Expert Systems with Applications 161 (2020) 113725

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance

Sandip Modha[a,*], Prasenjit Majumder[a], Thomas Mandl[a,b], Chintak Mandalia[c]

[a] DA-IICT, Gandhinagar, India
[b] University of Hildesheim, Germany
[c] infoAnalytica, Ahmedabad, India

ARTICLE INFO

Article history:
Received 24 November 2019
Revised 20 April 2020
Accepted 4 July 2020
Available online 13 July 2020

Keywords:
Natural language processing
Deep learning
Hate speech
Aggression detection
Text classification
Social media visualization

ABSTRACT

The multi-fold growth of the social media user-base fuelled a substantial increase in the amount of hate speech posts on social media platforms. The enormous data volume makes it hard to capture such cases and either moderate or delete them. This paper presents an approach to detect and visualize online aggression, a special case of hate speech, over social media. Aggression is categorized into overtly aggressive (OAG), covertly aggressive (CAG), and non-aggressive labels (NAG). We have designed a user interface based on a web browser plugin over Facebook and Twitter to visualize the aggressive comments posted on the Social media user's timelines. This plugin interface might help to the security agency to keep a tab on the social media stream. It also provides citizens with a tool that is typically only available for large enterprises. The availability of such a tool alleviates the technological imbalance between industry and citizens. Besides, the system might be helpful to the research community to create further tools and prepare weakly labeled training data in a few minutes using comments posted by users on celebrity's Facebook, Twitter timeline. We have reported the results on a newly created dataset of user comments posted on Facebook and Twitter using our proposed plugins and the standard Trolling Aggression Cyberbullying 2018 (TRAC) dataset in English and code-mixed Hindi. Various classifiers like Support Vector Machine (SVM), Logistic regression, deep learning model based on Convolution Neural Network (CNN), Attention-based model, and the recently proposed BERT pre-trained language model by Google AI, have been used for aggression classification. The weighted F1-score of around 0.64 and 0.62 is achieved on TRAC Facebook English and Hindi datasets while on Twitter English and Hindi datasets, the weighted F1-score is 0.58 and 0.50, respectively.

1. Introduction

The large fraction of hate speech and other offensive and objectionable content online poses a considerable challenge to societies (Seglow, 2016). Offensive language such as insulting, hurtful, derogatory, or obscene content directed from one person to another person and open for others undermines objective discussions (Assimakopoulos et al., 2017). Such type of language can be more increasingly found on social web platforms and can lead to the radicalization of debates (Pelzer et al., 2018; Kwok and Wang, 2013). The debate on hate speech falls within the broader topic of content moderation. The introduction of content moderation obligations for large online platforms in several countries, e.g., in Germany, has led to intense public discussions. Public opinion-forming requires rational-critical discourse with an exchange of arguments (Habermas et al., 1984). Objectionable content can pose a threat to opinion-forming for open societies, and ultimately to democracy. Although there is no definitive proof, several studies have claimed a deterioration of the climate of discussions online in different contexts, e.g., Canada (Braun, 2004) and the EU (Assimakopoulos et al., 2017). Recent analysis has shown differentiated results for many regions (Lu and Yu, 2020). At the same time, societies need to find adequate ways to react to such content without imposing rigid censorship regimes (Land, 2019). As a consequence, many platforms of social media websites monitor user posts. Content moderation is also a legal requirement in some countries.

The need for content moderation leads to a pressing demand for methods to automatically identify suspicious posts. Online communities, social media enterprises, and technology companies have

[*] Corresponding author.
E-mail addresses: sjmodha@gmail.com (S. Modha), prasenjit.majumder@gmail.com (P. Majumder), mandl@uni-hildesheim.de (T. Mandl), chintak.soni75@gmail.com (C. Mandalia).

https://doi.org/10.1016/j.eswa.2020.113725
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

been investing heavily in technology and processes to identify the offensive language to prevent abusive behavior in social media. Popular social media like Facebook and Twitter have deployed their hate speech detection tools to track offensive or hate-related content. However, such tools are not accessible to the common user or a law enforcement agency.

The amount of hate speech or and offensive content, in general, is a research problem of central attention within the natural language processing community. The research community is reshaping the hate speech problem into a very fine-grained issue, a problem like aggression, abusive, or offensive text. These fine-grained problems are also ambiguous for humans. Standard datasets are available at various forum like TRAC (Ritesh et al., 2018) for aggression detection (Ritesh et al., 2018), GermEval (Wiegand et al., 2018) for offensive content detection in German language, SemEval OffenEval 2019 for offensive content detection in the English language (Zampieri et al., 2019), and HASOC (Mandl et al., 2019), for hate and offensive content in English, German, and Hindi languages. An overview of hate speech datasets is also provided by MacAvaney et al. (2019).

The TRAC dataset (Ritesh et al., 2018) was the first inherently multilingual dataset. It contains Hindi, English, and also includes mixed script tweets. This makes the TRAC dataset especially interesting for multilingual societies. Therefore, it was used for this study. In addition, the aggression level in the text is categorized into three classes. The classes are defined as 'Non-aggressive' (NAG), 'Covertly Aggressive' (CAG), and 'Overtly Aggressive' (OAG). The posts belonging to the class CAG are difficult to identify. Some examples of highly aggressive tweets from the TRAC dataset are listed in Table 1. Hate speech annotation often reveals that hate is a very subjective concept. The three levels make TRAC more realistic than binary datasets. Furthermore, the training dataset is sampled from Facebook, while two test datasets are sampled from Facebook and Twitter. The reported weighted F1-score on the Twitter test dataset is substantially lower (around 10% for English and 20% for Hindi) than the F1-score for the Facebook test dataset (Ritesh et al., 2018). This shows that there is a bias depending on the social media platform from which the training data was collected.

Verbal aggression could be understood as any linguistic behavior which intends to damage the social identity of the target person and lower their status, and prestige (Ritesh et al., 2018; Culpeper, 2011). Any speech or text in which aggression is overtly expressed either through the use of specific kinds of lexical items or lexical features which are considered aggressive and or certain syntactic structures is overt aggression (Ritesh et al., 2018). Covertly aggression contains an indirect attack against the victim and is framed as a polite or sarcastic expression and might not contain any abusive/offensive/controversial word. Non-aggressive posts do not contain any aggressive content. Table 1 shows examples of overtly aggressive and covertly aggressive/sarcastic comments. Sometimes posts are written in Hindi language using a Roman script instead of Devanagari script. Therefore, multilingual text content poses serious challenges to hate speech detection tools.

In this paper, we tried to address issues related to how the detection results can be displayed in online environments. For that purpose, a tool that runs on live social media streams, such as Facebook and Twitter, is developed as a proof of concept. These are the objectives of this paper:

(1) Evaluate state-of-the-art classifiers for hate speech detection on the TRAC dataset.
(2) Visualize hate speech on the fly on Facebook & Twitter in a web browser.
(3) Explore use-cases for hate detection and visualization systems.

Table 1
Sample aggressive post.

Text	Class Label	Language
shut up you bloody actor, f*ck yourself!! why do you took Modi's biography which is useless for the whole country, better f*ck yourself for acting in this film	OAG	English
Hamare modi ji kisi Aladdin se kam hai ke? (Is our Modiji less than any Aladdin?)	CAG	Code-mixed Hindi

The rest of the paper is organized as follows: In Section 2, we review the relevant works in the area of hate speech detection. Section 3 lists the motivation behind the development of this tool. Various classification model are discussed in Section 4. We report results regarding the classification in Section 5. Section 6 contains the detailed architecture of the tool. We summarize the potential applications of the tool in Section 6.3. We conclude the discussion and provide insight for the future work in Section 7.

