and Wikipedia datasets. Meng et al. [25] applied a two-branch parallel neural network with multi-head self-attention mechanism (MHSA), capsule network (CapsNet) and independent recurrent neural network (IndRNN). Paul et al. [27] did the comparative analysis with the slot-gated or attention-based DL models using BERT for CB identification. Liu et al. [23] proposed a model for multi-label text classification using ELMo and attention with GRU on Kaggle's toxic comment classification data. Krešnáková et al. [24] carried out experimentation on the same Kaggle's dataset using different text pre-processing technique with DL models such as CNN, GRU, Bi-LSTM + CNN, Bi-GRU + CNN. Özel et al. [26] implemented various ML models such as SVM, kNN etc. on Formspring.me dataset using varied feature selection methods. Zhao et al. [39] performed automatic detection of cyberbullying using bag-of-words and latent semantic features on Twitter dataset.

3 The Bi-GAC model for cyberbullying detection

The Bi-GAC model which gets its nomenclature from its constituent core components, realizes the complexities of textual data in user-generated content where data representation is learned as real-valued vectors. The model has two core components, namely representation

learning and classification. A Bi-GRU encoder is trained using ELMo embedding to generate a sequence context feature vector. This feature vector is flawed due to the existence of redundant and irrelevant features. Consequently, a self-attention mechanism is added to capture significant information. Next the CapsNet generates semantic representation using a dynamic routing algorithm which is finally used for classification of the posts. Figure 2 depicts the architecture of Bi-GAC model.

3.1 Embedding layer

Word embeddings are based on the idea of distributional meaning: the fact that semantically (or morphologically) related words tend to appear in similar contexts. These represent each word using a continuously valued, lower-dimensional vector so that it reserves semantic information of the word. Once word embeddings have been trained, we can use them to derive similarities between words, as well as other relations. The pre-trained ELMo (Embeddings from Language Models) embedding [28] are used in this work. ELMo has an advantage above other conventional embeddings such as GloVe and word2vec as it encapsulates context in the word feature representations. These are high-dimensional representations of words, based on the contexts that different words appear in. ELMo uses a 2-layer bidirectional LSTM for learning words and their context. This design allows ELMo to learn more context-dependent aspects of word meanings in the higher layers along with syntax aspects in lower layers. Figure 3 shows an example of how an ELMo specific representation is generated by combining the bidirectional hidden representations.

3.2 Encoding layer

Recurrent neural network (RNN) models [6] have been popularly used in the field of natural language processing. But, seq2seq encoder/decoder architecture suffers from the vanishing gradients problem, that is the inability of the RNN unit to reminisce values

