Benchmarking Aggression Identification in Social Media

Ritesh Kumar[1], Atul Kr. Ojha[2], Shervin Malmasi[3], Marcos Zampieri[4]
[1]Bhim Rao Ambedkar University, [2]Jawaharlal Nehru University,
[3]Harvard Medical School, [4]University of Wolverhampton,

Abstract

In this paper, we present the report and findings of the Shared Task on Aggression Identification organised as part of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) at COLING 2018. The task was to develop a classifier that could discriminate between *Overtly Aggressive*, *Covertly Aggressive*, and *Non-aggressive* texts. For this task, the participants were provided with a dataset of 15,000 aggression-annotated Facebook Posts and Comments each in Hindi (in both Roman and Devanagari script) and English for training and validation. For testing, two different sets - one from Facebook and another from a different social media - were provided. A total of 130 teams registered to participate in the task, 30 teams submitted their test runs, and finally 20 teams also sent their system description paper which are included in the TRAC workshop proceedings. The best system obtained a weighted F-score of 0.64 for both Hindi and English on the Facebook test sets, while the best scores on the surprise set were 0.60 and 0.50 for English and Hindi respectively. The results presented in this report depict how challenging the task is. The positive response from the community and the great levels of participation in the first edition of this shared task also highlights the interest in this topic.

1 Introduction

In the last decade, with the emergence of an interactive web and especially popular social networking and social media platforms like Facebook and Twitter, there has been an exponential increase in the user-generated content being made available over the web. Now any information online has the power to reach billions of people within a matter of seconds. This has resulted in not only positive exchange of ideas but has also lead to a widespread dissemination of aggressive and potentially harmful content over the web. While most of the potentially harmful incidents like bullying or hate speech have predated the Internet, the reach and extent of Internet has given these incidents an unprecedented power and influence to affect the lives of billions of people. It has been reported that these incidents have not only created mental and psychological agony to the users of the web but has in fact forced people to deactivate their accounts and in extreme cases also commit suicides (Hinduja and Patchin, 2010). Thus the incidents of aggression and unratified verbal behaviour have not remained a minor nuisance, but have acquired the form of a major criminal activity that affects a large number of people. It is therefore important that preventive measures can be taken to cope with abusive behaviour aggression online.

One of the strategies to cope with aggressive behaviour online is to manually monitor and moderate user-generated content, however, the amount and pace at which new data is being created on the web has rendered manual methods of moderation and intervention almost completely impractical. As such the use (semi-) automatic methods to identify such behaviour has become important and has attracted more attention from the research community in recent years (Davidson et al., 2017; Malmasi and Zampieri, 2017).

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 1–11
Santa Fe, USA, August 25, 2018.

This paper reports the results of the first Shared Task on Aggression Identification which was organised jointly with the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) at COLING 2018.

2 Related Work

Verbal aggression *per se* has been rarely explored within the field of Natural Language Processing. However, previous research in the field has been carried out to automatically recognise several related behaviour such as trolling (Cambria et al., 2010; Kumar et al., 2014; Mojica, 2016; Mihaylov et al., 2015) , cyberbullying (Dinakar et al., 2012; Nitta et al., 2013; Dadvar et al., 2013; Dadvar et al., 2014; Hee et al., 2015), flaming / insults (Sax, 2016; Nitin et al., 2012), abusive / offensive language (Chen et al., 2012; Nobata et al., 2016; Waseem et al., 2017), hate speech (Pinkesh Badjatiya and Varma, 2017; Burnap and Williams, 2014; Davidson et al., 2017; Vigna et al., 2017; Djuric et al., 2015; Fortana, 2017; Gitari et al., 2015; Malmasi and Zampieri, 2018; Waseem and Hovy, 2016; Schmidt and Wiegand, 2017), radicalization (Agarwal and Sureka, 2015; Agarwal and Sureka, 2017), racism (Greevy and Smeaton, 2004; Greevy, 2004) and others. In addition to these, there have been some pragmatic studies on behaviour like trolling (Hardaker, 2010; Hardaker, 2013).