2. State of the art

Hate speech detection research attracts researchers from diverse backgrounds such as Computational Linguistics, Computer Science, Law, and Social Science. The actual term hate speech was coined by Warner and Hirschberg (2012). Various Authors used different notions such as offensive language (Razavi et al., 2010), Cyberbullying (Xu et al., 2012), or Aggression (Ritesh et al., 2018).

2.1. Defining and analyzing problematic content

Greenwell and Dengerink (1973) studied the role of a perceived versus an actual attack in human physical aggression. The authors concluded that verbal abuse or aggression is as damaging as physical abuse. Culpeper (2011) studied mechanisms of impoliteness through cross-cultural comparisons. The author discusses impoliteness within linguistics and puts it into the context of the debate in pragmatics. Schmidt and Wiegand (2017) define the term hate speech as an umbrella of various anti-social or impoliteness related utterance. These include flames, abusive/hostile messages, offensive content, cyberbullying. Fortuna and Nunes (2018) reported a detailed survey on hate speech detection from a computer science perspective. They have reported several definitions that range from profanity, flame, abusive messaging, extremism to radical content. A precise definition of hate speech and clear differentiation from related concepts cannot be found universally.

Waseem et al. (2017) study the abusive language on the web and categorized different hate speech-related problems in a 2-way topology based upon the type of the target (an individual or a group) and type of the attack (explicit and implicit). Although most definitions seem rather straightforward, it is difficult for humans to draw the line between hate speech and normal content exactly. Malmasi and Zampieri (2018) point out the issue of subjectivity for annotating offensive content. For annotation, agreement on a guideline and common rules seems to be necessary to have a common understanding. In particular, the annotation process is difficult for the content which does not belong to one of the extremes. For these middle cases or grey areas, the inter-annotator reliability drops significantly (Salminen et al., 2019). It is not even scientifically clear whether distributing guidelines increases the quality of the annotation process. One study could not prove any effect (Ross et al., 2016).

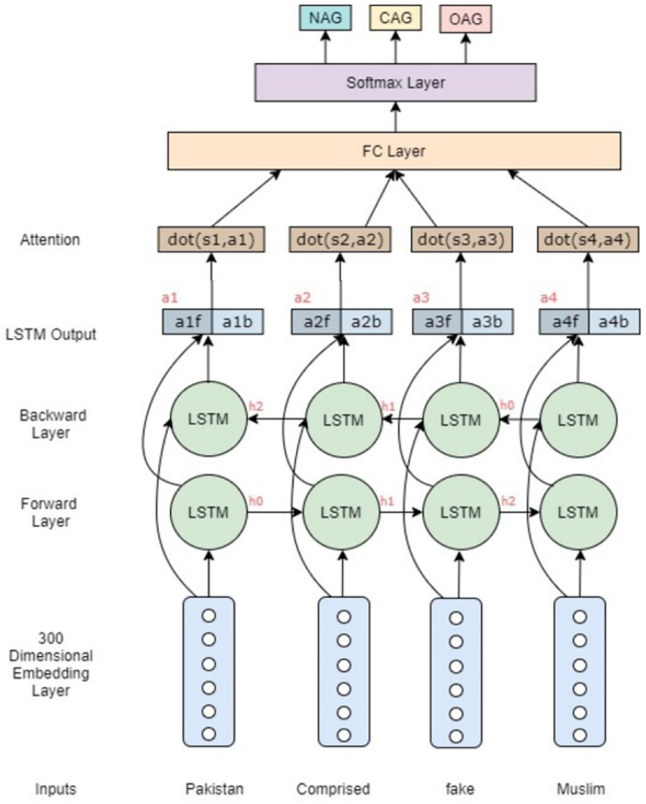


Fig. 1. Model Architecture with attention mechanism.

2.2. Text classification approaches for problematic content

Xu et al. (2012) introduces cyberbullying as a concept for classification. The authors ran various binary classification experiments on tweets text with a bullying perspective to determine whether the user is cyberbullied or not and reported binary classification accuracy around 81%. Kwok and Wang (2013) have tried to classify whether tweets are directed against colored people or not. They have collected two classes (racist and non-racists) of tweets and used a Naive Bayes classifier for binary classification. The average accuracy for the binary classification was around 76%.

There has been a variety of methods applied to hate speech detection. Fortuna and Nunes (2018) have presented a search technique to select papers on hate speech detection from the Internet. Authors have reported results along with a set of features of the various highly-cited paper. Authors describe different datasets and summarize the research challenges mentioned by the various authors. In the last few years, deep learning models started to mark their footprint in all NLP tasks (Vaswani et al., 2017; Ritesh et al., 2018). Yang et al. (2016) proposed a hierarchical attention network for document classification. The proposed model focused on the hierarchical structure of documents using two levels of attention mechanisms applied at the word and sentence level. This enables the model to construct a document representation differentiating between more and less important content.

Pitsilis et al. (2018) experimented in this direction by designing classifiers for different tasks using different input features. The use of the input features and the analysis of the performance of classifiers allows the first step toward interpretability. Bourgonje et al. (2017) also worked toward that goal, and they identified the most relevant features in a dataset. In that manner, they detected words that are most typical for hate speech. A similar approach is taken by Malmasi and Zampieri (2018).

2.3. Visualizing problematic content

Open systems can allow users to classify their content and experiment with the input. Seeing the effects of a classifier may contribute to a better understanding of classification and more trust in algorithms. Many authors (e.g., Ribeiro et al. (2016)) have stressed the importance of better understanding and explaining machine learning algorithms, which are often perceived as black boxes. The visualization of hate speech or aggression for users within social media is an emerging area, but the systems proposed in the scientific literature are still very limited. The system proposed by Kurniasih et al. (2018) is limited to simple pattern matching for words that can not capture the complexity of most natural language processing tasks, including hate expressions.

As mentioned, there are very few approaches that filter or label content on Facebook or other social media systems. The Facebook Inspector, developed by Dewan and Kumaraguru (2017), filters out malicious content in real-time using a Random Forest classifier. Unfortunately, the plugin is not available anymore. It is obvious that there is still a need to search for optimal machine learning systems that are trained on standard collections and which also encompass flexible software architectures and usable interfaces.

