International Journal of Informatics Technology (INJIT)
Vol. 1., No. 3, November 2023
e-ISSN : 2985-9646
https://jurnal.amrillah.net/index.php/injit

Analysis of Hate Speech Detection based on Obstacles and Solutions using Deeplearning Methods

Shayuti Rohmah[a,1,*], Hopa Kapi[b,2]

[a]Department of Computer Engineering, College of Computer Engineering Sciences,
Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[b]Department of Computer Science and Engineering, School of Engineering Sciences and Technology,
Jamia Hamdard, New Delhi, India
[1]shayut91@gmail.com *, [2]hopa_kapi@gmail.com
* Corresponding Author

ABSTRACT

Hate speech is one type of harmful online content which directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation. With online hate speech on the rise, its automatic detection as a natural language processing task is gaining increasing interest. However, it is only recently that it has been shown that existing models generalise poorly to unseen data. This survey paper attempts to summarise how generalisable existing hate speech detection models are and the reasons why hate speech models struggle to generalise, sums up existing attempts at addressing the main obstacles, and then proposes directions of future research to improve generalisation in hate speech detection.

INJIT International Journal of Informatics Technology

KEYWORDS
Artificial Intelligence
Explorer
Symbolic
Glosglow

1. Introduction

The Internet saw a growing body of user-generated content as social media platforms flourished (Schmidt & Wiegand, 2017; Chung et al., 2019). While social media provides a platform for all users to freely express themselves, offensive and harmful contents are not rare and can severely impact user experience and even the civility of a community (Nobata et al., 2016). One type of such harmful content is hate speech, which is speech that directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation (Waseem & Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Sharma, Agrawal & Shrivastava, 2018). Major social media companies are aware of the harmful nature of hate speech and have policies regarding the moderation of such posts. However, the most commonly used mechanisms are very limited. For example, keyword filters can deal with profanity, but not the nuance in the expression of hate (Gao, Kuppersmith & Huang, 2017). Crowd-sourcing methods (e.g., human moderators, user reporting), on the other hand, do not scale up. This means that by the time that a hateful post gets detected and taken down, it has already made negative impacts (Chen, McKeever & Delany, 2019).

The automatic detection of hate speech is thus an urgent and important task. Since the automatic detection of hate speech was formulated as a task in the early 2010s (Warner & Hirschberg, 2012), the field has been constantly growing along the perceived importance of the task.

Hate speech, offensive language, and abusive language Although different types of abusive and offensive language are closely related, there are important distinctions to note. Offensive language and abusive language are both used as umbrella terms for harmful content in the context of automatic detection studies. However, while "strongly impolite, rude" and possible use of profanity are seen in the definitions of both (Fortuna & Nunes, 2018), abusive language has a strong component of intentionality (Caselli et al., 2020). Thus, offensive language has a broader scope, and hate speech falls in both categories.

Because of its definition mentioned above, hate speech is also different from other sub-types of offensive language. For example, personal attacks (Wulczyn, Thain & Dixon, 2017) are characterised by being directed at an individual, which is not necessarily motivated by the target's identity. Hate speech is also different from cyberbullying (Zhao, Zhou & Mao, 2016), which is carried out repeatedly and over time against vulnerable victims that cannot defend themselves.1 This paper focuses on hate speech and hate speech datasets, although studies that cover both hate speech and other offensive language are also mentioned.

Most if not all proposed hate speech detection models rely on supervised machine learning methods, where the ultimate purpose is for the model to learn the real relationship between features and predictions through training data, which generalises to previously unobserved inputs (Goodfellow, Bengio & Courville, 2016). The generalisation performance of a model measures how well it fulfils this purpose.

To approximate a model's generalisation performance, it is usually evaluated on a set-aside test set, assuming that the training and test data, and future possible cases come from the same distribution. This is also the main way of evaluating a model's ability to generalise in the field of hate speech detection.

The ultimate purpose of studying automatic hate speech detection is to facilitate the alleviation of the harms brought by online hate speech. To fulfil this purpose, hate speech detection models need to be able to deal with the constant growth and evolution of hate speech, regardless of its form, target, and speaker.

