Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19
Shivang Agarwal, C. Ravindranath Chowdary

Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, 221005, India

ARTICLE INFO
Keywords:
Hate speech detection
Ensemble learning
Social media

ABSTRACT
Social media platforms generate an enormous amount of data every day. Millions of users engage themselves with the posts circulated on these platforms. Despite the social regulations and protocols imposed by these platforms, it is difficult to restrict some objectionable posts carrying hateful content. Automatic hate speech detection on social media platforms is an essential task that has not been solved efficiently despite multiple attempts by various researchers. It is a challenging task that involves identifying hateful content from social media posts. These posts may reveal hate outrageously, or they may be subjective to the user or a community. Relying on manual inspection delays the process, and the hateful content may remain available online for a long time. The current state-of-the-art methods for tackling hate speech perform well when tested on the same dataset but fail miserably on cross-datasets. Therefore, we propose an ensemble learning-based adaptive model for automatic hate speech detection, improving the cross-dataset generalization. The proposed expert model for hate speech detection works towards overcoming the strong user-bias present in the available annotated datasets. We conduct our experiments under various experimental setups and demonstrate the proposed model's efficacy on the latest issues such as COVID-19 and US presidential elections. In particular, the loss in performance observed under cross-dataset evaluation is the least among all the models. Also, while restricting the maximum number of tweets per user, we incur no drop in performance.

1. Introduction
Hate speech on social media is defined as an online post that demonstrates hatred towards a race, colour, sexual orientation, religion, ethnicity or one’s political inclination. Hate speech is not a trivial task to define, mainly because it is subjective. The classification of content as hate speech might be influenced by the relationships between individual groups, communities, and language nuances.

Davidson et al. (2017) define hate speech as “the language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group”. The critical point to note here is that hate speech is usually expressed towards a group or a community and causes may cause social disorder.

Psychologists claim that the anonymity provided by the Social Media Platforms (SMPs) is one of the reasons why people tend to be more aggressive in such environments (Burnap & Williams, 2015; Fortuna & Nunes, 2018). This aggression sometimes turns into hate speech. Also, people tend to be more involved in heated debates on social media rather than involving in a face to face discussion. The impact of social media cannot be underestimated. The propagation of hate speech has the potential for societal impact. It has been observed in the past that posts shared on social platforms or textual exchanges may instigate individuals or groups affecting the democratic process (Fortuna & Nunes, 2018). Along with the societal impact of the phenomenon, hate speech makes it uncomfortable for users who use social media only for entertainment.

According to the European Union Commission directives, hate speech is a criminal offence that needs to be legislated (Fortuna & Nunes, 2018). European Union has instructed the various SMPs to improve their automatic hate speech detection mechanisms so that no objectionable content remains available online for more than 24 h.

Hate speech detection is a challenging task because of the nature of hate speech. As human editorial methods are not feasible on enormous volumes of tweets, automated approaches or expert models are required to address the issue. The scientific study of hate speech involves analysing the existing and proposing novel automated approaches that

https://doi.org/10.1016/j.eswa.2021.115632
Received 27 August 2020; Received in revised form 10 June 2021; Accepted 18 July 2021
Available online 27 July 2021
0957-4174/© 2021 Elsevier Ltd. All rights reserved.
programmatically classify social media posts as hate speech (Gomez et al., 2020). The current procedures for hate speech detection consider it an application of supervised learning with an assumption that the ground truth is available (MacAvaney et al., 2019). The state-of-the-art methods achieve excellent performance within specific datasets (Agrawal & Awekar, 2018; Badjatiya et al., 2017). Unfortunately, the performance of these methods degrades drastically when tested on cross-datasets (i.e., similar but not same dataset) (Arango et al., 2019). Thus, we claim that to incorporate the data bias, the model requires to be adaptive towards the properties of data.

Therefore, we propose an adaptive model for automatic hate speech detection, which can overcome the data bias and perform well on cross-datasets. Our proposed method is based on our previous work A-Stacking (Agrawal & Chowdary, 2020), an ensemble-based classifier used originally for spoof fingerprint detection. A-Stacking is an adaptive classifier that uses clustering to conform to the dataset’s features and generate hypotheses dynamically.