3. Purposes of the hate speech visualization tool

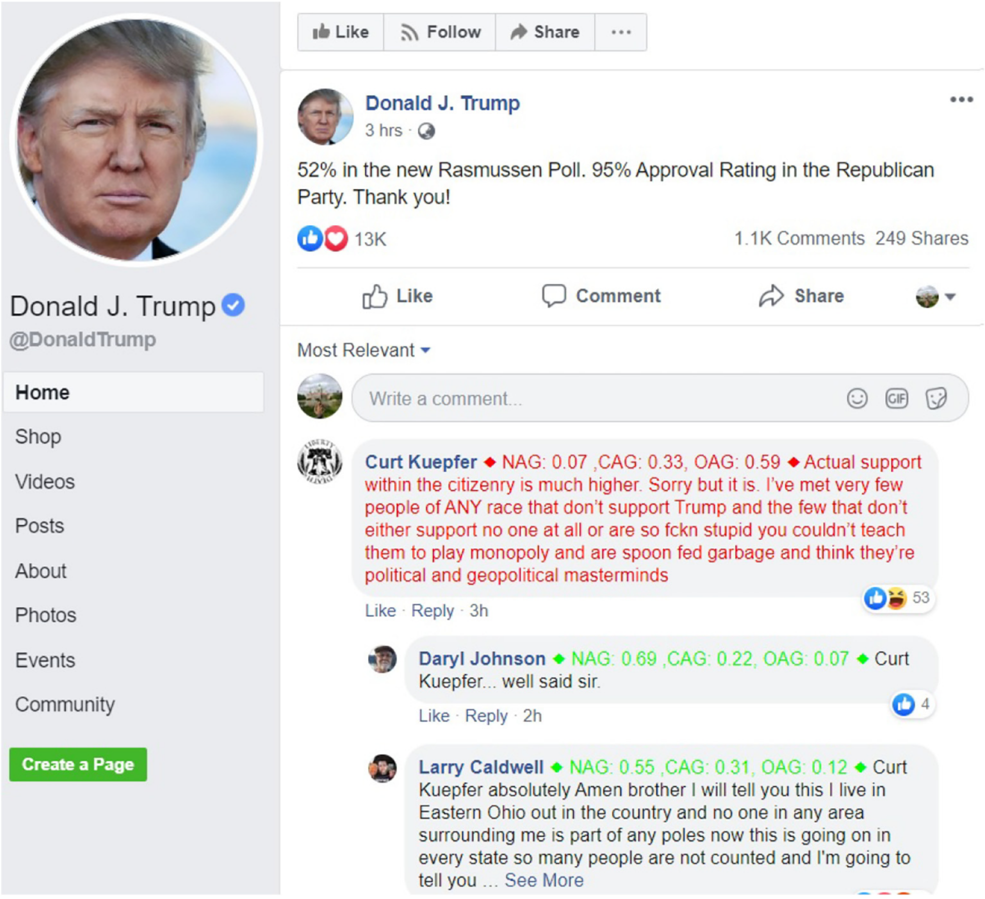
The main objective behind the development of the tool is to identify and visualize aggressive or offensive content posted by social media users. As such, our approach goes beyond previous research which purely identifies undesired content and suggests to delete it. In this section, we will present various motivations behind the development of the visualization interface for aggressive and hateful content.

3.1. Existing benchmarks

Facebook and Twitter have deployed their proprietary online hate speech detection tools to track aggressive or hate-related content internally. However, on many occasions, such tools fail to detect such content (Kottasová, 2017). It is nearly impossible for the common user or law-enforcement agencies to evaluate or regulate these tools. Multilingual content and the massive volume of the social media stream pose serious challenges for hate speech detection tools.

3.2. Supporting safety during browsing

Social Media has tremendous power to viral sensitive information to billions of people within a short period of time. Sometime, this sensitive information might include aggressive and potentially harmful content, which is then disseminated over the web. Social media also became a platform where bullying now occurs. Bullying in online platforms is often referred to as cyberbullying. Incidents of trolling and cyberbullying badly affect the life of common people. Hinduja and Patchin (2010) reported that victims feel extreme physical or mental suffering, and that may lead to the deactivation of their social media account. In rare cases, such incidents led the victim to commit suicide. Our plugin interfaces as shown in Figs. 2–4 changes the color of the contents from black to red or yellow if the model detects aggressive content in the comments. As soon as comments load in the user timeline, the plugin cautions the user before she read such content. Overtly aggressive (OAG) comments are rendered in red color while covertly/sarcastic comments are rendered in yellow color. This gives the user a choice before accessing and reading content.



52% in the new Rasmussen Poll. 95% Approval Rating in the Republican Party. Thank you!

Curt Kuepfer ◆ NAG: 0.07 ,CAG: 0.33, OAG: 0.59 ◆ Actual support within the citizenry is much higher. Sorry but it is. I've met very few people of ANY race that don't support Trump and the few that don't either support no one at all or are so fckn stupid you couldn't teach them to play monopoly and are spoon fed garbage and think they're political and geopolitical masterminds

53

Like · Reply · 3h

Daryl Johnson ◆ NAG: 0.69 ,CAG: 0.22, OAG: 0.07 ◆ Curt Kuepfer... well said sir.

4

Like · Reply · 2h

Larry Caldwell ◆ NAG: 0.55 ,CAG: 0.31, OAG: 0.12 ◆ Curt Kuepfer absolutely Amen brother I will tell you this I live in Eastern Ohio out in the country and no one in any area surrounding me is part of any poles now this is going on in every state so many people are not counted and I'm going to tell you ... See More

Fig. 2. Web Browser plugin on President Donald Trumps' Facebook Page.

3.3. Protecting victims of aggression

Social media gives freedom of speech and anonymity to its users. However, often, social media users exploit this liberty to spread abuse and hate through posts or comments. On many occasions, these user-generated content is offensive or actively aggressive. Such content is written in a way that might defame or insult individuals or groups of people without actually using any explicit hate-related or abusive words. Such posts cannot be found by using merely lexical approaches. Social media users may become the victim of such abusive or hateful comments. Our classifier model can classify many sarcastic comments into the covertly aggressive (CAG) category. The plugin changes the color of the text of such a covertly aggressive comment to yellow, which enables the user to delete content or block the other user without reading the content.

3.4. Visualization of aggressive contents

To the best of our knowledge, we did not come across any interface or tool which enables a user to visualize online hate or textual aggression present in the text. This is our novel approach to empower the common user by visualizing hate over the internet. The user needs to add our plugin inside the browser extension.

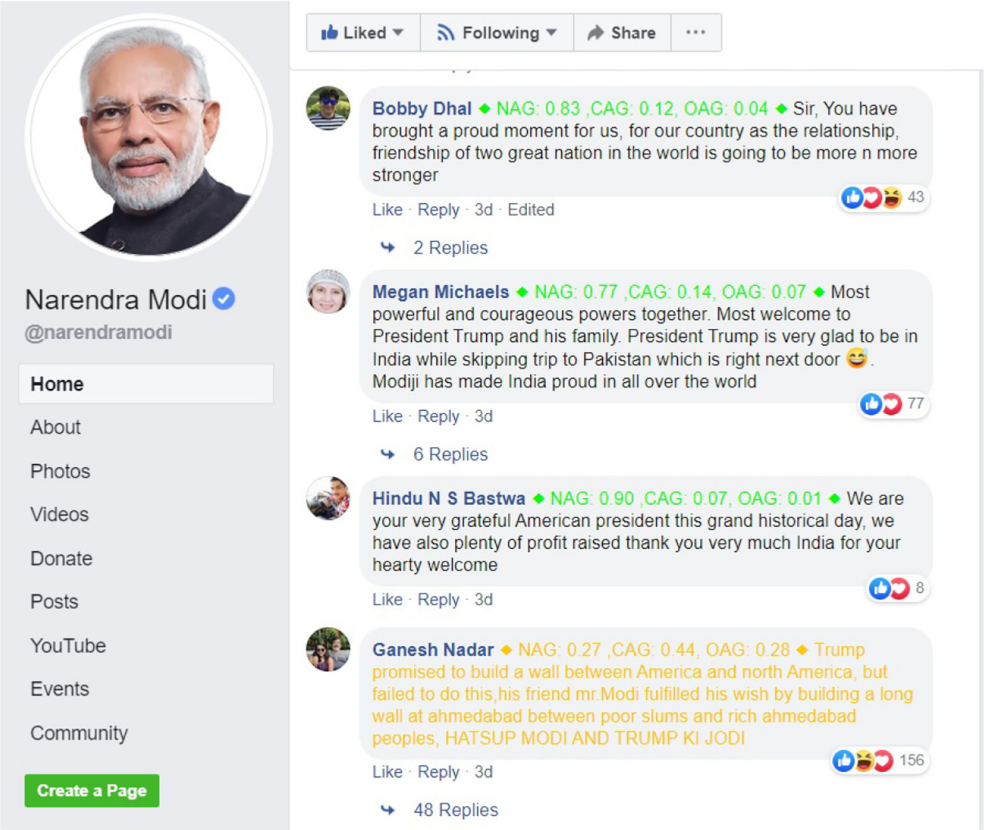
3.5. Sovereignty for individuals: classification of own input