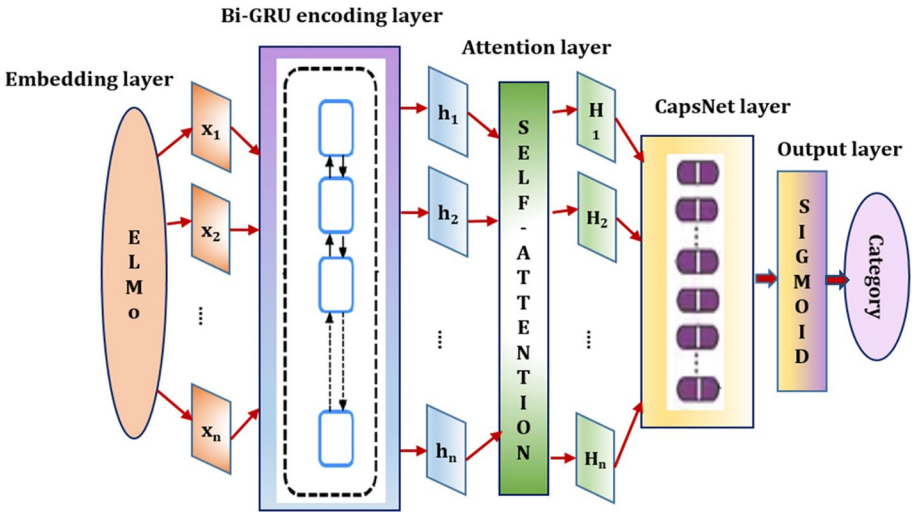

Figure 2 System architecture of proposed model

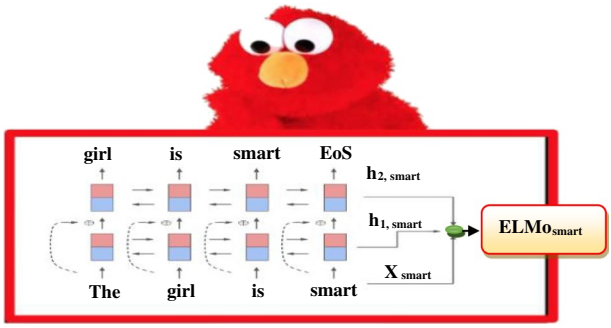

Fig. 3 ELMo-specific representation for "smart"

that showed initially in the sequence [11]. But this is vital within a typical NLP task like text classification, as few words rely on words that appear very early in the sentence. As a solution, both gated recurrent unit (GRU) & long-short-term memory (LSTM) solve the problem of vanishing gradients in RNN, by replacing the neurons of the hidden layer with the memory-units to store early sequence data [33]. GRU uses few training parameters, less memory, needs fewer data to generalize, execute faster and train faster than LSTM whereas LSTM is more accurate on dataset using longer sequence. As social media posts are shorter sequence, GRU is an apt choice for encoding. A GRU is a gating mechanism in RNN similar to a LSTM unit but without an output gate. The two gates of a GRU are as follows:-

- Update Gate: It determines how much of the past knowledge needs to be passed along into the future.
- Reset Gate: It determines how much of the past knowledge to forget.

The reset gate sits between the previous activation and the next candidate activation to forget the previous state, and the update gate decides how much of the candidate activation to use in updating the cell state. These gates aid in dealing with the long-term-dependencies. They forward and backward pass the information from the previous state to the next state. The update gate keeps a track of retaining all the important features and also aids in solving the long term temporal dependency issues. A vector having values from 0 to 1 is received via update gate with point-wise multiplication-operation (PMO). It uses sigmoid activation function for squashing values b/w 0 and 1. It basically helps in updating or forgetting (or disappearing) the data as the PMO would result in 0 for any multiplication with a vector of zeros. In such a scenario, the resulting values would be considered to have 'disappeared or be forgotten'. Contrastingly, for any multiplication with a vector of ones would result in the 'same value or be kept'. This mechanism would eventually help in retaining the relevant data and forgetting the non-relevant ones. The other gate which is the reset gate is normally used to ensure the amount of past information that can be forgotten. Figure 4 depicts the architecture of a basic GRU cell.

In this research, we use a bidirectional GRU's (Bi-GRU) encoder which takes the input sequence and encapsulates the information as the internal state vectors. A Bi-GRU allows capturing information from both previous time steps and later time steps to make predictions about the current state. Bi-GRU enables apprehending meaning and context for the sentences than a simple GRU. The forward GRU $\vec{f}$ reads the sentence s_i from w_{il} to w_{iT} as given in (1):

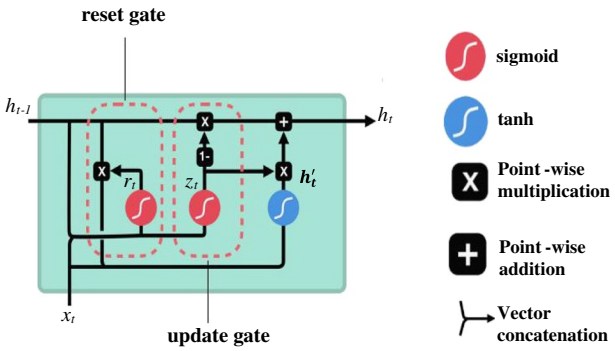

Figure 4 GRU Cell architecture

$$\vec{h}_{\mathrm{it}} = \overrightarrow{GRU}\left(X_{it}\right),\ \mathrm{t} \in [1, \mathrm{T}] \tag{1}$$

The backward GRU $f^{\leftarrow}$ reads sentences from w_{iT} to W_{i1}, as given in (2):

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}\left(X_{it}\right),\ t \in [T, 1] \tag{2}$$

The annotation of the word w_{it} is calculated by combining the forward and backward hidden states i.e., $h_{\mathrm{it}} = [\vec{h}_{\mathrm{it}}, \overleftarrow{h}_{\mathrm{it}}]$.