In Table 1, we highlight the difference in the proposed work, Agrawal and Awekar (2018)’s and Badjatiya et al. (2017) methods for hate speech detection. As reported by Arango et al. (2019), the papers, as mentioned above, have flaws in their experimental settings, which is why they overestimate their results. Specifically, the authors have mishandled data with overfitting and oversampling techniques and overstated the results. Arango et al. (2019) corrected the experimental settings and reported the actual results. In addition, they also found out a strong user bias in the popular datasets. They observed a drastic change in the performance of the state-of-the-art methods on cross-datasets and when user bias is removed from the datasets. We take care of overfitting issues and compare our performance with the corrected results reported in Arango et al. (2019) and observe that our adaptive model proves to be outperforming on cross-datasets environments while maintaining a decent performance on the within-dataset environment. We list our contributions as follows:

• We provide a comprehensive study on the importance of automatic hate speech detection in the times of COVID’19 and the US presidential election. We highlight the worrying rise in hate speech on SMPs during the pandemic and the democratic process and the need for non-discriminatory access to digital platforms.
• Unlike the state-of-the-art methods that fail to achieve cross-dataset generalization, our proposed adaptive model yields adequate performance under cross-dataset environments.
• We perform our experiments on standard high-dimensional datasets. We use multiple experimental settings to explore the model’s behaviour while considering the user-overfitting effect, data bias, and restricting the number of tweets per user.

1.1. Definitions

1. Hate Speech: European Union defines hate speech as “All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic”. International minorities associations (ILGA) define hate speech as “Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups”.

2. User bias: User bias is a phenomenon due to which a dataset is constituted majorly by a single user (or a small set of users), thereby increasing the chance of model overfitting.

3. Within-dataset environment: It is an experimental setting where the train and test data are two disjoint sets from the same dataset.

4. Cross-dataset environment: In this experimental setting, the train and test datasets are two disjoint sets from different but similar datasets. This environment evaluates the generalization abilities of the learning model.

2. Characteristics of hate speech

The key characteristics of hate speech are:

- Hate speech is meant to target a group or a community based on their ethnicity, religion, origin, sexual orientation, physical appearances or political inclinations.
- Hate speech has the potential to instigate violence or social disorder. Hate speech is commonly observed when people engage themselves in heated arguments based on their political inclinations. In extreme cases, it may also affect the democratic processes.
- There is a fine line between humorous content and hate speech. Although humorous content may offend people, its nature is only to entertain people and not to cause a stir in society. Facebook differentiates between hate speech and humorous content as “…We do, however, allow clear attempts at humour or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (e.g. jokes, stand-up comedy, popular song lyrics, etc.)”.

Studies have been conducted to determine the more affected or frequently targeted groups by online hate speech. Some of the common phenomena observed in the studies are:

- Racism: The majority of the hate speech content online is based on racism where people are attacked based on their race (Sap et al., 2019). In Kwok and Wang (2013), the authors conducted a study on why social media contents are flagged as racist. They found out that in most cases (86%), the reason is “the presence of offensive words”. Other than that, “the presence of stereotypes and threatening” and “references to painful historical contexts” make the content racist.
- Sexism: Another category of hate speech is based on sexism which is mostly caused by the misogynistic language used on SMPs (Anzovino et al., 2018; Fersini, Nozza et al., 2018; Fersini, Rosso et al., 2018; Frenena et al., 2019; Hewitt et al., 2016). Studies conducted on UK-based Twitter profiles found around 100,000 instances of the word “rape”, out of which around 12% of the cases were threatening. The sad part is the observation that in about 29% of these instances, the word “rape” is used casually or metaphorically.

The same study also shows that women are as likely as men to post offensive tweets against women.

Frequent use of swearing in the text does not necessarily imply spreading hate speech (Salawu et al., 2020). At the same time, hateful comments can be propagated in subtle and sarcastic ways.