Social media user often gives an opinion on the controversial topic discussed across the thousands of people. Some of the examples of such controversial topics are #Brexit, #Refugeecrisis, #tradewar. The fine line between maintaining freedom of speech and blocking hate speech is highly debatable in society and needs more constructive discussion. Our web user interface, as shown in Fig. 6 helps the user to ensure that their critical comments do not fall into any hate speech or aggressive content category. In this way, our web UI helps the user to uphold its information sovereignty on social media.

4. Proposed approaches for the classification

Most of the approaches available in the literature for hate speech detection are based on supervised learning. Machine learning approaches present a limitation concerning the learning process. They work on a specific data sample that has been annotated according to rules representing a specific human bias (Davidson et al., 2019). These classifiers may not provide accurate predictions on new and unseen expressions that are likely to appear in such a complex and subjective concept like hate speech. Human creativity can develop highly distinct patterns to express hate in many forms. Nevertheless, deep learning approaches are state of the art for such complex and ambiguous text classification tasks. Another limitation of the hate speech datasets lies in the distribution of problematic and non-problematic classes. The problematic classes, like hate, are more frequent in the datasets of all benchmark than in a realistic distribution (Wiegand et al., 2019). However, the classifiers cannot be trained if there are too few examples provided.

Initially, we set up classification experiments with traditional machine learning classifiers like Support Vector Machines (SVMs) and Logistic Regression to create the baseline results. After that,



Bobby Dhal ◆ NAG: 0.83 ,CAG: 0.12, OAG: 0.04 ◆ Sir, You have brought a proud moment for us, for our country as the relationship, friendship of two great nation in the world is going to be more n more stronger

Megan Michaels ◆ NAG: 0.77 ,CAG: 0.14, OAG: 0.07 ◆ Most powerful and courageous powers together. Most welcome to President Trump and his family. President Trump is very glad to be in India while skipping trip to Pakistan which is right next door 😅. Modiji has made India proud in all over the world

Hindu N S Bastwa ◆ NAG: 0.90 ,CAG: 0.07, OAG: 0.01 ◆ We are your very grateful American president this grand historical day, we have also plenty of profit raised thank you very much India for your hearty welcome

Ganesh Nadar ◆ NAG: 0.27 ,CAG: 0.44, OAG: 0.28 ◆ Trump promised to build a wall between America and north America, but failed to do this,his friend mr.Modi fulfilled his wish by building a long wall at ahmedabad between poor slums and rich ahmedabad peoples, HATSUP MODI AND TRUMP KI JODI

Fig. 3. Web Browser plugin on Narendra Modis' Facebook Page.

experiments are performed using deep learning models based on Convolutional Neural Network (CNN), Bidirectional LSTM with attention, and the recently proposed Bidirectional Encoder Representations from Transformers (BERT) transformer (Devlin et al., 2018).

4.1. Traditional classifiers

Logistic Regression and Support Vector Machines (SVMs) are popular conventional classifiers used in the various text classification work (Malmasi and Zampieri, 2018; Davidson et al., 2017). From our previous work (Modha and Majumder, 2019), Support Vector Machines (SVMs), and Logistic regression are found to be the best classifiers among all conventional classification algorithms. Our feature vector includes word unigrams with tf-idf weight, char 5-grams, length of the post, and no special characters. Tables 4 and 5 show results on the Logistic Regression and Support Vector Machines (SVM).

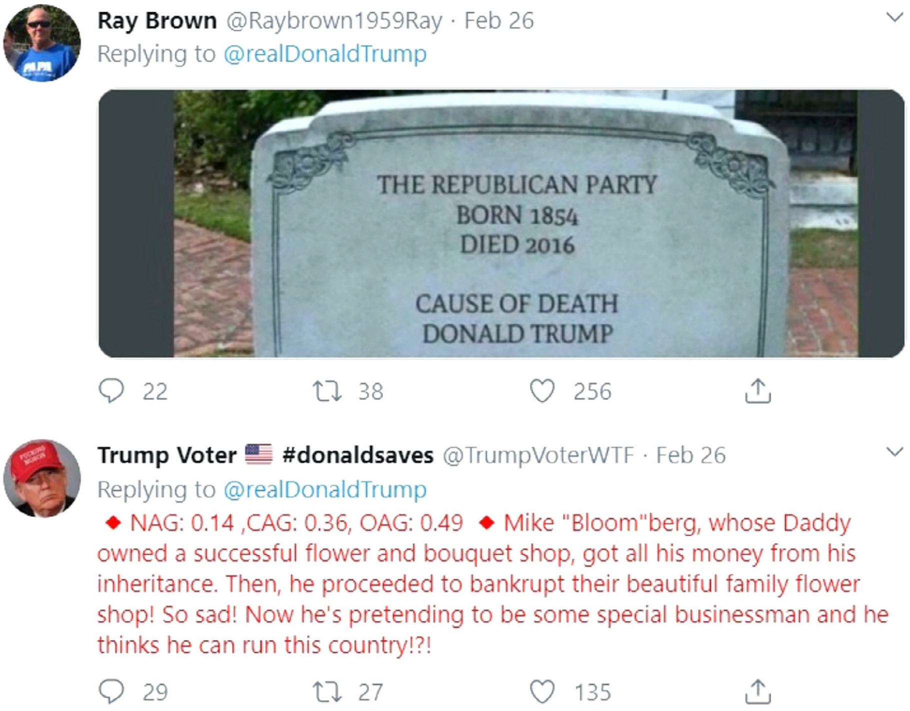
4.2. Deep learning model based on CNN

Deep learning models with distributed word representation are a widely used approach for text classification (Badjatiya et al., 2017; Modha et al., 2018). We did not perform any sort of Facebook or Twitter-specific text pre-processing or stemming on the text. The restrictions imposed on the length of a tweet by Twitter trigger a large variety of short and irregular forms introduced to tweets, which leads to the data sparsity problem (Saif et al., 2012). Because of the sparsity problem, a classifier based on BoW features might not be appropriate as compared to distributed word representations. This model includes one embedding layer whose weights are initialized with a fastText pre-trained vector with 300 dimensions (Mikolov et al., 2018). The word vectors were trained on 600 billion tokens of the Common Crawl corpus (Simonite, 2013). The Common Crawl is a nonprofit organization that crawls the web and freely provides its archives and datasets to the public. The freely available Common Crawl[1] corpus consists of petabytes of data collected using 9 years of web crawling. The embedding layer is followed by a one-dimensional convolution layer with 100 filters of height 2 and stride 1 to target character bigrams. In addition to this, a Global Max Pooling layer is added to obtain the maximum value from the filters, which are fed to the fully connected hidden layer with size 256. At last, the output layer follows. The ReLU and the softmax activation functions are used within the hidden layer and the output layer, respectively.