Recent research has raised concerns on the generalisability of existing models (Swamy, Jamatia & Gambäck, 2019). Despite their impressive performance on their respective test sets, the performance significantly dropped when the models are applied to a different hate speech dataset. This means that the assumption that test data of existing datasets represent the distribution of future cases is not true, and that the generalisation performance of existing models have been severely overestimated (Arango, Prez & Poblete, 2020). This lack of generalisability undermines the practical value of these hate speech detection models.

So far, existing research has mainly focused on demonstrating the lack of generalisability (Gröndahl et al., 2018; Swamy, Jamatia & Gambäck, 2019; Wiegand, Ruppenhofer & Kleinbauer, 2019; Fortuna, Soler-Company & Wanner, 2021), apart from a handful of studies that made individual attempts at addressing aspects of it (Karan & Šnajder, 2018; Waseem, Thorne & Bingel, 2018; Arango, Prez & Poblete, 2020). Recent survey papers on hate speech and abusive language detection (Schmidt & Wiegand, 2017; Fortuna & Nunes, 2018; Al-Hassan & Al-Dossari, 2019; Mishra, Yannakoudakis & Shutova, 2019; Vidgen et al., 2019; Poletto et al., 2020; Vidgen & Derczynski, 2020) have focused on the general trends in this field, mainly by comparing features, algorithms and datasets. Among these, Fortuna & Nunes (2018) provided an in-depth review of definitions, Vidgen et al. (2019) concisely summarised various challenges for the detection of abusive language in general, Poletto et al. (2020) and Vidgen & Derczynski (2020) created extensive lists of resources and benchmark corpora while Al-Hassan & Al-Dossari (2019) focused on the special case of the Arabic language.

This survey paper thus contributes to the literature by providing (1) a comparative summary of existing research that demonstrated the lack of generalisability in hate speech detection models, (2) a systematic analysis of the main obstacles to generalisable hate speech detection and existing attempts to address them, and (3) suggestions for future research to address these obstacles. This paper is most relevant to any researcher building datasets of, or models to detect, online hate speech, but can also be of use for those who work on other types of abusive or offensive language.

We started with a pre-defined set of keywords. Then, titles of proceedings of the most relevant recent conferences and workshops (Workshop on Abusive Language Online, Workshop on Online Abuse and Harms) were skimmed, to refine the set of keywords. We also modified the keywords during the search stages as we encountered new phrasing of the terms. The above keywords shown are the final keywords. Before starting to address the aims of this paper, an initial coarse literature search involved searching for the general keywords, skimming the titles and abstracts. During this stage, peer-reviewed papers with high number of citations, published in high-impact venues were prioritised. Existing survey papers on hate speech and abusive language detection (Schmidt & Wiegand, 2017; Fortuna & Nunes, 2018; Al-Hassan & Al-Dossari, 2019; Mishra, Yannakoudakis & Shutova, 2019; Vidgen et al., 2019; Poletto et al., 2020; Vidgen & Derczynski, 2020) were also used as seed papers. The purpose of this stage was to establish a comprehensive high-level view of the current state of hate speech detection and closely related fields.

For the first aim of this paper—building a comparative summary of existing research on generalisability in hate speech detection—the search mainly involved different combinations of the general and generalisation-related keywords. As research on this topic is sparse, during this stage, all papers found and deemed relevant were included.

Building upon the first two stages, the main obstacles towards generalisable hate speech detection were then summarised: (1) presence of non-standard grammar and vocabulary, (2) paucity of and biases in datasets, and (3) implicit expressions of hate. This was done through extracting and analysing the error analysis of experimental studies found in the first stage, and comparing the results and discussions of the studies found in the second stage. Then, for each category of obstacles identified, another search was carried out, involving combinations of the description and paraphrases of the challenges and the general keywords. The search in this stage is the most fine-grained, in order to ensure coverage of both the obstacles and existing attempts to address them. After the main search stages, the structure of the main findings in the literature was laid out. During writing, for each type of findings, the most representative studies were included in the writing up. We defined the relative representativeness within studies we have found, based on novelty, experiment design and error analysis, publishing venues, and influence. We also prioritised studies that addressed problems specific to hate speech, compared to better-known problems that are shared with other offensive language and social media tasks.