Table 1

Comparison with state-of-the-art.

| Method                  | Classifier | Cross-dataset evaluation | Data-Bias control |
|-------------------------|------------|--------------------------|--------------------|
| Badjatiya et al. (2017) | GBDT       | ✓                        | ✗                  |
| Agrawal and Awekar (2018)| BiLSTM     | ✗                        | ✓                  |
| Proposed work           | A-Stacking | ✓                        | ✓                  |

https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_17_403.

https://www.facebook.com/communitystandards/objectionable_content.

https://demosuk.wpengine.com/files/MISOGYNY_ON_TWITTER.pdf?7

5 https://ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech.

6 https://www.facebook.com/communitystandards/objectionable_content.

7 https://demosuk.wpengine.com/files/MISOGYNY_ON_TWITTER.pdf?1399567516.
It is a general misconception to see offensive language as hate speech. While the former is morally wrong and flouts ill-mannerism, the latter is a crime and prohibited by law8 (Hate speech is prohibited in the United States under the free speech provisions of the first amendment9 (Paschalides et al., 2020). Countries like United Kingdom, France, Canada, etc., have imposed laws on propagating hate speech that affects minorities and may result in community violence). As these laws extend to social media and digital platforms, it becomes essential for the SMPs to administer their provisions for hate speech detection. SMPs provide guidelines for removing offensiveness or harmful content.

While the manual review of the content is accurate to a greater extent, it is relatively slow because of which the content in question can be available online for a long time. In addition, in the unprecedented times caused by COVID19, it is unfeasible for SMPs to flag all the hateful content manually. Therefore, the need of the hour is an efficient expert model for automatic hate speech detection.

The rise in hate speech propagation on social media has alerted the SMPs to take decisive actions against it. Twitter, for instance, has declared an update on its rules for flagging content as hate speech. The examples given below are now considered as hateful conduct and will be removed if reported.

- “All [Age Group] are leeches and do not deserve any support from us”.
- “People with [Disease] are rats that contaminate everyone around them”.
- “[Religious Group] should be punished. We are not doing enough to rid us of those filthy animals”.

3. Importance of automatic hate speech detection

Hate speech detection has become a popular research area in recent years. The need for automatic mechanisms for detecting hate speech on social media becomes urgent in the difficult times of global pandemic and during democratic processes such as the US presidential elections. We describe the importance of automatic hate speech detection in the following subsections:

3.1. Hate speech in the times of COVID-19

While Internet communication has helped the world survive this unprecedented period, it has also grown more toxic than before. After the widespread outbreak of COVID-19, an AI-based start-up L1ght analysed the SMPs and observed that the propagation of hate speech has increased by 900% during the pandemic.11 They have also observed a 70% rise in toxicity among teens and youngsters during the period of December 2019 to June 2020. Although we do not endorse these numbers, it is evident that there has been a worrying rise in hate speech propagation during the pandemic. Most of the hateful conduct regarding the novel coronavirus situation is directed towards China and the Chinese. As we have defined earlier, targeting a community is directed towards China and the Chinese.

The scarcity of datasets on hate speech detection makes it challenging to analyse the models proposed in the past. Currently, Waseem and Hovy (2016) and SemEval2019 (Basile et al., 2019) are the most popular datasets for the task. As we shall see in Section 6, these datasets have certain limitations and may introduce bias in the performance of hate speech detectors. Detecting hate speech in a text is not an easy task, as it depends on various factors. The primary issue is its subjectiveness. It is a complex phenomenon that affects differently to different people based on their nationality, religion, ethics or even language nuances (Chopra et al., 2020; Corazza et al., 2020; Ousidhoum et al., 2019).

In the past, the problem of hate speech detection has been tried to solve by using general mechanisms for text feature extraction. These mechanisms include the Bag-of-Words (BOW) representation, the Term Frequency–Inverse Document Frequency (TF–IDF), word embeddings (Lilleberg et al., 2015), deep learning models (Leng & Jiang, 2016; Liang et al., 2017), etc.

BOW is a text representation that creates a corpus of words present in the training data. In most cases, the frequency of words available in the training data. In most cases, the frequency of words available in the training data. In most cases, the frequency of words available in the training data.