4.3. Deep learning model with attention

Dzmitry et al. (2014) proposed attention mechanisms in machine translation using neural machine translation. Yang et al. (2016) used attention mechanisms for the text classification. Authors claimed that some of the words in the sentence or posts are responsible for determining the correct label of the post. Bag-of-Words (BoW) is the traditional text representation techniques to model the text numerically. tf-idf and count-vector are the most popular text representation technique based on BoW. It can capture the important word but the lost context or ignore the sequential structure of the text while the deep neural model considers word order but failed to give a higher preference to the important word. The attention mechanism addresses both these issues while forming the sentence vector.

[1] https://commoncrawl.org.



Ray Brown @Raybrown1959Ray · Feb 26
Replying to @realDonaldTrump

Trump Voter 🇺🇸 #donaldsaves @TrumpVoterWTF · Feb 26
Replying to @realDonaldTrump

◆ NAG: 0.14 ,CAG: 0.36, OAG: 0.49 ◆ Mike "Bloom"berg, whose Daddy owned a successful flower and bouquet shop, got all his money from his inheritance. Then, he proceeded to bankrupt their beautiful family flower shop! So sad! Now he's pretending to be some special businessman and he thinks he can run this country!?!

Fig. 4. Web browser plugin for Twitter.

Our attention-based model consists of two LSTM layers as a forward layer and a backward layer. The model includes an embedding layer with 300 dimensions. It converts each word from the post into a fixed-length vector. Short posts are padded with zero values. Subsequent layers include 2 LSTM layers with 128 memory units, respectively. They are followed by the attention layer and dense layer with a size of 64 units and an output layer with softmax activations. The ReLU activation function is used for the hidden layer activation. The hyper-parameters are set as follows: the sequence length is fixed at 1073 words; that is the maximum length of posts in the dataset. The number of features is equal to half of the total vocabulary size. The models are trained for 10 epochs with batch size 128. Adam optimization algorithm is used to update network weights. The attention mechanism uses an average of the encoded state's output by the LSTM. But all of the encoded states of the LSTM are equally valuable. Thus, we are using a weighted sum of these encoded states to make our prediction. Fig. 1 shows our model architecture with an attention mechanism.

4.4. BERT

Devlin et al. (2018) proposed the Bidirectional Encoder Representations from Transformers (BERT), which has achieved significant improvement in various NLP tasks. The biggest challenge of the NLP task is the unavailability of the labeled data. The main objective behind BERT is to address this issue. BERT is based on a language model trained on a large corpus. It considers the previous and next tokens when predicting the next word as opposed to traditional language models, which only consider the previous n tokens and predict the next one. BERT can generate different vector representations of the same word that might appear in multiple contexts. BERT creates vectors on-the-fly by passing text into the model instead of dictionary lookup. Originally, BERT embeddings were trained with a next sentence prediction task and trained with large amounts of text. As with any pre-trained model, word representation using BERT also depends on the training dataset. It may not be a good representation for a specific subset of the language used in social media. It can not account for unexpected usage scenarios that are typical for the dynamics of language.

We have fine-tuned pre-trained BERT representations for the text classification task (Sun et al., 2019). BERT is based on a transformer architecture for encoding the text. Often, it performs better for small training datasets. Many variants of pre-trained BERT models are available. We have used the uncased BERT base model with 12 layers and 110 million parameters to classify the aggression on the English dataset. While for the Hindi dataset, the multilingual version of BERT is used with 110 million parameters. The hyper-parameters are set as follows: Sequence length is 128. The model is trained for three epochs with a batch size of 32, and the learning rate is set to 0.00002.

5. Evaluation and results

We have evaluated our classifier models on two different social media datasets: (i) Trolling Aggression and Cyberbullying (TRAC) dataset (ii) Comments which are posted on political leaders' Facebook pages and Twitter posts. The second dataset is created using our proposed web browser plugins.

5.1. TRAC dataset

The TRAC (Trolling, Aggression, and Cyberbullying) dataset consists of 15,001 aggression-annotated Facebook posts and comments in Hindi (Romanized and Devanagari script) and English (Ritesh et al., 2018). Tables 2 and 3 show a detailed description of the datasets. The weighted F1 measure has been used for evaluation. However, the interpretation should take into account that the F1 measure also can introduce bias for unbalanced datasets. (see (Powers, 2011) for a discussion).

Tables 4 and 5 presents results on TRAC dataset (Ritesh et al., 2018) on various classifier discussed in the previous section. A

Table 2
TRAC English Dataset statistics.

Class	# Training	# Facebook test	# Twitter test
NAG	6285	630	483
CAG	5297	142	413
OAG	3419	144	361
Total	15001	916	1257

Table 3
TRAC Hindi Dataset statistics.

Class	# Training	# Facebook test	# Twitter test
NAG	2813	413	459
CAG	6115	362	381
OAG	6073	195	354
Total	15001	970	1194

model based on BERT has produced a good weighted F1-score, but it failed to outperform a CNN based deep neural model.

The deep neural model based on the CNN model (Team name DA-LD-Hildesheim, (Modha et al., 2018)) has achieved the best results for the TRAC Twitter Hindi Dataset (Ritesh et al., 2018). Based on the results reported in Table 4 and 5, we have deployed the CNN model on an AWS server to classify the aggressive posts.

5.2. Dataset sampled from Facebook/Twitter plugins

Facebook and Twitter Live are popular features that enable politicians, celebrities to broadcast their speeches delivered during public meetings to their followers. Fans, followers, and critics present their views by posting comments during the live broadcast. We have deployed our Facebook/Twitter plugin and extracted comments from the popular worldwide social media profiles, such as president Trumps' timeline. Table 6 shows results on the Facebook/Twitter dataset. The weighted F1-score is around 0.75. Out of 652 comments, 464 were annotated as non-aggressive, 100 were from the OAG category, and 88 annotated as CAG by human annotators. Our model performs better for posts belonging to the NAG category.

The weighted F1 score on Facebook/Twitter dataset is around 0.75, which is substantially higher than the TRAC dataset, where the weighted F1 score is around 0.64. It is important to note that posts belonging to NAG classes are much higher than CAG, and OAG classes in the Facebook/Twiter live datasets compared to the TRAC dataset. Therefore, one can infer that the model may be able to perform better in predicting non-aggressive (NAG) messages rather than classifying overall well. It needs to be noted that the weighted F1 values can not be easily compared. Although the weighted F1 measure balances the influence of the classes based on their frequency, the distribution of aggressive and non-aggressive content is very different in the two datasets.

5.3. Result analysis

In this subsection, we will present an analysis of the experiments carried out in the previous sub-section. Overall, the CNN model with pre-trained fastText word embeddings marginally outperform traditional classifiers with respect to weighted F_1- score on the Facebook English corpus, The CNN model, substantially outperforms traditional classifiers on the English Twitter corpus. Similar results are observed on code-mixed Hindi corpus as shown in Tables 4, and 5. We can conclude that the CNN classifier tackle bias, which is introduced by the training corpus, better than traditional classifiers.