3.3 Attention layer

Attention mechanism allows output to focus attention on input while producing output [12, 21] whereas a self-attention model allows inputs to interact with each other, that is, calculate attention of all other inputs with respect to one input. Self-attention is good at modeling dependencies between different parts of the sequence. Self- attention includes both location and observation value information and replaces conditioning on the entire sequence with pairwise comparisons (the importance of one word and its position conditional on some other word and its location) given as vector representations of both. Consider the sentence "*The dog didn't eat the food because it was too full*". The word *"it"* refers to the dog. If we replace *"full"* with *"much"*, the word *"it"* now refers to the food. Attention mechanism helps to understand this, that is, in the former case there's high attention linking *"it"* and *"dog"* but in the latter case high attention shifts to *"food"*. In this research, a self-attention mechanism [32] as shown in Figure 5 is applied on the outputs of the GRU layer.

3.4 CapsNet layer

Conventionally, artificial neurons output a scalar, real-valued activation that loosely represents the probability of an observation. CapsNets [15, 40] replace the scalar output feature detectors with vector output capsules and max-pooling with routing-by-agreement. A capsule is a collection of neurons which represent distinctive properties of the same entity as outputs. Capsule's activity vector represents the instantiation parameters of a specific type of entity which adequately captures the important and relevant information (features). The CapsNet is divided into two layers: primary capsule layer and digit capsule layer. As the

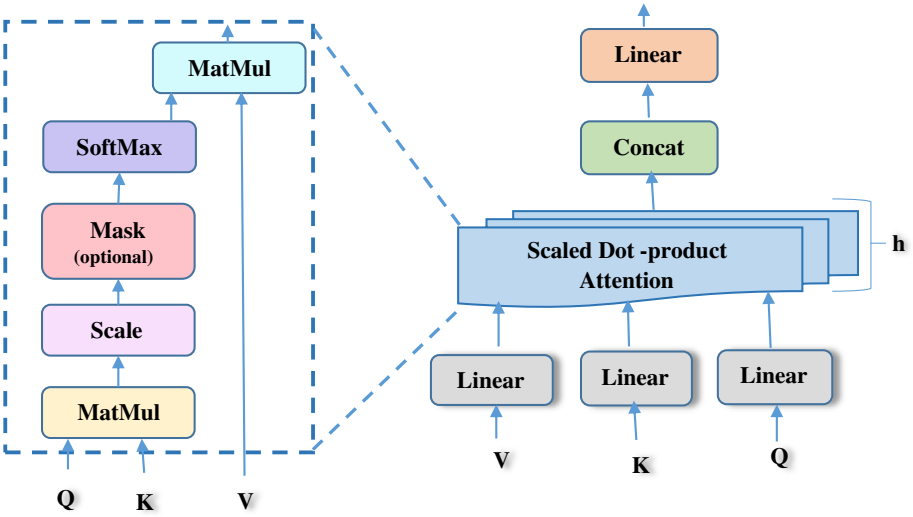

Figure 5 Self-Attention [32]

preceding Bi-GRU attention layer outputs high-level feature representation, these feature maps are input to the primary caps and subsequent digit caps layer as shown in Figure 6. Capsules with vector outputs are generated by the primary caps layers as given (3).