Testing a model on a different dataset from the one which it was trained on is one way to more realistically estimate models' generalisability (Wiegand, Ruppenhofer & Kleinbauer, 2019). This evaluation method is called cross-dataset testing (Swamy, Jamatia & Gambäck, 2019) or cross-application (Gröndahl et al., 2018), and sometimes cross-domain classification (Wiegand, Ruppenhofer & Kleinbauer, 2019) or detection (Karan & Šnajder, 2018) if datasets of other forms of offensive language are also included.

As more hate speech and offensive language datasets emerged, a number of studies have touched upon cross-dataset generalisation since 2018, either studying generalisability per se, or as part of their dataset validation. The datasets they use (Table 1) to some extent reflect the best-known datasets in hate speech and other types of offensive language. These studies are further compared in Table 2 in terms of the models and datasets they used. As different datasets and models were investigated, instead of specific performance metrics, the remainder of this section will discuss the general findings of these studies, which can be roughly grouped into those on models and those on training and evaluation data.

2. Method

First of all, model performance had been severely over-estimated. This includes existing "state-of-the-art" models and common baselines. Models used in the experiments ranged from neural networks—deep or shallow—to classical machine learning methods, including mixtures of both. When applied cross-dataset, all show a significant performance drop. Performance on a different dataset highlights that the test set of the same dataset does not realistically represent the distribution of unseen data. Earlier (before 2019) state-of-the-art models often involved recurrent neural networks (Gröndahl et al., 2018).

For example, the CNN-GRU model by Zhang, Robinson & Tepper (2018) first extracts 2 to 4-gram features using convolutional layers with varying kernel sizes on word embeddings, then captures the sequence orders of these features with a gated recurrent unit (GRU) layer. This model outperformed previous models on six datasets when tested in-dataset. However, when tested cross-dataset by Gröndahl et al. (2018), the model's performance dropped even more than an LSTM, by over 30 points in macro-averaged F1.

Similarly, Badjatiya et al. (2017)'s model was once considered state-of-the-art when trained and evaluated on Waseem. Their two-stage training first produces word embeddings using a Long Short-Term Memory (LSTM) network through the same hate speech classification task, based on which another Gradient-Boosted Decision Tree (GBDT) classifier was trained. Arango, Prez & Poblete (2020) showed a similar F1 drop of around 30 points when applied on HatEval, and discussed a crucial methodological flaw—overfitting induced by extracting features on the combination of training and test set. Gröndahl et al. (2018) also reported that they failed to reproduce Badjatiya et al. (2017)'s results. Both Gröndahl et al. (2018) and Arango, Prez & Poblete (2020) also tested a Long Short-Term Memory (LSTM) network, which had been commonly used as a strong baseline. The performance drop was similar to the above two state-of-the-art models by Zhang, Robinson & Tepper (2018) and Badjatiya et al. (2017).

Since the introduction of BERT (Devlin et al., 2019), itself and its variants have been established as the new state-of-the-art. This is seen through the comparison to other neural networks (Swamy, Jamatia & Gambäck, 2019) and on the leaderboards of shared tasks, such as Zampieri et al. (2020); Fersini, Nozza & Rosso (2020). The general approach is to fine-tune a model, which had been pre-trained on domain-general data, on a target classification dataset. Yet, BERT and its variants are no exception to the lack of generalisation, although the cross-dataset performance drop is seemingly smaller. In cross-dataset experiments with four datasets, macro-averaged F1 scores decreased by 2 to 30 points (Swamy, Jamatia & Gambäck, 2019), which is less drastic compared to earlier state-of-the-art neural networks tested in other studies (Gröndahl et al., 2018; Arango, Prez & Poblete, 2020). Pamungkas, Basile & Patti (2020) and Fortuna, Soler-Company & Wanner (2021) also found that BERT and ALBERT tended to generalise the best across the models they experimented with.

Building upon BERT, a handful of recent studies suggest that additional hate-specific knowledge from outside the fine-tuning dataset might help with generalisation. Such knowledge can come from further masked language modelling pre-training on an abusive corpus (Caselli et al., 2021), or features from a hate speech lexicon (Koufakou et al., 2020).