This huge interest in the field from different perspectives has created a conglomeration of terminologies as well as understandings of the phenomenon. On the one hand, this provides us with a very rich and extensive insight into the phenomena yet, on the other hand, it has also created a theoretical gap in the understanding of interrelationship among these. Moreover, it has also resulted in duplication of research, to certain extent, and a certain kind of lack of focus and reusability of datasets across different strands of research. In order to make improvements towards solving a complex phenomenon like this, it is of utmost importance that some kind of uniform understanding of problem be achieved so that, at least, standardised datasets and an understanding of different approaches to solving the problem may be developed.

While a large part of the research has focused on any one of these phenomena and their computational processing, it seems there is a significant overlap among these phenomenon in the way they are understood in these studies - and because of this underlying overlap, insights from different studies might prove useful for solving these seemingly different phenomena. All of these behaviours are considered undesirable, aggressive and detrimental for those on the receiving end. So, trolling is intended "to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement" (Hardaker, 2010). Cyberbullying is "humiliating and slandering behavior towards other people" (Nitta et al., 2013). Flaming intends "to offend someone through e-mail, posting, commenting or any statement using insults, swearing and hostile, intense language, trolling, etc." (Krol, 1992).

Waseem et al. (2017) makes an attempt to unify these different trends of research in what may be considered a significantly overlapping field and proposes a 2-way typology for understanding what they call 'abusive language' over the web. They propose 2 scales on which abusive language could be categorised - the target of the abuse (an individual or a group) and the nature of the language (explicit or implicit). Our classification of aggression into overt and covert aggression is largely similar to the explicit-implicit distinction. However, we make a more detailed distinction in relation to the target of the abuse (Kumar et al., 2018b) and it is not made along the axis of individual vs. group. This is so because we noticed in a large number of instances both individuals and groups are simultaneously targeted - in such cases individuals are targeted as members of certain groups or the individuals' actions were considered those of the group and became the locus of attack. As such it was not feasible to distinguish between the individual and group attack in lot of instances while annotating the dataset. The distinction that we made was related to the "locus" of attack and included such targets as gender, religion, caste, country of origin, race, etc. This classification, on the one hand, gave scope for focusing on different kinds of attack (for example, racial attacks or communal attacks) and, on the other hand, each of these targets may actually be attacked using a different set of vocabulary, thereby, making these more natural classes that could be classified using the surface-level linguistic features. Of course, it cannot be denied that these targets are not mutually exclusive and, as such, it makes the problem not just a multi-class classification problem but also multi-label classification problem. In addition to this, we also make use of a different terminology taking into account its use within socio-pragmatics. This was done with an understanding that huge amount of literature within the field of aggression and impoliteness studies might be able to contribute and provide insights to understanding the phenomenon in a better way.

The aim of this shared task was much simpler than the one discussed in the previous para. It only involved classification of the texts into 3 categories - overt aggression, covert aggression and non-aggression. We wanted to use the dataset for experimenting with different approaches to make the most top-level classification of aggression on social media.

3 Task Setup and Schedule

The participants interested in competing in the shared task were required to register using a Google Form. The form gave them an option to participate for either English or Hindi or both the languages. All the registered participants were sent the links to the annotated dataset in the language(s) of their choice, along with a description of the format of the dataset. The participants were allowed to use additional data for training the system, with the condition that the additional dataset should be either publicly available or make available immediately after submission (and well before the submission of the system papers) and this must be mentioned in the submission. Use of non-public additional data for training was not allowed. The participants were given around 6 weeks to experiment and develop the system. However, since more than half of the participants registered after the first release of the data, most of them got less time than this. Initially, the dataset was not released publicly but was emailed only to the registered participants. After the 6 weeks of release of train and dev sets, the test set was released and the participants had 5 days to test and upload their system. The complete timeline of the shared task is given in Table 1. We made use of CodaLab [1] for the evaluation. Each team was allowed to submit up to 3 systems for evaluation. We used the best of the 3 runs for the final ranking and evaluation of the systems.