$$u_{j|i} = W_{ij} u_i \tag{3}$$

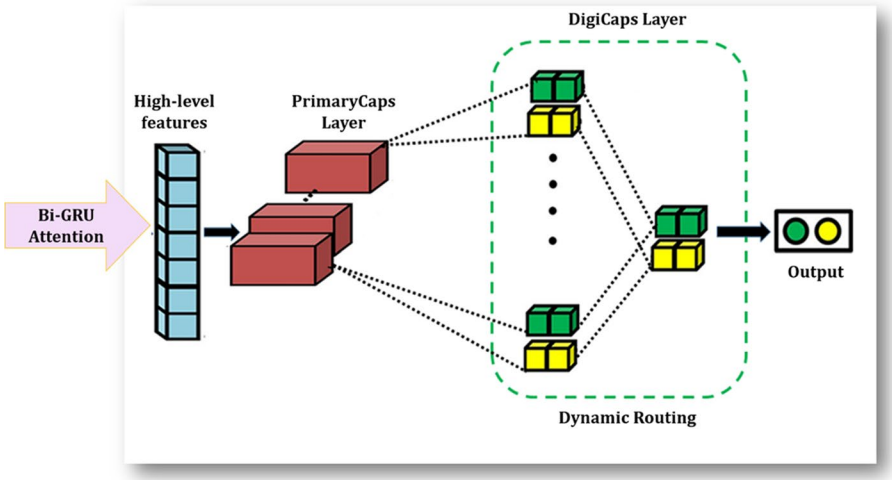

Figure 6 CapsNet Architecture

where, 'i' is in the current lower-level primary-caps and 'j' is in the next-level layer. Connection-weight updating (as shown in Eq. 4) will occur during network training using dynamic routing algorithm.

$$s_j = \sum_i c_{ij} u_{j|i} \tag{4}$$

The key operations within the capsule are as follows [19]:

- Multiplying input vector matrix and weights to encode significant features and their relationships within the text.
- Dynamic routing for sending output from one low-level capsule to a higher level capsule.
- Weighted input vectors summation.
- "Squash" function to add non-linearity using. This function takes a vector and "squashes" it to have a maximum length of 1, and a minimum length of 0 while retaining its direction.

Figure 7 summarizes the operations within a capsule.
The dynamic routing algorithm [30] is given below (algorithm 1).

Algorithm 1. Dynamic routing algorithm

procedure ROUTING($\hat{u}_{j|i}$, r, l)
  for all capsule i in layer l and capsule j in layer (l + 1): $b_{ij} \leftarrow 0$.
  for r iterations **do**
    for all capsule i in layer l: $c_i \leftarrow \text{softmax}(b_i)$
    for all capsule j in layer (l + 1): $s_j \leftarrow \sum_i c_{ij} \hat{u}_{j|i}$
    for all capsule j in layer (l + 1): $v_j \leftarrow \text{squash}(s_j)$
    for all capsule i in layer l and capsule j in layer (l + 1): $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} . v_j$
return v_j

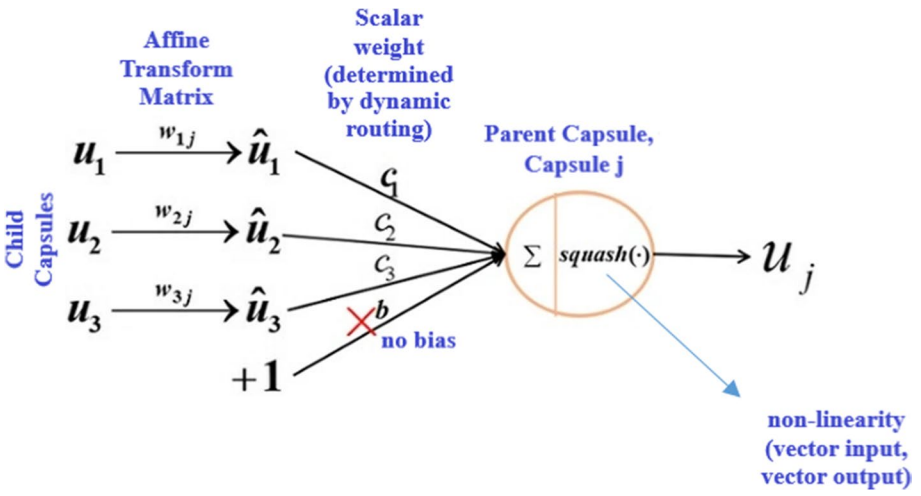

Figure 7 Operations within a Capsule [22]

Table 1 Sample distribution in datasets

Dataset	Bullying samples	Non-Bullying samples	Total samples
Formspring.me [26]	836	12,288	13,124
MySpace [5]	357	1396	1753

Table 2 F-1 Score for Bi-GAC

Dataset	F1-Score
Formspring.me	94.03
MySpace	93.89

3.5 Output layer

The final prediction layer of the proposed model is the fully-connected layer with sigmoid activation that eventually helps in obtaining the probabilities for the binary classification (target classes: bullying and non-bullying). It basically maps a real valued number using a threshold to a probability (i.e. to a no. b/w 0 to 1).