Other models that have been studied include traditional machine learning models, such as character n-gram Logistic Regression (Gröndahl et al., 2018), character n-gram Multi-Layer Perceptron (MLP) (Gröndahl et al., 2018; Waseem, Thorne & Bingel, 2018), Support Vector Machines (Karan & Šnajder, 2018; Fortuna, Soler-Company & Wanner, 2021; Pamungkas & Patti, 2019; Pamungkas, Basile & Patti, 2020), and shallow networks with pre-trained embeddings, e.g., MLP with Byte-Pair Encoding (BPE)-based subword embeddings (Heinzerling & Strube, 2018; Waseem, Thorne & Bingel, 2018) and FastText (Joulin et al., 2017a; Wiegand, Ruppenhofer & Kleinbauer, 2019; Fortuna, Soler-Company & Wanner, 2021).

Generally, these simpler models do not perform as good as deep neural networks, such as LSTM (Pamungkas & Patti, 2019) and especially BERT and its variants (Pamungkas, Basile & Patti, 2020; Fortuna, Soler-Company & Wanner, 2021), in- or cross-dataset. However, exceptions exist in some dataset combinations, especially when it comes to generalising. For example, n-gram Logistic Regression when comparing to LSTM (Gröndahl et al., 2018), SVM when comparing to LSTM and BERT (Pamungkas & Patti, 2019; Pamungkas, Basile & Patti, 2020), and FastText when comparing to BERT (Fortuna, Soler-Company & Wanner, 2021).

These cross-dataset studies only cover some of the more representative and/or recent hate speech detection models, but one can expect that the generalisation problem go beyond this small sample, and is far more ubiquitous in existing models than what these studies cover. Despite the significance of the problem, systematic studies that compared a variety of models with datasets controlled are very limited (Arango, Prez & Poblete, 2020; Pamungkas & Patti, 2019; Pamungkas, Basile & Patti, 2020; Fortuna, Soler-Company & Wanner, 2021); there is also limited overlap in the datasets used between different studies (Table 2). Thus, one should be careful when drawing conclusions on the relative generalisability of models.

Training data has a pronounced influence on generalisation. The performance drops in models highlight the differences in the distribution of posts between datasets (Karan & Šnajder, 2018), yet some datasets are more similar to each other. Furthermore, certain attributes of a dataset could lead to more generalisable models. Similarity between datasets varies, as there are groups of datasets that produce models that test much better on each other. For example, in Wiegand, Ruppenhofer & Kleinbauer (2019)'s study, FastText models (Joulin et al., 2017a) trained on three datasets (Kaggle, Founta, Razavi) achieved F1 scores above 70 when tested on one another, while models trained or tested on datasets outside this group achieved around 60 or less. In Swamy, Jamatia & Gambäck (2019)'s study with fine-tuned BERT models (Devlin et al., 2019), Founta and OLID produced models that performed well on each other. The source of such differences are usually traced back to search terms (Swamy, Jamatia & Gambäck, 2019), topics covered (Nejadgholi & Kiritchenko, 2020; Pamungkas, Basile & Patti, 2020), label definitions (Pamungkas & Patti, 2019; Pamungkas, Basile & Patti, 2020; Fortuna, Soler-Company & Wanner, 2021), and data source platforms (Glavaš, Karan & Vulić, 2020; Karan & Šnajder, 2018).

Another way of looking at generalisation and similarity is by comparing differences between individual classes across datasets (Nejadgholi & Kiritchenko, 2020; Fortuna, Soler & Wanner, 2020; Fortuna, Soler-Company & Wanner, 2021), as opposed to comparing datasets as a whole. In both Nejadgholi & Kiritchenko (2020) and Fortuna, Soler-Company & Wanner (2021)'s experiments, the best generalisation is achieved for more general labels such as "toxicity", "offensive", or "abusive". Generalisation is not as good for finer-grained hate speech labels. All in all, these findings are indicative of an imbalance of the finer-grained subclasses, particularly owing to disagreements in the definition of what constitutes hate speech, which proves more difficult than defining what constitutes offensive language.