Date	Event
1 February, 2018	Shared Task Announcement and Start of Registration
13 March, 2018	Release of train and dev sets
25 April, 2018	Release of test set
30 April, 2018	Deadline for Submission of System
2 May, 2018	Declaration of Results
28 May, 2018	Deadline for Submission of System Description Paper

Table 1: Timeline of the Aggression Identification Shared Task at TRAC - 1.

4 Dataset

The participants of the shared task were provided with a dataset of 12,000 randomly sampled Facebook comments for training and 3,000 comments for development and in English and Hindi each, annotated with 3 levels of aggression - Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-Aggressive (NAG). For test, 916 English comments and 970 Hindi comments were provided. Additionally, 1,257 English tweets and 1,194 Hindi tweets were given as the surprise test set [2]. The dataset released for the task is a subset of a larger dataset discussed in Kumar et al. (2018b).

4.1 Issues with the Dataset

While most of the participants considered the dataset to be of high quality, two major problems came up during the task -

- **The language issue**: Some of the comments in English dataset contained code-mixed Hindi-English data as well as data from other languages like German. These formed a minuscule proportion of the data but nevertheless these need to be filtered out.

[1] https://competitions.codalab.org/

[2] The complete dataset used for the shared task can be downloaded here - http://trac1-dataset.kmiagra.org/

- **The annotation issue**: The second and more serious issue that was raised by some participants is related to the the annotation itself. Several instances of supposedly inaccurate annotation were pointed out. Despite the fact that aggression is a highly subjective phenomenon and different annotators may have different judgments about the same comment, some of the annotation indeed looked highly implausible and consequently it needs further scrutiny and validation.

5 Participants and Approaches

The shared task gave the participants an option to register for either one of the two languages - English or Hindi - or both. A total of 131 participants registered for the shared task, with 73 teams registering to participate only in English track, 2 teams only in Hindi track and 56 teams registered to participate in both the tracks. Out of these, finally a total of 30 teams submitted their systems - 15 teams for both English and Hindi and 30 teams for only English track. All the systems who submitted their system were invited to submit the system description paper, describing the experiments conducted by them. 18 participants submitted the final description paper which are included in the workshop proceedings - it included papers by majority of the top 10 teams. Table 2, lists the participating teams and the language they took part in.

Team	Hindi	English	System Description Paper
saroyehun		✓	(Aroyehun and Gelbukh, 2018)
EBSI-LIA-UNAM		✓	(Arroyo-Fernández et al., 2018)
DA-LD-Hildesheim	✓	✓	(Modha et al., 2018)
TakeLab		✓	(Golem et al., 2018)
sreeIN		✓	(Madisetty and Desarkar, 2018)
Julian	✓	✓	(Risch and Krestel, 2018)
taraka_rama	✓	✓	
uOttawa		✓	(Orabi et al., 2018)
Isistanitos		✓	(Tommasel et al., 2018)
hakuchumu		✓	
DataGeeks	✓	✓	
na14	✓	✓	(Samghabadi et al., 2018)
dinel		✓	(Orasan, 2018)
vista.ue	✓	✓	(Raiyani et al., 2018)
MANITBHOPALINDIA	✓	✓	
IRIT		✓	(Ramiandrisoa and Mothe, 2018)
quine	✓	✓	(Nikhil et al., 2018)
IIIT-Delhi		✓	
PMRS	✓	✓	(Maitra and Sarkhel, 2018)
resham	✓	✓	
IreneR		✓	
Nestor	✓	✓	
UAEMex-UAPT1	✓	✓	
forest_and_trees		✓	(Galery et al., 2018)
groutar		✓	(Fortuna et al., 2018)
Shusrut	✓	✓	(Roy et al., 2018)
malaypramanick		✓	
UAEMex-UAPT-TAC2	✓	✓	
Unito	✓	✓	
bhanodaig		✓	(Kumar et al., 2018a)
Total	**15**	**30**	**18**

Table 2: The teams that participated in the Aggression Identification Shared Task at TRAC - 1.