4 Computational results

Various social media datasets from Twitter, YouTube, MySpace, Kongregate, Formspring and Slashdot have been created for automatic cyberbullying detection. In this work, two benchmark datasets, namely, question-answering Formspring.me and thread-style MySpace social network cyberbullying classification datasets are used for experimentation. The datasets are labelled for cyberbullying and non-bullying type and the sample distribution is as given in Table 1.

Evidently, the datasets suffer from class imbalance and we over-sampled the data from the bullying class (sporadic class) thrice to deal with this skewness. This technique is usually used for balancing the corpus in similar studies [1]. The training: test ratio was taken as 70:30 and tenfold cross validation is used on each dataset. Using 10 folds means that in each iteration of cross-validation the validation-set would be approximately 10% of the size of the total dataset. The results are then averaged across the folds, using suitable performance measures. We used the Scikit-learn library and Keras deep learning library with Theano backend. The hyperparameters in experiments were set as follows: The Bi-GRU layer had 50 units and a dropout value of 0.2. The CapsNet layer also had a drop out of 0.2 and had 3 iterations for dynamic routing. Adam optimizer was used with the learning rate of 0.0001. The Bi-GAC model was evaluated using the F-1 score. The results for both datasets are shown in Table 2.

The results are also compared with the existing techniques used on both datasets. The performance comparison of the proposed Bi-GAC model with the existing techniques is shown in Tables 3 and 4 for MySpace and Formspring.me datasets respectively. A superior performance is observed for the proposed Bi-GAC model.

Table 3 Comparison of Bi-GAC with existing works on MySpace dataset

Reference Study	Techniques	F1-score
Kumar & Sachdeva [18]	k-nearest neighbour	67.7
Kumar & Sachdeva [18]	J48	69.6
Zhang et al. [38]	Logistic Regression	78
Zhang et al. [38]	Support Vector Machine	79
Zhang et al. [38]	CNN	85
Proposed Model	Bi-GAC	93.89

Table 4 Comparison of Bi-GAC with existing works on formspring.me dataset

Reference Study	Techniques	F1-score
Agrawal & Awekar [1]	Random Forest	29.8
Agrawal & Awekar [1]	Naive Bayesian	35.9
Agrawal & Awekar [1]	Support Vector Machine	42.2
Agrawal & Awekar [1]	Logistic Regression	44.8
Agrawal & Awekar [1]	Bi-LSTM	86
Paul & Saha [27]	RNN + LSTM	88
Paul & Saha [27]	Bi-LSTM with Attention	91
Paul & Saha [27], Agrawal & Awekar [1]	CNN	91
Proposed Model	Bi-GAC	94.03

Table 5 Ablation architectures

Dataset	Bi-LSTM + Attention + CapsNet	Bi-GRU + Attention + CNN	Bi-GRU + Attention + CapsNet
Formspring.me	92.67	91.83	94.03
MySpace	93.10	92.35	93.89

We also perform an ablation study to learn about the network by removing and/or replacing parts of the complex neural network architecture and study the model performance. The two variations studied are: using Bi-LSTM instead of Bi-GRU (Bi-LSTM + Attention + CapsNet) and using CNN instead of CapsNet (Bi-GRU + Attention + CNN). Table 5 presents the F1-score of these variations where the proposed Bi-GAC (Bi-GRU + attention + CapsNet) outperforms the other two.

Figure 8 shows the ROC (Receiver Operating Characteristic)—AUC (Area under curve) curve. ROC curve tells us about how good the model can distinguish between two classes. The AUC score gives a good idea of how well the model performs.

5 Conclusion

Social media serves as a virtual playground for bullying and allows the offenders to be mysterious, tough to trace and shielded from confrontation making the detection of bullying instances a challenging task. It can become a source of misery for victims and