8 http://www.mandola-project.eu/m/file/filer_public/8d/b3/8db3c018-4e9f-4120-95ec-b80b5f71749/mandola_d221b.pdf.
9 https://en.wikipedia.org/wiki/Hate_speech_in_the_United_States.
10 https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html.
11 https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf.
this corpus is used as the features and acts as the input to the classifier. BOW representation is limited in determining the context, and if the context is changed or used differently, it might lead to misclassification. TF-IDF is a numeric statistic that measures each word’s importance in the corpus by its frequency in the training data.

Badjatiya et al. (2017) used a Recurrent Neural Network to extract features from the tweets and later classified these tweets using Gradient Boosted Decision Trees (GBDT) (Ye et al., 2009). Their architecture is constituted by an embedding layer followed by a Long Short Memory (LSTM) network (Hochreiter & Schmidhuber, 1997), a fully connected layer with three neurons and finally, a softmax activation. The authors use this architecture only as a feature extractor to turn a tweet (which is a sequence of words, $t = (w_1, w_2, ..., w_k)$) into a sequence of vectors $E_t = (e_{w_1}, e_{w_2}, ..., e_{w_k})$ using the embedding layer. These vectors are averaged and fed into a GBDT classifier as inputs, and the resulting class label is achieved. The authors claim to yield a 93% F1 score at their best, but as corrected by Arango et al. (2019), the actual best performance is 82% F1 score.

Agrawal and Awekar (2018) use Bidirectional LSTMs (BLSTM) as the recurrent layer that processes the input in both directions. The rest of their architecture is similar to Badjatiya et al. (2017) The authors report their best performance as 94.4% F1 score, but as corrected by Arango et al. (2019), the actual performance is 84.7%. The major flaw in their experimental setting is the method used for oversampling the minority class. The authors performed oversampling over the whole dataset and later partitioned it into train and test sets, which introduced a performance bias. Arango et al. (2019) considered an example of oversampling the minority class three times and then partitioning the whole dataset into 15 − 85% test-train split and observed that there is 38% probability that a particular instance may simultaneously belong to both train and test sets, which eventually increases the model’s performance.

The problem of automatic hate speech detection is relatively new as compared to other NLP tasks. The basis of this problem can be established by studying similar tasks such as sentiment analysis on social media which has been thoroughly explored by the researchers. In sentiment analysis, the task is to classify tweets based on their sentiments, i.e., positive, negative or neutral (Bravo-Marquez et al., 2016; Catal & Nangir, 2017; Daniel et al., 2017; Ghiasii & Lee, 2018; Hassan et al., 2020; Hassounah et al., 2020; Jain et al., 2020; Symeonidis et al., 2018; Tellez et al., 2017; Wu et al., 2019).

In Hassounah et al. (2020), the authors use two feature selection methods, ReliefF and Multi-Verse Optimizer (MVO), along with SVM classifier, to accomplish the task. In Hassan et al. (2020), the authors propose a model for predicting the early impact of scientific research articles based on the sentiments expressed about them on Twitter. This is relatively a new research topic, and it can benefit the researchers with examining the impact of their published articles. The study has been conducted on more than one Million research articles which are significant in itself. The authors claim that an article with positive and neutral sentiments expressed about it has a substantial impact on the research community. We expect that early estimation of the impact can be helpful in hate speech detection, where we can predict the overall effect of hateful posts by examining their early trends.

The more advanced research in the field of sentiment analysis is being conducted on Aspect-Based Sentiment Analysis (ABSA) (Wu et al., 2019), which provides fine-grained sentiments based on specific aspects. Another advanced area in this field is satirical text detection, which may influence the polarity of the statement in a considerable manner. In del Pilar Salas-Zárate et al. (2017), the authors propose a mechanism based on psycholinguistic features to distinguish between satirical and non-satirical texts.

The above literature survey summarizes the work done so far on automatic hate speech detection and the related fields. It is evident that there is room for improvement in the cross-dataset generalization abilities of the models. Also, it is required to consider the user distribution present in the annotated datasets to estimate the performance correctly. Therefore, we conduct a study to address these issues and analyse the proposed model thoroughly under various experimental environments.