Next we give a short description of the approach taken by each team for building their system. More details about the approaches could be found in the paper submitted by the respective teams.

- **saroyehun** system gives the best performance with LSTM and they resorted to translation as data augmentation strategy. With the surprise twitter set, a combination of the representations of the RNN and CNN as features, along with additional preprocessing like spelling correction, translation of emoji, and computation of sentiment score gave the best performance. In this case, the dataset was also augmented using translation and pseudolabelled using an external dataset on hate speech.[3] This is the only approach in the competition that performs better on the Twitter dataset, despite being trained the Facebook dataset, thereby, depicting the ability of the approach to generalise across domain.

- **EBSI-LIA-UNAM** system uses a combination of the Passive-Aggressive (PA) and SVM classifiers with character based n-gram (1 - 5 grams) TF-IDF for feature representation.

- **DA-LD-Hildesheim** uses LSTM with pretrained Fasttext vector for embeddings for classifying English Facebook texts. For all other datasets including Twitter data in English and both Facebook and Twitter dataset in Hindi, CNN performs better.

- **TakeLab** uses a Bidirectional LSTM on Glove embeddings to give the best performance.

- **sreeIN** system uses a voting-based ensemble method with 3 classifiers - CNN with 4 layers, LSTM and Bidirectional LSTM.

- **Julian** team uses translation as data augmentation strategy and use an ensemble of TF-IDF based approaches, using character n-grams (2 - 6) and word n-grams (1 - 2) with a bi-directional RNN, using fasttext embeddings, to get the best performance in the task..

- **taraka_rama** uses different systems for different datasets. For English Facebook dataset and Hindi Twitter dataset, the team uses a stacked ensemble classifier that uses a SVM on top of the ensemble of SVM classifiers. The SVMs were trained on 1 - 6 character n-grams and word unigrams. For Hindi Facebook and English Twitter dataset, however, a plain SVM trained using character and word bag-of-n-grams gave the best performance. In this case, the overlapping character and word n-gram features are weigthed with sublinear tf-idf before being used for training and testing. The system is tuned using 5-fold CV on the combined training and develpment sets for maximum number of character and word n-grams included, case normalization, and SVM margin (regularization) parameter C.

- **uOttawa** system is trained using a novel deep-learning architecture for text classification based on Multi-task learning (MTL). The approach, MTL, is evaluated using three neural network models. MultiCNN, multiple convolution structure with a trainable embedding layer, gives the best performance.

- **Isistanitos** system uses a soft voting (average the class probabilities of other models) of two models - a recurrent neural network, and an SVM. The recurrent neural network uses 3 preprocesed set of features. The first set uses an ad-hoc glove model for representing the words, the second is a sentiwornet based model, and the third is a traditional TfIdf plus Vader Sentiment analysis and sentiments associated with the emojis. The SVM model is trained on a TF-IDF of the post stemmed terms, excluding stopwords, and 3 - 5 character n-grams.

- **hakuchumu** system makes use of a Random Forest classifier with some preprocessing including removal of urls and non letter characters and stop words. Along with the bag-of-word, the approach uses multiple occurrences of letters, exclamation marks and question marks in a row and emoticons as binary features.

- **DataGeeks** system uses Logistic Regression classifier with some preprocessing on the data such as removing non-ascii characters, replacing new line with '.', replacing n't with not, removing stopwords and 1 - 3 word n-grams and 2 - 6 character n-grams for training the classifier.

[3] `https://github.com/ZeerakW/hatespeech`

- **na14** also uses Logistic Regression classifier with preprocessing involving replacing URLs, numbers, email addresses and spelling correction. The classifier is trained using word unigrams, tf-idf vectors of word unigram, character 4-gram, character 5-gram and Google news pre-trained word embedding model. For the Hindi dataset, Devanagari texts were transliterated into Roman at the preprocessing stage.