Table 7 presents a detailed comparative analysis of the results of two classifiers: the CNN model with fastText pre-trained vectors (Mikolov et al., 2018) and the Logistic Regression model with tf-idf weighting on the English TRAC Facebook dataset. The CNN model classifies Facebook posts better than the Logistic regression at the individual class level and overall. It can be observed that posts belonging to the CAG class are more difficult to classify. Table 8 shows some example tweets which belong to the CAG class and incorrectly classified under the NAG class. Table 9 shows the comparison with the peers (Ritesh et al., 2018).

6. Cyber watchdog: interface for aggression visualization on social media

This paper focuses on the real-time visualization of aggressive social media posts. We have developed a web browser plugin that runs inside the browser, read and send the real-time comments posted on any user's Facebook or Twitter timeline to the server using REST API. The server contains a classifier based on the CNN model and classify comment into NAG, CAG, and OAG categories and revert the labels to the plugin. Plugin rendered the comment according to the aggression predicted by the model deployed on the server. Fig. 2 shows the comment rendered by the plugin posted on president Donald trumps' Facebook page. Similarly, Fig. 3 shows the Indian Prime minister Narendra Modis' Facebook Page. Overtly aggressive (OAG) comments are rendered in red color, covertly aggressive comments are rendered in yellow color, and non-aggressive comments are flagged by a green diamond by the web browser plugin and the color of the comment's text remains unchanged. Fig. 4 shows the Twitter plugin that visualizes aggressive content on Twitter.

6.1. User interface architecture

In this section, we will describe our user interface deployed over Facebook, Twitter, and a standalone Interactive Web UI. Fig. 5 shows the detailed architecture of the system. We will describe each of the components of the UI in the following subsections.

6.1.1. Web browser plugin

A plugin is a program component that runs over another program like a web browser to add the specific functionality to it. As soon as the Facebook post is loaded into the web browser, a plugin reads the comments and sends them to the API server. The API server returns the label of the comments predicted by the aggression classifier model. The green diamond is used for the NAG label, the

Table 4
Results on the TRAC English Dataset (P-Precision, R-Recall, F1:weighted F1-score).

Model	Facebook English			Twitter English		
	P	**R**	**F1**	**P**	**R**	**F1**
CNN with fastText	0.6996	0.6091	**0.6407**	0.5664	0.5911	0.5520
BiLSTM with attention	0.7059	0.5098	0.5476	0.5766	0.5839	**0.5782**
BERT	0.7156	0.5840	0.6184	0.5623	0.5807	0.5683
Logistic Regression	0.6811	0.5710	0.6046	0.5232	0.5179	0.4890
Support Vector Machine	0.6795	0.5524	0.5902	0.4899	0.4924	0.4853

Table 5
Results on the TRAC Hindi Dataset (P-Precision, R-Recall, F1:weighted F1-score).

Model	Facebook Hindi			Twitter Hindi		
	P	R	F1	P	R	F1
CNN with fastText	0.6261	0.6124	0.6081	0.5206	0.4975	**0.4992**
BiLSTM with attention	0.5904	0.5690	0.5651	0.4861	0.4623	0.4639
BERT	0.6474	0.6216	**0.6182**	0.5101	0.3852	0.3353
Logistic Regression	0.6607	0.6185	**0.6133**	0.3779	0.3731	0.3723
Support Vector Machine	0.5998	0.5886	0.5861	0.3942	0.3874	0.3886

Table 6
Results on Dataset created using Facebook/Twitter plugin using CNN model.

Social Media	Precision	Recall	F1 (weighted)
Facebook	0.7941	0.7344	0.7549
Twitter	0.8116	0.7386	0.7593

red color is used for the OAG label, and the yellow color is used to render comments belonging to the CAG class. Figs. 2 and 3 shows a screenshot of the Facebook page rendered by the plugin. Similarly, Fig. 4 shows a screenshot of the Twitter page rendered by the plugin. We have recorded the screen while the plugin was marking comments on Facebook. The sample videos are uploaded on Youtube for Facebook [2] and for Twitter. [3]

6.1.1.1. Instructions to download the plugins. We have published our web browser plugins for Facebook and Twitter for the community. The Facebook plugin can be downloaded from this URL:https://bit.ly/2XTONqx and the URL for the Twitter plugin is:https://bit.ly/2KdMw1j. These plugins are tested on Google Chrome browser version id 81.0.4044.92 under Windows 10 and Ubuntu operating system. It has been observed that firewalls might block the plugin call to the server, and in such cases, the plugin might not work within the browser. Therefore, the user is requested to connect the Internet using mobile data. After adding the plugin in the browser, users can visit Twitter [4] and Facebook[5] for a quick view.

6.1.1.2. Disclosure. Once the user adds these plugins in the browser, then the data which will be rendered by our plugins are stored in the server. The purpose of the storing data is to create the weakly labeled dataset for future research. So, we request the user to use the plugin at her discretion.

6.1.2. API server

The API server is responsible for network communication. It handles requests and responses across different clients. The API server aggregates comments received from multiple clients and passes the text to the classifier model. The API server receives the predicted label from the model and sends it back to the client. The API server is deployed on an Amazon AWS cloud system.

6.1.3. Aggression classifier model

A deep neural model based on a convolutional neural network (CNN), as presented in Section 4.2, is deployed in the Amazon AWS cloud. The model classifies social media comment text into the three classes, namely: NAG, OAG, and CAG. The system returns these labels to the API server. All the comments, along with predicted labels, are stored in the MySQL database to prepare a weakly labeled training dataset for future research.

6.1.4. Web user interface

In addition to the web browser plugins, we developed an interactive, standalone web user interface (UI) where a user can input text, and the predicted label will be displayed on the web UI. The web UI can be accessed through this link: http://3.16.1.236:8000. Fig. 6 shows a snapshot of this web UI.

6.2. Social media dashboard for hate visualization

We have developed a PoC (proof of concept) for a personalized standalone dashboard for Twitter and Instagram. Demonstration versions of such PoCs can be found on these YouTube link[6].[7]

6.3. Applications

Facebook and Twitter are very popular across different age groups of users. They give them a real-time platform to disseminate their opinion about current events. Many celebrities regularly use these social media to connect with their followers. But at the same time, they might become victims of trolling and hate speech. In the following subsection, we will discuss the potential application of our proposed UI.

6.3.1. Social media surveillance

Many countries have passed laws against criminal hate speech and offensive content. Agencies are installed to monitor content and to act on complaints. On many occasions, persons posting criminal hate speech posts might not be prosecuted by these agencies in case the victim does not report an incident. Using our plugins, the security establishments and also social media companies themselves can monitor a particular set of users' timelines and trending hashtags to identify aggressive posts.

Post1: If you have nothing to hide, release the transcripts you corrupt b*sterd!.

Post2: Trump is worried about Joe . . .that's why he's abusing his power of the Oval Office. Trump has been a lying cheating cr*ok his entire life.

The above tweets are highly abusive and offensive. Twitters' hate speech detector failed to capture such offensive posts (Kottasová, 2017) while our plugin quickly classifies them into overtly aggressive posts. Law establishment agency could use this label as a piece of prima facie evidence to judge posts and launch investigations.

6.3.2. Creation of weakly labeled data

The creation of the labeled data for any classification task is expensive and time-consuming. As we look at Fig. 5, the API server stores each comment which it received from the plugins along with the aggression label into a MySQL database. Our Facebook and Twitter plugins can be used to create weakly annotated multilingual training data for any classification task in a crowdsourcing

[2] https://www.youtube.com/watch?v=Ytg0ZRo5O0c.