5. Proposed model for automatic hate speech detection

In this section, we describe the proposed methodology used for automatic hate speech detection. As mentioned earlier, the major limitation of the existing methods is their inability to perform on cross-dataset environments, i.e., when the trained model is tested on a similar but different dataset, the performance drops drastically. Therefore, the existing models show poor generalization abilities and the need to adapt to the changing environment.

We make use of our previous work A-Stacking (Agarwal & Chowdary, 2020), which is an adaptive ensemble learning model proposed initially for spoof fingerprint detection. We claim that the required adaptiveness for hate speech detection can be achieved by using the A-Stacking classifier.

A-Stacking is a hybrid classifier based on ensemble learning that uses clustering to form weak hypotheses that are carefully integrated using a meta-classifier in a later stage. The model is adaptive because it studies the properties of data for generating the hypotheses to be used by base-classifiers. We have shown in Agarwal and Chowdary (2020) that A-Stacking adapts to the features of fingerprint images and efficiently classifies the new images on which the model has not been trained yet. Therefore, the motivation is to use this adaptive classifier for achieving better cross-dataset generalization for automatic hate speech detection.

We divide the whole architecture into two phases, as represented in Fig. 1. Phase-I is responsible for feature extraction, and phase-II is for classification. We use the Recurrent Neural Network (RNN) (Wang et al., 2016) for generating dense-vector representations for the tweets, also known as word-embeddings. RNN makes it possible to consider the context of words while generating embeddings. The embedding layer is followed by Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997) and a softmax activation. We use binary cross-entropy (Mannor et al., 2005) as the loss function and a dropout of 0.3 along with the Adam optimizer. The embedding dimension is 200. This process is carefully done while following the guidelines by Arango et al. (2019) so that the overfitting bias in the performance can be avoided.

As described earlier, a tweet can be seen as a vector of words, $t = (w_1, w_2, ..., w_k)$. The task is to create a dictionary of words and pass it to the embedding layer that generates a sequence of vectors, $E_t = (e_{w_1}, e_{w_2}, ..., e_{w_k})$. These vectors are averaged and a single vector $\bar{E}$ for each tweet is generated. In this study, we do not perform end-to-end classification, rather we generate these sets of vectors for train set $D_{Train}$ and test set $D_{Test}$ and later classify $D_{Test}$ in phase-II, $D_{Train}$ is used to train the A-Stacking model. First, $D_{Train}$ is partitioned into two parts: one is used for validation $D_{Train_{val}}$ and the rest of it is used for actual training. In this study, we have considered $D_{Train_{val}}$ to be 20% of $D_{Train}$. This way, train, test and validation sets are disjoint from each other.

The classifier studies the data and constitutes a set of clusters $C = \{c_1, c_2, ..., c_n\}$, where the number of clusters $n$ is decided a priori.
Later, base classifiers are trained on these individual clusters and a set of hypotheses \( H = (h_1, h_2, \ldots, h_n) \) is generated. Each hypothesis \( h_i \) is tested on multiple base classifiers (e.g., SVM, GBDT, etc.) and the best performing base classifier is chosen based on its performance on \( D_{val} \) and fixed for the particular hypothesis.

Later, each instance \( i \) of \( D_{val} \) is passed through the set \( H \) and the individual decisions \( v_j \) of each \( h_j \) are recorded and sent to a meta-classifier \( M \). This meta-classifier is responsible for carefully integrating the set of \( v_j \)'s and coming up with the final decision. This final decision is the class label for the tweet, i.e., sexist, racist, non-hateful etc.

6. Experimental setup, analysis and discussion

In this section, we describe the experimental results performed under various settings along with a discussion. We explore the proposed model's behaviour thoroughly by creating multiple environments, such as within dataset, cross-dataset, and by limiting the number of tweets per user (to avoid user-bias). Next, we give details of the datasets used in the study:

6.1. Datasets

As mentioned earlier, there is a scarcity of annotated datasets for hate speech detection. SMP such as Facebook does not impose a restriction on the size of text, making it challenging to analyse. In this study, we focus on textual data from the micro-blogging site Twitter. The most popular datasets are Waseem and Hovy (2016) and SemEval2019 (Basile et al., 2019). Waseem and Hovy dataset comes with tweet identifiers along with their associated class labels, i.e., sexist, racist and non-hateful. The actual tweets can be extracted using any tweet crawler. 16k tweet identifiers constitute the dataset; the actual number of tweets was lesser than that. Despite its popularity, the dataset is highly biased. It has been observed that only a few users have communicated the majority of the hateful tweets in this dataset. Specifically, 65% of the hateful tweets are posted by only two users, which causes a significant user-overfitting effect. Therefore, it is advised to restrict the number of tweets per user and then conduct another set of experiments. Ideally, if the model is capable of overcoming the user-overfitting effect, it must not observe a performance drop.