- **dinel** achieves the best accuracy on the Facebook test set using a Random Forest classifier while SVMs performed better for the surprise Twiiter test set. Both the classifiers were trained using 300 semantic features which represent the vector representation of the text, average scores of the top emojis for each of the classes and positive and negative sentiment scores.

- **vista.ue** system is developed using dense neural networks.

- **MANITBHOPALINDIA** system for English is developed using SVM while for English it is trained using deep neural networks.

- **IRIT** system gets the best performance for the English Facebook test set by using a combination of two models - a doc2vec model and a logistic regression classifier. For the Twitter test set, it uses a combination of CNN and LSTM to get the best performance.

- **quine** system is trained using an LSTM with attention and simple embeddings (word to index) instead of pre-trained embeddings.

- **IIIT-Delhi** system uses a Single channel CNN for this task. Bayesian Optimization is used for tuning the parameters.

- **PMRS** system employs a winner-takes- all autoencoder, called Emoti-KATE for Twitter sentiment classification. Each input dimension of Emoti-KATE is a log-normalized, sentiwordnet-score weighted word-count vector. A binary cross-entropy loss function is used to train the network.

- **resham** system for English has been made using an open vocabulary approach and ensemble model of two predictors with soft voting. The first predictor is a Naive Bayes model with CountVectorizer for preprocessing. The second predictor is a recurrent neural network with one embedding layer and two LSTM layers. Pre-trained word vectors have been used for the embedding layer. For Hindi dataset, a Naive Bayes classifier is trained using the dataset augmented with English translations.

- **IreneR** system is based on a Multinomial Naive Bayes classifier that uses unigrams, bigrams, hedging bigrams and trigrams such as 'do you', someone who is','to see that', that potentially signal covert aggressivity, identified with chi-squared test as features. It also includes features from LIWC2015 (list of anger and swear words).

- **Nestor** uses an approach that combines Neural Networks and a new word representation model. The patterns obtained from the word model representation are used for training the back propagation neural network with fix parameters. The length of the post was fixed and the word model representation is language independent, so it was used for both the English and the Hindi tasks.

- **UAEMex-UAPT1** uses the same approach as used by the team Nestor.

- **forest_and_trees** system uses a Pooled Recurrent Unit architecture combined with pre-trained English and Hindi fasttext word embeddings as a representation of the sequence input. In this approach, Hindi and English vectors were aligned using pre-computed SVD matrices that pulls representations from different languages into a single space. This enabled the same model to be used for both the languages, thereby, making data re-utilization and model deployability easier.

- **groutar** system is trained using random forests. The dataset is augmented with an external toxicity dataset [4]. The approach involved understanding the effects of new data on aggression identification.

[4] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

- **Shusrut** system uses an ensemble of CNN 2D with MAXPOOL classifier and a SVM classifier. The ensemble model is passed through 3 dense layers to finally predict the output. Softmax activation is used in the outer layer for classification.

- **malaypramanick** system uses a random forest classifier trained using a set of surface-level features including number of line,s uppercase and lowercase letters, digits, named entities, unicode characters, etc.

- **UAEMex-UAPT-TAC2** system is generated by combination of twelve distance measures, through a K Nearest Neighbors classification algorithm and a canonical genetic algorithm.

- **Unito** is the only unsupervised system submitted in the task. It is based only on a multilingual lexicon of aggressive words. The lexicon is obtained by automatic translation from an handmade lexicon of offensive words in Italian, with minimal human supervision. The original words are expanded into a list of their senses. The senses are manually annotated to filter out senses that are never used in an offensive context. Finally, all the lemmas of the remaining senses are generated with BabelNet in 50+ languages. The words in the lexicon are divided in those translating sense that can be used in an offensive context (but not necessarily are) and words translating senses that are directly offensive. This distinction is mapped to the Overtly Aggressive and Covertly Aggressive classes respectively. The classification of sentences is straightforward: a sentence that does not contain any word from the lexicon is tagged as NAG, a sentence containing more directly offensive words than potentially offensive words is tagged as OAG, and the other cases are tagged as CAG.