[3] https://www.youtube.com/watch?v=DxNIJCxCWo4.

[4] https://twitter.com/realDonaldTrump.

[5] https://www.facebook.com/DonaldTrump/.

[6] https://www.youtube.com/watch?v=HJY0EOluzUk.

[7] https://www.youtube.com/watch?v=2xwDj9CvWus.

Table 7
Model Comparison: CNN vs Logistic Reg. TRAC Facebook English Test dataset.

Classes	CNN model			Logistic Regression			#Posts
	P	R	F1	P	R	F1	
NAG	0.86	0.64	0.73	0.83	0.60	0.70	630
CAG	0.28	0.46	0.35	0.23	0.54	0.32	142
OAG	0.42	0.61	0.50	0.46	0.39	0.42	144
Overall	0.70	0.61	0.64	0.68	0.57	0.60	916

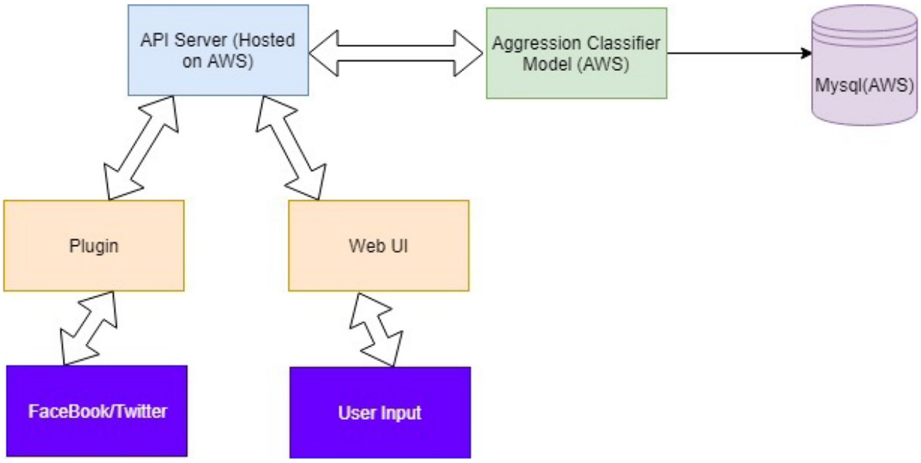
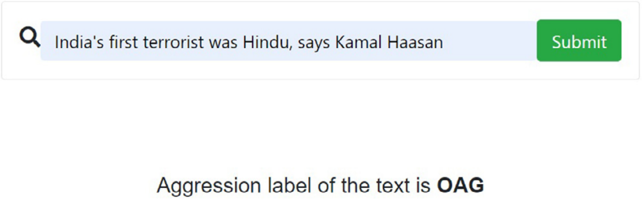


Fig. 5. User interface architecture.



India's first terrorist was Hindu, says Kamal Haasan

Aggression label of the text is **OAG**

Fig. 6. Web UI architecture.

manner. Thus, one can save time and money for the data annotation when a certain level of ambiguity is accepted.

6.3.3. Empowering citizens

Politicians or celebrities often use Facebook Live, or Twitter broadcasts to connect with their fans and followers. However, on many occasions, some of the users post abusive, offensive, sarcastic, or covertly aggressive comments responding to their posts. The same can happen to any user who writes about a controversial topic. Most users of social media will not like abusive or offensive content to be visible on their timeline when their profile is accessed. Our proposed plugin helps them to identify such comments by raising the appropriate flag, which enables them to delete or hide such abusive comments from their timeline using personalized Social media dashboards[8,9].

Using our plugin, ordinary users also have a tool which enables them to predict the offensiveness of text. Typically, such advanced AI tools are only available to large companies like Google or Facebook who develop such sophisticated technology and apply it for content moderation. Our tool gives every citizen the power to use the same kind of technology. They can use it to check text, which they want to enter into a social media system to predict the probability of it being criticized and moderated by the companies. Also, further tools can be developed based on the technology described, e.g., a system calculating an offensiveness profile for a particular user. Thus, the plugin alleviates technological imbalance and give citizens more sovereignty regarding the digital world.

Table 8
Examples of misclassified Tweets.

Text	Actual Class	Predicted Class
Yes yes ..traffic population pollution unlivability index ..bridging the gap between poor n middle class by bringing middle-class down	CAG	NAG
The name and the meaning has changed of RBI to reverse bank of India.	CAG	NAG

7. Conclusion and future work

In this paper, we have presented a user interface based on web browser plugins to visualize aggressive content expressed either implicitly or explicitly. The plugins are developed for the two most popular media: Facebook and Twitter. The work shows which kind of problems are moving into the center of attention for research in NLP area. Using deep learning models, there is great potential, yet still, the performance is far from perfect. The general complexity of language remains unsolved. There are no ideal text representations available yet. The use of pre-trained models like BERT can not solve all use cases perfectly. In particular, hate and aggression are dynamic concepts that evolve and for which users develop creative forms and language use patterns (Jaki et al., 2019). The social media-specific challenges like data sparsity, spontaneous use, and frequent violation of language norms make it hard to acquire sufficient training data.

[8] https://www.youtube.com/watch?v=2xwDj9CvWus.
[9] https://www.youtube.com/watch?v=HJY0EOluzUk.

Table 9
Weighted F1-score on TRAC Test Data: Peer comparisons (Ritesh et al., 2018).

System	English		Hindi	
	Facebook Dataset	Twitter Dataset	Facebook Dataset	Twitter Dataset
Our system result	**0.6407**	0.5541	0.6081	**0.4992**
DA-LD-hildesheim	0.6178	0.552	0.6081	**0.4992**
saroyehun	**0.6425**	0.5920	NA	NA
EBSI-LIA-UNAM	0.6315	0.5715	NA	NA
TakeLab	0.5920	0.5651	NA	NA
taraka rama	0.6008	0.5656	**0.6420**	0.40
vista.ue	0.5812	**0.6008**	0.5951	0.4829
na14	0.5920	0.5663	0.6450	0.4853

Social media posts should not be considered as a standalone text. It is interactive, and posts are context-dependent. Often, knowing at least the preceding posts is crucial to recognizing aggression. Again, future research needs to focus also on the nature of the discourse between users in social media. Existing research has modeled information flow between users by estimating entropy (Kwon et al., 2016) and conversational patterns as a discourse within social networks (Lipizzi et al., 2016). Such approaches need to be also applied to hate speech detection in the future.

The inter-annotator agreement of the current benchmark collections is at a moderate level might be the outcome of the high ambiguity and a lack of a commonly accepted definition of hate speech. Each annotator might have a different threshold level to decide on potential hateful and sensitive topics. The systems currently can not model this high level of disagreement. That makes it difficult to interpret the state of the performance reached by hate speech identification systems. Future collections need to find ways to incorporate divergent opinions. The accuracy of classification methods is not high, which means that users will be confronted with misclassified content. Also, there is further need for research on the visualization of the outcomes of text classification algorithms and the need to develop advanced systems providing more functions that include live relevance judgment on misclassified posts. We have implemented this functionality in the proposed PoC of the Twitter dashboard. The tool can be seen in a video.[10] In the future, we want to develop a separate personalized interface that aggregates all feeds from different social media platforms for a particular celebrity. Our proposed interface will have features like the visualization of hate and sentiment, and the identification and blocking of hateful individuals.