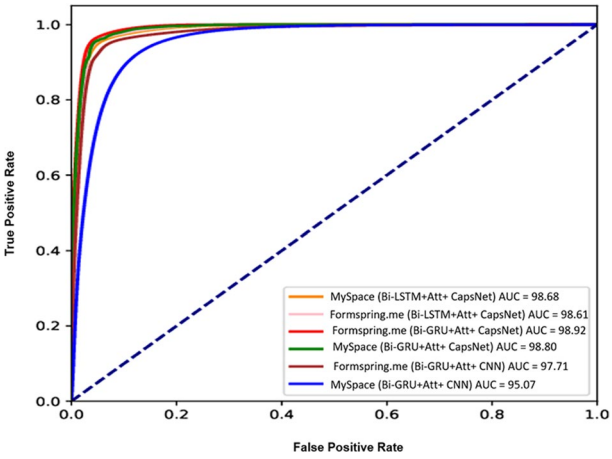

Figure 8 ROC-AUC for MySpace and Formspring.me datasets

often lead to mental distress and depression, and in extreme situations, even suicide. It is necessary to be both responsive and proactive to such a toxic environment on social media platforms. This research proposed a hybrid deep learning model for cyberbullying classification task which combines the advantages of self-attention based Bi-GRU encoder and capsule network. ELMo contextual embeddings are used as input. The model achieves a superior F1-score of 94.03 and 93.89 for Formspring.me and MySpace benchmark cyberbullying datasets respectively.

As a future direction, we would like to use transformer-based models like BERT, especially on multimodal and multilingual cyberbullying datasets. Fine-grain classification into cyberbullying categories has not been explored as yet and we would like to extend the model to multi-class cyberbullying categories. Cyberbullying divides the stakeholders into categories of offender, victim and by-stander (assistants, reinforcers, defenders and outsiders) and therefore combining user-based features with text-based features to train models for real-time cyberbullying needs investigation.

Author's contributions All the authors have contributed equally in the research and manuscript preparation.

Data availability Benchmark publicly available datasets have been used.

Code availability Can be made available on request.

Declarations

Ethics approval The work conducted is not plagiarized. No one has been harmed in this work.

Consent to participate All the authors have given consent to submit the manuscript.

Consent for publication Authors provide their consent for the publication.

Conflict of interest The authors certify that there is no conflict of interest in the subject matter discussed in the manuscript.

References

1. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. In: European Conference on Information Retrieval, pp. 141–153. Springer, Cham (2018)
2. Ballal, N, Saritha, SK: A study of deep learning in text analytics. In: Shukla, R, Agrawal, J, Sharma, S, Chaudhari, N, Shukla, K (eds.) Social networking and computational intelligence. Lecture notes in networks and systems, vol. 100. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2071-6_16
3. Bounegru, L, Gray, J, Venturini, T, Mauri, M: A field guide to 'fake news' and other information disorders. A field guide to" fake news" and other information disorders: a collection of recipes for those who love to cook with digital methods. Public Data Lab, Amsterdam (2018)
4. Campbell, MA: Cyber bullying: an old problem in a new guise? Aust. J. Guid. Couns. **15**(1), 68–76 (2005)
5. Çiğdem, A.C.I., Çürük, E., Eşsiz, E.S.: Automatic detection of cyberbullying in FORMSPRING. Me, Myspace and Youtube Social Networks. Turk J Eng **3**(4), 168–178 (2019)
6. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
7. Dadvar, M., Eckert, K.: Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv preprint arXiv:1812.08046 (2018)
8. Deng, J., Cheng, L., Wang, Z.: Self-attention-based BiGRU and capsule network for named entity recognition.arXiv preprint arXiv:2002.00735 (2020)
9. Gangwar, A. K., Ravi, V.: A novel BGCapsule network for text classification. arXiv preprint arXiv:2007.04302 (2020)
10. Hang, O.C., Dahlan, H.M.: Cyberbullying lexicon for social media. In: 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), pp. 1–6. IEEE (2019)
11. Hochreiter, S: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems **6**(02), 107–116 (1998)
12. Jain, D, Kumar, A, Garg, G: Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Appl. Soft Comput. **91**, 106198 (2020)
13. John, A., Glendenning, A. C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., ... Hawton, K.: Self-harm, suicidal behaviours, and cyberbullying in children and young people: systematic review. J. Med. Internet Res. **20**(4), e129 (2018)
14. Kowalski, RM, Limber, SP: Psychological, physical, and academic correlates of cyberbullying and traditional bullying. J. Adolesc. Heal. **53**(1), S13–S20 (2013)
15. Kim, J, Jang, S, Park, E, Choi, S: Text classification using capsules. Neurocomputing **376**, 214–221 (2019)
16. Kumar, A., Sachdeva, N.: Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimed Tools Appl **78**(17), 23973–24010 (2019)
17. Kumar, A, Jaiswal, A: A deep swarm-optimized model for leveraging industrial data analytics in cognitive manufacturing. IEEE Trans. Industr. Inf. **17**(4), 2938–2946 (2020)
18. Kumar, A., Sachdeva, N.: Cyberbullying checker: online bully content detection using Hybrid Supervised Learning. In: International Conference on Intelligent Computing and Smart Communication 2019, pp. 371–382. Springer, Singapore (2020)
19. Kumar, A., Sachdeva, N.: Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia Systems, 1–15 (2020)
20. Kumar, A., Sachdeva, N.: Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimedia Systems, 1–10 (2021)
21. Kumar, A., Sangwan, S.R., Arora, A., Nayyar, A., Abdel-Basset, M.: Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access **7**, 23319–23328 (2019)
22. Kumar, A., Srinivasan, K., Cheng, W.H., Zomaya, A.Y.: Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Inf Process Manag **57**(1), 102141 (2020)
23. Liu, W., Wen, B., Gao, S., Zheng, J., Zheng, Y.: A multi-label text classification model based on ELMo and attention. In: MATEC Web of Conferences, Vol. 309, pp. 03015. EDP Sciences (2020)
24. Maslej-Krešňáková, V, Sarnovský, M, Butka, P, Machová, K: Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Appl. Sci. **10**(23), 8631 (2020)