- **bhanodaig** system uses a bidirectional LSTM.

6 Results

In this section, we present the results of the experiments carried out by different teams during the shared task. The results of the top 15 teams on English dataset is given in Figure 1 and that on Hindi dataset is in Figure 2.

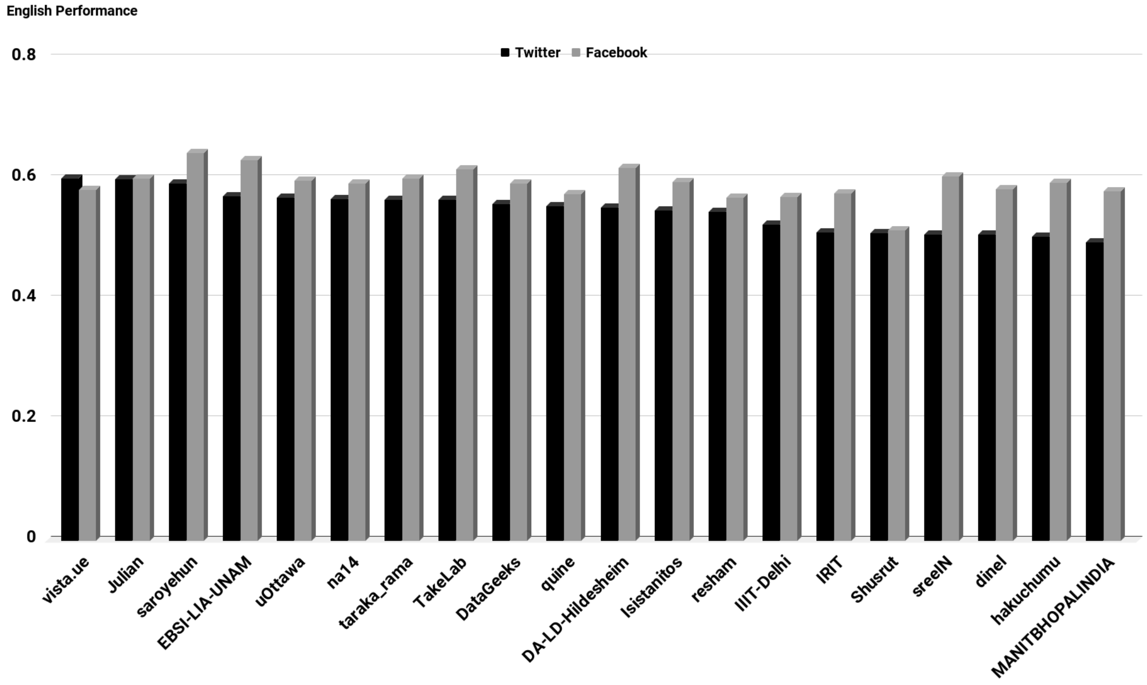

Figure 1: Performance of top 15 teams on English Dataset

The participants were allowed to use other datasets, in addition to the one provided by the organizers of the task. However, because of the lack of similar alternative datasets, all the groups, except ’groutar’ and ’saroyehun’ team, used only the dataset provided for the task. As we mentioned earlier, the participants were given two kinds of test sets for the final testing of the system - one from Facebook and a surprise test set from Twitter.