In addition, there was a massive loss of tweets when we extracted them using the identifiers. We crawled the tweets in mid-2020 and achieved only 10,147 tweets. Out of which, only four were from the “racist” class. While considering the three categories of the dataset, the results could not reflect the models' behaviours correctly. Therefore, we requested (Arango et al., 2019) to provide us with the extracted data used in their paper. The authors generously supported us and provided useful links.

SemEval 2019 dataset is from “Multilingual detection of hate speech against immigrants and women in Twitter” (Basile et al., 2019). The dataset comprises 9k tweets with only two labels: hateful and non-hateful. It provides the tweets and the associated labels, but not the user identifiers. The user information is protected under the General Data Protection Regulation (EU GDPR).

For providing case studies on hate speech propagation during the ongoing pandemic and US elections 2020, we used some samples from the Covid-Hate dataset (Ziems et al., 2020). The dataset contains a collection of tweets related to COVID-19. A part of the dataset includes 2400 hand labelled tweets (Covid–Hatex). The other version consists of over 30M machine labelled tweets. We sampled around 10,000 tweets (Covid–HatexML) to test our models' efficacy. Note that we do not consider the tweets belonging to ‘counterhate’ and ‘non-Asian aggression’ categories in this study. For the US election-related experiments, we handpicked some tweets from the Covid-Hate datasets related to the election. The resulting dataset is the combination of hand labelled and machine labelled tweets. In all these datasets, retweets are eliminated as they have little relevance in training the models.

The details of the datasets are given in Table 2.

6.2. Experimental setup

Next, we describe the implementation details and experimental setup use for the study. As defined earlier, we work in two phases: Phase-I makes use of RNNs to get word embeddings while considering the textual context. Phase-II uses the A-Stacking (Agarwal & Chowdary, 2020; Agarwal et al., 2021) model for classifying the input vectors into predefined labels. A-Stacking is an ensemble-based classifier that uses multiple base classifiers with a combination of a meta-classifier. In this study, we have used Support Vector Machine Classifier (SVM) (Hearst, 1998), Gradient Boosting Decision Trees (GBDT) (Ye et al., 2009), Multi-Layer Perceptron Classifier (MLP) (Pal & Mitra, 1992), kNeighbors Classifier (Zhang & Zhou, 2007), ELM classifier along with Logistic Regression (Le Cessie & van Houwelingen, 1992) as the meta-classifier. For clustering, we have used SimpleKMeans (Arthur & Vassilvitskii, 2007) clustering algorithm with varying values of \( k \).

6.2.1. Evaluation metrics

**Precision**: Precision is used for evaluating the correct positive predictions. It is calculated as the ratio of correctly classified positive instances to total predicted positive instances.

**Recall**: Recall is the fraction of total correctly predicted instances among all positive instances.

**F1**: F1-measure conveys the balance between precision and recall. It is calculated as \( 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \).

A macro-average computes the metric independently for each class and then takes the average, whereas a micro-average aggregates all classes' contributions to computing the average metric.

6.3. Results

We performed our experiments under three categories. Category-I is for testing the within dataset performance of the model, where we train and test the model on the same dataset. The dataset is partitioned into train and test sets using 10-fold cross-validation. Category-II explores the model's behaviour in the cross-dataset environment, where we train it on a dataset and test it on another dataset and vice-versa. We have designed Category-III for restricting the user distribution among the dataset so that we can avoid the user-overfitting effect. For this purpose, we have limited the number of tweets per user to 250. In general, the average number of tweets per day (TPD) is 4.42 for an active user. Waseem & Hovy dataset (Waseem & Hovy, 2016) was collected in the span of two months. Therefore 250 is the ideal choice for such restriction. If a user constitutes more than 250 tweets in the dataset, we randomly select 250 of their tweets.