25. Meng, Z, Tian, S, Yu, L: Regional bullying text recognition based on two-branch parallel neural networks. Autom. Control. Comput. Sci. **54**(4), 323–334 (2020)
26. Özel, SA, Sarac, E: Effects of feature extraction and classification methods on cyberbully detection. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi **21**(1), 190–200 (2016)
27. Paul, S., & Saha, S. (2020). CyberBERT: BERT for cyberbullying identification.*Multimedia Systems*, 1–8.
28. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations.arXiv preprint arXiv:1802.05365 (2018)
29. Rosa, H, Pereira, N, Ribeiro, R, Ferreira, PC, Carvalho, JP, Oliveira, S, Coheur, L, Paulino, P, Simão, AV, Trancoso, I: Automatic cyberbullying detection: a systematic review. Comput. Hum. Behav. **93**, 333–345 (2019)
30. Sabour, S., Frosst, N., Hinton, G. E.: Dynamic routing between capsules. arXiv preprint arXiv:1710.09829 (2017)
31. Sangwan, S.R., Bhatia, M.P.S.: D-BullyRumbler: a safety rumble strip to resolve online denigration bullying using a hybrid filter-wrapper approach. Multimedia Systems, 1–17 (2020)
32. Shao, Y, Lin, JCW, Srivastava, G, Jolfaei, A, Guo, D, Hu, Y: Self-attention-based conditional random fields latent variables model for sequence labeling. Pattern Recogn. Lett. **145**, 157–164 (2021)
33. Sherstinsky, A: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D **404**, 132306 (2020)
34. Shrivastava, G., Kumar, P., Ojha, R.P., Srivastava, P.K., Mohan, S., Srivastava, G.: Defensive modeling of fake news through online social networks. IEEE Trans Comput Soc Syst **7**(5), 1159–1167 (2020)
35. Smith, PK, Mahdavi, J, Carvalho, M, Fisher, S, Russell, S, Tippett, N: Cyberbullying: Its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry **49**(4), 376–385 (2008)
36. Van Hee, C, Jacobs, G, Emmery, C, Desmet, B, Lefever, E, Verhoeven, B, Hoste, V: Automatic detection of cyberbullying in social media text. PLoS ONE **13**(10), e0203794 (2018)
37. Young, T, Hazarika, D, Poria, S, Cambria, E: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)
38. Zhang, A., Li, B., Wan, S., Wang, K.: Cyberbullying detection with birnn and attention mechanism. In: International Conference on Machine Learning and Intelligent Communications, pp. 623–635. Springer, Cham (2019)
39. Zhao, R., Zhou, A., Mao, K.: Automatic detection of cyberbullying on social networks based on bullying features. In: Proceedings of the 17th international conference on distributed computing and networking, pp. 1–6 (2016)
40. Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., Zhao, Z.: Investigating capsule networks with dynamic routing for text classification. arXiv preprint arXiv:1804.00538 (2018)