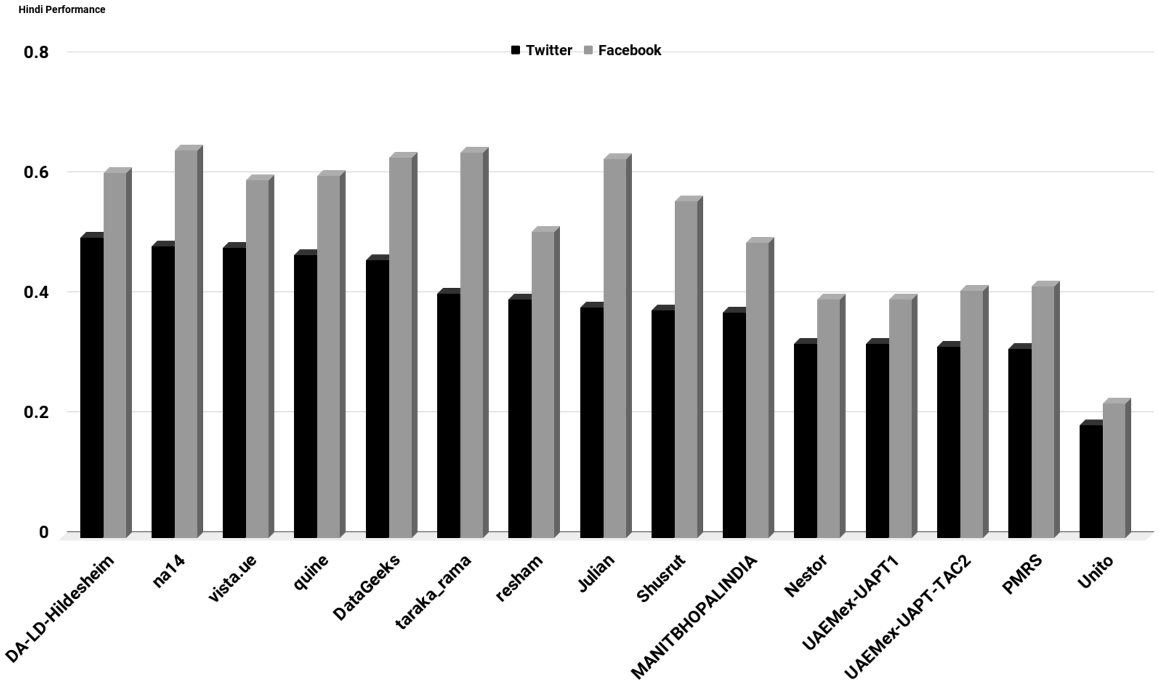

Figure 2: Performance of teams on Hindi Dataset

7 Conclusion

In this paper, we have presented the report of the First Shared task on Aggression Identification organized with the TRAC workshop at COLING 2018. The shared task received a very encouraging response from the community which underlines the relevance and need of the task. More than 100 teams registered and 30 teams finally submitted their system.

The performance of the best systems in the task show that aggression identification is a hard problem to solve. Moreover, the performance of the neural networks-based systems as well as the other approaches do not seem to differ much. If the features are carefully selected then classifiers like SVM and even random forest and logistic regression perform at par with deep neural networks. On the other had, we find quite a few neural networks-based systems not performing quite well in the task. Nonetheless, 14 systems were trained using one or the other architectures of deep neural networks - either solely or as part of an ensemble. Moreover, 8 systems out of the top 15 are trained on neural networks, which shows the efficacy of the approach but at the same time does not rule out the usefulness and relevance of linear models for the task. There was only one system, Unito, that made use of a lexicon-based approach to solve the task. A few participants of the task pointed out the apparent “inconsistencies” in the annotation. It points towards the need to get the annotations validated by multiple human annotators.

Acknowledgements

We would like to thank Microsoft Research India for providing grants to prepare the dataset and to our annotators who worked very hard to finish the annotations within a strict deadline.

We would also like to thank the participants of the Shared Task for their participation and feedback and the TRAC workshop PC members for thoroughly reviewing the shared task papers within a very short span of time.

References

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431 – 442. Springer.