CRediT authorship contribution statement

Sandip Modha: Conceptualization, Methodology, Software, Writing - original draft. **Prasenjit Majumder:** Supervision, Validation. **Thomas Mandl:** Supervision, Writing - review & editing, Validation. **Chintak Mandalia:** Visualization, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We acknowledge Mr. Daksh Patel, Ms. Mohana Dave, and Mr. Devarshi Raval for the design of the front end of PoC for Twitter and Instagram dashboards.

References

Assimakopoulos, S., Baider, F. H., & Millar, S. (2017). *Online hate speech in the European Union: A discourse-analytic perspective*. Springer.

Badjatiya, P., Gupta, S., Gupta, M. & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on world wide web companion (pp. 759–760). International World Wide Web Conferences Steering Committee.

Bourgonje, P., Moreno-Schneider, J., Srivastava, A., & Rehm, G. (2017). Automatic classification of abusive language and personal attacks in various forms of online communication. In *International conference of the German society for computational linguistics and language technology* (pp. 180–191). Cham: Springer.

Braun, S. (2004). *Democracy off balance: Freedom of expression and hate propaganda law in Canada*. University of Toronto Press.

Culpeper, J. (2011). *Impoliteness: Using language to cause offence* (Volume 28). Cambridge University Press.

Davidson, T., Bhattacharya, D. & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In Proceedings of the third workshop on abusive language online (pp. 25–35). Florence, Italy: Association for Computational Linguistics. URL:https://www.aclweb.org/anthology/W19-3504. doi: 10.18653/v1/W19-3504.

Davidson, T., Warmsley, D., Macy, M. & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of ICWSM..

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dewan, P., & Kumaraguru, P. (2017). Facebook inspector (fbi): Towards automatic real-time detection of malicious content on facebook. *Social Network Analysis and Mining, 7*. URL:https://doi.org/10.1007/s13278-017-0434-5.

Dzmitry, B., Kyunghyun, C. & Yoshua, B. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*. 51, URL:https://doi.org/10.1145/3232676.

Greenwell, J., & Dengerink, H. (1973). The role of perceived versus actual attack in human physical aggression. *Journal of Personality and Social Psychology, 26*, 66–71. URL:https://doi.org/10.1037/h0034223.

Habermas, J. et al. (1984). *Preliminary studies and supplements to the Theory of Communicative Action*. Suhrkamp Frankfurt/ M.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research, 14*, 206–221.

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict, 7*, 240–268.

Kottasová, I. (2017). Europe gives Facebook, Twitter final warning on hate speech. URL:https://money.cnn.com/2017/09/28/technology/hate-speech-facebook-twitter-europe/index.html [Accessed: 19 April 2020].

Kurniasih, N., Abdillah, L. A., Sudarsana, I. K., Yogantara, I., Astawa, I., Nanuru, R. F., Miagina, A., Sabarua, J. O., Jamil, M., Tandisalla, J., et al. (2018). Prototype application hate speech detection website using string matching and searching algorithm. *International Journal of Engineering & Technology, 7*, 62–64.

Kwok, I. & Wang, Y. (2013). Locate the hate: Detecting Tweets Against Blacks. In Twenty-seventh AAAI conference on artificial intelligence.

Kwon, H., Kwon, H. T., & Yoon, W. C. (2016). An information-theoretic evaluation of narrative complexity for interactive writing support. *Expert Systems with Applications, 53*, 219–230.

Land, M. K. (2019). Against privatized censorship: Proposals for responsible delegation. Virginia Journal of International Law.

Lipizzi, C., Dessavre, D. G., Iandoli, L., & Marquez, J. E. R. (2016). Towards computational discourse analysis: A methodology for mining twitter backchanneling conversations. *Computers in Human Behavior, 64*, 782–792. URL:https://www.aclweb.org/anthology/N19-1060.pdf.

Lu, J., & Yu, X. (2020). Does the internet make us more intolerant? a contextual analysis in 33 countries. *Information, Communication & Society, 23*, 252–266.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One, 14*.

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence, 30*, 1–16.

[10] https://youtu.be/HJY0EOluzUk?t=60.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C. & Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation (pp. 14–17).

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (LREC)* (pp. 2018).).

Modha, S. & Majumder, P. (2019). An empirical evaluation of text representation schemes on multilingual social web to filter the textual aggression. arXiv preprint arXiv:1904.08770.

Modha, S., Majumder, P. & Mandl, T. (2018). Filtering aggression from the multilingual social media feed. In Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018) (pp. 199–207).

Pelzer, B., Kaati, L. & Akrami, N. (2018). Directed digital hate. In 2018 IEEE international conference on intelligence and security informatics (ISI) (pp. 205–210). IEEE.

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence, 48*, 4730–4742.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology, 2*, 37–63.

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian conference on artificial intelligence* (pp. 16–27). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ritesh, K., Kr., O. A., Shervin, M. & Marcos, Z. (2018). Benchmarking aggression identification in social media. In Proceedings of the first workshop on trolling, aggression and cyberbulling (TRAC) (pp. 1–11). Santa Fe, USA.

Ritesh, K., N., R. A., Akshit, B. & MaheshwariTushar (2018). Aggression-annotated corpus of hindi-english code-mixed data. In Proceedings of the 11th language resources and evaluation conference (LREC) (pp. 1–11). Miyazaki, Japan.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N. & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In Proceedings of the workshop on natural language processing for computer-mediated communication (NLP4CMC). Bochum, Germany.

Saif, H., He, Y. & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In #MSM2012 making sense of microposts. Proceedings of the WWW'12 Workshop on 'Making Sense of Microposts' Lyon, France, April 16, 2012. CEUR Workshop Proceedings (CEUR-WS. org). URL:http://ceur-ws.org/Vol-838/paper_01.pdf.

Salminen, J., Almerekhi, H., Kamel, A.M., Jung, S.-g., & Jansen, B.J. (2019). Online hate ratings vary by extremes: A statistical analysis. In Proceedings of the 2019 conference on human information interaction and retrieval (pp. 213–217). ACM.

Schmidt, A. & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. In Proceedings of the fifth international workshop on natural language processing for social media. Association for Computational Linguistics (pp. 1–10). Valencia, Spain.

Seglow, J. (2016). Hate speech, dignity and self-respect. *Ethical Theory and Moral Practice, 19*, 1103–1116.

Simonite, T. (2013). A free database of the entire web may spawn the next google. URL:https://www.technologyreview.com/s/509931/a-free-database-of-the-entire-web-may-spawn-the-next-google/ [Accessed: 19 April 2020].

Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019). How to fine-tune BERT for text classification? CoRR, abs/1905.05583. URL:http://arxiv.org/abs/1905.05583. arXiv:1905.05583.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

Warner, W. & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media (pp. 19–26). Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D. & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In Proceedings of the first workshop on abusive language online.

Wiegand, M., Ruppenhofer, J. & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 602–608).

Wiegand, M., Siegel, M. & Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In 14th Conference on Natural Language Processing. KONVENS 2018. Proceedings of the GermEval 2018 Workshop. Austrian Academy of Sciences, Vienna, September 21, 2018. URL:https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf.

Xu, J. -M., Jun, K. -S., Zhu, X. & Bellmore, A. (2012). Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 656–666). ACL.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). 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).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).