Within the hate speech labels, the relative similarity also varies. Fortuna, Soler & Wanner (2020) used averaged word embeddings (Bojanowski et al., 2017; Mikolov et al., 2018) to compute the representations of classes from different datasets, and compared classes across datasets. One of their observations is that Davidson's "hate speech" is very different from Waseem's "hate speech", "racism", "sexism", while being relatively close to HatEval's "hate speech" and Kaggle's "identity hate". This echoes with experiments that showed poor generalisation of models from Waseem to HatEval (Arango, Prez & Poblete, 2020) and between Davidson and Waseem (Waseem, Thorne & Bingel, 2018; Gröndahl et al., 2018).

In terms of what properties of a dataset lead to more generalisable models, there are frequently mentioned factors, but also inconsistency across different studies. Interactions between factors, which contribute to the inconsistency, are also reported. The proportion of abusive posts in a dataset, first of all, plays a part. Swamy, Jamatia & Gambäck (2019) holds that a larger proportion of abusive posts (including hateful and offensive) leads to better generalisation to dissimilar datasets, such as Davidson. This is in line with Karan & Šnajder (2018)'s study where Kumar and Kolhatkar generalised best, and Waseem, Thorne & Bingel (2018)'s study where models trained on Davidson generalised better to Waseem than the other way round. In contrast, in Wiegand, Ruppenhofer & Kleinbauer (2019)'s study, the datasets with the least abusive posts generalised the best (Kaggle and Founta). Similarly, Fortuna, Soler-Company & Wanner (2021) could not confirm the impact of class proportions. Nejadgholi & Kiritchenko (2020) offered an explanation to this: there exists a trade-off between true positive and true negative rates dictated by the class proportions, which impacts the minority class performance the most but this is not always reflected in the overall F1 score.

Biases in the samples are also frequently mentioned. Wiegand, Ruppenhofer & Kleinbauer (2019) hold that less biased sampling approaches produce more generalisable models. This was later reproduced by Razo & Kübler (2020) and also helps explain their results with the two datasets that have the least positive cases. Similarly, Pamungkas & Patti (2019) mentioned that a wider coverage of phenomena lead to more generalisable models. So do topics that are more general rather than platform-specific (Nejadgholi & Kiritchenko, 2020).

A larger training data size is generally believed to produce better and more generalisable models (Halevy, Norvig & Pereira, 2009). It is mentioned as one of the two biggest factors contributing to cross-dataset performance in Karan & Šnajder (2018)'s study. Caselli et al. (2020) also found that, on HatEval, their dataset (AbuseEval) produced a model even better-performing than the one trained on HatEval end-to-end. They partially attributed this to a bigger data size, alongside annotation quality. However, the benefit of having more data is counterbalanced by data distribution differences (Karan & Šnajder, 2018), as discussed above. Moreover, its relative importance compared to other factors seems to be small, when the latter are carefully controlled (Nejadgholi & Kiritchenko, 2020; Fortuna, Soler-Company & Wanner, 2021).

Most of these studies only worked with English data. Yet, it is worth stressing that hate speech is a universal problem that exists in many languages, and generalisation studies focused on languages other than English are to date very sparse, despite the importance of the problem. Thus, research on cross-lingual generalisation is still in early stages.

One way to look at generalisation in non-English hate speech detection is applying the same cross-dataset evaluation on multiple datasets in another language. However, such studies do not yet exist. This is related to the fact that the majority of datasets are in English, which reflects linguistic and cultural unevenness in this field of research (Poletto et al., 2020; Vidgen & Derczynski, 2020).

3. Results and Discussion

The use of pre-trained embeddings (discussed earlier) and parameter dropout (Srivastava et al., 2014) have been accepted as standard practice in the field of NLP to prevent over-fitting, and are common in hate speech detection as well. Nonetheless, the effectiveness of domain-general embedding models is questionable, and there has been only a limited number of studies that looked into the relative suitability of different pre-trained embeddings on hate speech detection tasks (Chen, McKeever & Delany, 2018; Mishra, Yannakoudakis & Shutova, 2018; Bodapati et al., 2019).