6.3.1. Category-I: within dataset performance

In Category-I, we test the models' behaviour following the within-dataset environment, where the train and the test sets are from the same dataset. We use 10-fold cross-validation for partitioning the data. The results of various models on Waseem & Hovy and SemEval 2019 datasets under category-I are given in Table 3. The best results for the proposed model were achieved while considering five clusters on (a) and ten clusters on (b). As shown in Table 3, the proposed model

---

15 https://scikit-elm.readthedocs.io/en/latest/user_guide.html.
16 https://blog.hubspot.com/blog/tabid/6307/bid/4594/is-22-tweets-per-day-the-optimum.
represents the Micro average, and the second represents the Macro average.

As mentioned earlier, the available annotated datasets are highly biased, with only a few users constituting the majority of the hateful tweets (Poletto et al., 2020). Therefore, we claim that it is vital to highlight the model’s behaviour while restricting the number of tweets tested on new data which were not seen while training the model. We design this setup by training the models on Waseem & Hovy dataset and testing them on SemEval 2019 dataset. For managing the compatibility issues, we merge the three classes of the Waseem and Hovy dataset into two: Hateful and Non-Hateful. It can be observed from Table 6 that the performance of the models degrade drastically and incurs a loss of about 20%. The proposed model performs adequately well and manages a good score.

In addition, we also tested the efficacy of the proposed model while using SemEval 2019 dataset for training and Waseem & Hovy dataset for testing. We achieve a better performance than the rivals, which is adequate for the cross-dataset generalization. The best results for the proposed model were achieved while considering ten clusters on both datasets.

6.3.3. Category-III: controlling the user overfitting effect

As mentioned earlier, the available annotated datasets are highly biased, with only a few users constituting the majority of the hateful tweets (Poletto et al., 2020). Therefore, we claim that it is vital to highlight the model’s behaviour while restricting the number of tweets...
in the changing environment. Speech detection to avoid a performance drop after the limit is imposed. It was a prerequisite for the hate proposed model’s performance when the dataset is not restricted v/s the generalization ability of the state-of-the-art.

slightest deviation. Therefore, the model is able to overcome the poor accuracy degrades significantly, but the proposed model shows the same environment v/s—tested in a cross environment. It is evident that the difference in the performance of various models when tested in the change.

per user. Therefore, we explored the model’s ability to adapt to this effect. As shown in Table 7, there is no significant drop in the performance of the proposed model compared to the performance mentioned in Table 3. Therefore, we can establish that the proposed model is able to overcome the user-overfitting effect.

6.4. Discussion

We worked towards proposing an adaptive automatic hate speech detection model that can perform reasonably well on cross-datasets. The results reported in Tables 3–7 show that our proposed model is capable of adapting to the properties of data and behave accordingly when the test environment is changed. We emphasized that the available annotated datasets have a strong bias in them. For the correct assessment of the model, it is necessary to restrict the number of tweets per user. Therefore, we explored the model’s ability to adapt to this change.

In the bar diagram represented in Fig. 2, we show the percentage difference in the performance of various models when tested in the same environment v/s—tested in a cross environment. It is evident that the accuracy degrades significantly, but the proposed model shows the slightest deviation. Therefore, the model is able to overcome the poor generalization ability of the state-of-the-art.

The bar diagram represented in Fig. 3 shows the difference in the proposed model’s performance when the dataset is not restricted v/s the performance after imposing the limit. It was a prerequisite for the hate speech detector to avoid a performance drop after the limit is imposed. It is evident that our model shows no drop and performs reasonably well in the changing environment.