Swati Agarwal and Ashish Sureka. 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Ignacio Arroyo-Fernández, Dominic Forest, Juan-Manuel Torres-Moreno, Mauricio Carrasco-Ruiz, Thomas Legeleux, and Karen Joannette. 2018. Cyber-bullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Peter Burnap and Matthew L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In *Proceedings of Internet, Policy & Politics*, pages 1 – 18.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *ISWC, Shanghai*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. privacy, security, risk and trust (passat). In *International Conference on Social Computing (SocialCom)*, pages 71–80.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence*, pages 275–281. Springer, Berlin.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Karthik Dinakar, Birago Jones, Catherine Havasi Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18:1–18:30.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29 – 30.

Paula Fortana. 2017. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master's thesis, Faculdade de Engenharia da Universidade do Porto.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Thiago Galery, Efstathios Charitos, and Ye Tian. 2018. Aggression identification and multi lingual word embeddings. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon- based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215 – 230.

Viktor Golem, Mladen Karan, and Jan najder. 2018. Combining traditional machine learning models with deep learning for aggressive text detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Edel Greevy and Alan F. Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468 – 469. ACM.

Edel Greevy. 2004. *Automatic text categorisation of racist webpages*. Ph.D. thesis, Dublin City University.

Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research. Language, Behaviour, Culture*, 6(2):215–242.

Claire Hardaker. 2013. uh. . . . not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug. an overview of trolling strategies. *Journal of Language Aggression and Conflict*, 1(1):58–86.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Vronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, Cyberbullying, and Suicide. *Archives of suicide research*, 14(3):206–221.

E. Krol. 1992. *The whole internet: User's guide & catalog*. O'Reilly & Associates, Inc., Sebastopol, CA.

Sudhakar Kumar, Francesca Spezzano, and VS Subrahmanian. 2014. Accurately detecting trolls in slashdot zoo via decluttering. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 188–195.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018a. Trac-1 shared task on aggression identification: Iit(ism)@coling18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated corpus of hindi-english code-mixed data. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Sreekanth Madisetty and Maunendra Sankar Desarkar. 2018. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Promita Maitra and Ritesh Sarkhel. 2018. Emoti-kate: a k-competitive autoencoder for aggression detection in social media text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1 – 16.

Todor Mihaylov, Georgi D Georgiev, AD Ontotext, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning, CoNLL*, pages 310–314.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Luis G Mojica. 2016. Modeling trolling in social media conversations.

Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. 2018. Lstms with attention for aggression detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Nitin, Ankush Bansal, Siddhartha Mahadev Sharma, Kapil Kumar, Anuj Aggarwal, Sheenu Goyal, Kanika Choudhary, Kunal Chawla, Kunal Jain, and Manav Bhasinar. 2012. Classification of flames in computer mediated communications.

Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. 2013. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of IJCNLP*, pages 579–586.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Ahmed Husseini Orabi, Mahmoud Husseini Orabi, Qianjia Huang, Diana Inkpen, and David Van Bruwaene. 2018. Cyber-aggression detection using cross segment-and-concatenate multi-task learning from text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Constantin Orasan. 2018. Aggressive Language Identification Using Word Embeddings and Sentiment Features. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Manish Gupta Pinkesh Badjatiya, Shashank Gupta and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759 – 760. International World Wide Web Conferences Steering Committee.

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vitor Beires Nogueira. 2018. Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Faneva Ramiandrisoa and Josiane Mothe. 2018. Irit at trac 2018. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Julian Risch and Ralf Krestel. 2018. Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Arjun Roy, Prashant Kapil, Kingshuk Basak, and Asif Ekbal. 2018. An ensemble approach for aggression identification in english and hindi text. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Niloofar Safi Samghabadi, Deepthi Mave, Sudipta Kar, and Thamar Solorio. 2018. Ritual-uh at trac 2018 shared task: Aggression identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Sasha Sax. 2016. Flame Wars: Automatic Insult Detection. Technical report, Stanford University.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10, Valencia, Spain.

Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. 2018. Textual aggression detection through deep learning. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC – 1)*, Santa Fe, USA.

Fabio Del Vigna, Andrea Cimino, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity*, pages 86 – 95.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88 – 93.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.