7. Conclusions and future work

Automatic hate speech detection on social media is an important task that provides users with non-discriminatory access to digital platforms. We emphasize the importance of the task during the global outbreak of COVID19, where users are spending more time on social media and information is getting sensitive. With factual information, we have shown that there is a rapid growth in hate speech posts on social media, specifically targeting a community or a country and its citizens. We have also emphasized the importance of this research area during the US presidential elections. We reported the level of sincerity the social media platforms seek to improve their mechanisms for detecting hate speech without manual inspection. To address all these issues, we proposed an adaptive expert model for the task, keeping the motivation that the model must perform in cross-dataset environments. Our proposed model is constituted by deep learning methods for feature extraction and an ensemble-based adaptive classifier for predicting the class labels for the tweets. We performed our experiments on standard datasets along with latest datasets on COVID-19 and US elections and reported the performance under various experimental setups. We also reported the fallacies present in the available datasets; therefore, we strive to develop a dataset that is more fine-grained and free from user overfitting effect in the future.

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.

References

Agarwal, S., & Chowdary, C. R. (2020). A stacking and a-bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection. Expert Systems with Applications, 146, Article 113160. http://dx.doi.org/10.1016/j.eswa.2019.113160, URL http://www.sciencedirect.com/science/article/pii/S0957417419308784.

Agarwal, S., Rattani, A., & Chowdary, C. R. (2021). A comparative study on handcrafted features v/s deep features for open-set fingerprint liveness detection. Pattern Recognition Letters, 147, 34–40. http://dx.doi.org/10.1016/j.patrec.2021.03.032, URL https://www.sciencedirect.com/science/article/pii/S0167865521001276.

Agarwal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), Advances in information retrieval - 40th European conference on IR research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (pp. 141–153). Springer, http://dx.doi.org/10.1007/978-3-319-76941-7_11.

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In M. Silberztein, F. Atigui, E. Kornyshova, & A. Hanbury (Eds.), Natural language processing and information systems - 23rd international conference on applications of natural language to information systems, NLDB 2018, Paris, France, June 13-15, 2018, proceedings (pp. 57-64). Springer, http://dx.doi.org/10.1007/978-3-319-91947-8_6.

---

Table 7

| Method               | F1   | Prec. | Rec. |
|----------------------|------|-------|------|
| Proposed model       | 81.54| 82.21 | 81.38|
| Badjatiya et al. (2017) | 82.13| 82.16 | 82.96|
| Agrawal and Awekar (2018) | 73.89| 77.71 | 72.39|
| rival-1              | 67.11| 66.63 | 70.76|
| rival-2              | 67.11| 66.63 | 70.76|
| proposed             | 81.54| 82.21 | 81.38|

Fig. 2. Calculating the percentage drop in F1-score (micro) of various models when tested under cross-dataset environment in comparison with within-dataset environment. The significant drop in performance justifies the need of cross-dataset generalization. rival-1 = Badjatiya et al. (2017), rival-2 = Agrawal and Awekar (2018).

Fig. 3. Performance comparison of various models on Waseem and Hovy dataset under two environments: dataset with no restrictions on the number of tweets and dataset with a cap of 250 tweets per user. We show the F1 score-micro values for this comparison. Proposed model observes no drop.
S. Agarwal and C.R. Chowdary

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 1668–1678). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1163, URL https://www.aclweb.org/anthology/P19-1163.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, 298–310. http://dx.doi.org/10.1016/j.eswa.2018.06.022, URL http://www.sciencedirect.com/science/article/pii/S0957417418303683.

Tellez, E. S., Miranda-Jíménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017). A case study of spanish text transformations for twitter sentiment analysis. Expert Systems with Applications, 81, 457–471. http://dx.doi.org/10.1016/j.eswa.2017.03.071, URL http://www.sciencedirect.com/science/article/pii/S0957417417302312.

Wang, X., Jiang, W., & Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers (pp. 2428–2437). Osaka, Japan: The COLING 2016 Organizing Committee, URL https://www.aclweb.org/anthology/C16-1229.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88–93). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-2013, URL https://www.aclweb.org/anthology/N16-2013.

Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. In Proceedings of the 18th ACM conference on information and knowledge management (pp. 2061–2064). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1645953.1646301.

Zhang, M., & Zhou, Z. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 40, 2038–2048. http://dx.doi.org/10.1016/j.patcog.2006.12.019.

Ziemys, C., He, B., Soni, S., & Kumar, S. (2020). Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv:2005.12423.