In Swamy, Jamatia & Gambäck (2019)'s study of model generalisability, abusive language-specific pre-trained embeddings were suggested as a possible solution to limited dataset sizes. Alatawi, Alhothali & Moria (2020) proposed White Supremacy Word2Vec (WSW2V), which was trained on one million tweets sourced through white supremacy-related hashtags and users. Compared to general word2vec (Mikolov et al., 2013) and GloVe (Pennington, Socher & Manning, 2014) models trained on news, Wikipedia, and Twitter data, WSW2V captured meaning more suitable in the hate speech context –e.g., ambiguous words like "race" and "black" have higher similarity to words related to ethnicity than sports or colours. Nonetheless, their WSW2V-based LSTM model did not consistently outperform Twitter GloVe-based LSTM model or BERT (Devlin et al., 2019). They did not consider cross-dataset testing for generalisablity, either.

The pre-training for BERT (and its variants) is both data and computationally-heavy, which limits the feasibility of training the hate speech equivalent of BERT from scratch. A reasonable compromise to that is performing further Masked Language-Modelling pre-training before the fine-tuning stage. By further pre-training RoBERTa (Liu et al., 2019), Wiedemann, Yimam & Biemann (2020) achieved first place at the Offenseval 2020 shared task (Zampieri et al., 2020). Caselli et al. (2021) pre-trained BERT further on a larger-scale dataset of banned abusive subreddits and observed improvement over standard BERT on three Twitter datasets (OLID, AbuseEval, HatEval), in-dataset for all cases and cross-dataset for most cases. Both studies show that abusive language-specific pre-training, built upon generic pre-training, can be beneficial for both in-dataset performance and cross-dataset generalisation. The main downside is that the improvement gains, ranging from less than 1% to 4% in macro F1, seem disproportionate to the computational cost—Wiedemann, Yimam & Biemann (2020) only did the training on a small sample due to hardware limitations; it took Caselli et al. (2021) 18 days to complete 2 million training steps on one Nvidia V100 GPU. There also exists a trade-off between precision and recall for the positive class due to the domain shift (Caselli et al., 2021).

Research on transfer learning from other tasks, such as sentiment analysis, also lacks consistency. Uban & Dinu (2019) pre-trained a classification model on a large sentiment dataset (https://help.sentiment140.com/), and performed transfer learning on the OLID and Kumar datasets. They took pre-training further than the embedding layer, comparing word2vec (Mikolov et al., 2013) to sentiment embeddings and entire-model transfer learning. Entire-model transfer learning was found to be always better than using the baseline word2vec (Mikolov et al., 2013) model, but the transfer learning performances with only the sentiment embeddings were not consistent.

More recently, Cao, Lee & Hoang (2020) also trained sentiment embeddings through classification as part of their proposed model. The main differences are: the training data was much smaller, containing only Davidson and Founta datasets; the sentiment labels were produced by VADER (Gilbert & Hutto, 2014); their model was deeper and used general word embeddings (Mikolov et al., 2013; Pennington, Socher & Manning, 2014; Wieting et al., 2015) and topic representation computed through Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003) in parallel. Through ablation studies, they showed that sentiment embeddings were beneficial for both Davidson and Founta datasets.

Use of existing knowledge from a more mature research field like that of sentiment analysis has the potential to be used to jumpstart the relatively newer field of hate speech detection. It also offers a compromise between hate speech models, which might not be generalisable enough, and completely domain-general models, which lack knowledge specific to hate speech detection. Nonetheless, more investigation into the conditions in which transfer learning works best to increase generalisability in particular still needs to be done. In addition to a limited size, datasets are also prone to biases. Non-random sampling and subjective annotations introduce individual biases, and the different sampling and annotation processes across datasets further increase the difficulty of training models that can generalise across heterogeneous data.

Hate speech and, more generally, offensive language generally represent less than 3% of social media content (Zampieri et al., 2019b; Founta et al., 2018). To alleviate the effect of scarce positive cases on model training, all existing social media hate speech or offensive content datasets used boosted (or focused) sampling with simple heuristics.

Turthermore, machine learning models should be considered as part of a sociotechnical system, instead of an algorithm which only exists in relation to the input and outcomes (Selbst et al., 2019). Thus, more future work can be put into studying hate speech detection models in a wider context of application. For example, can automatic models practically aid human moderators in content moderation? In that case, how can human moderators make use of the outputs or post-hoc feature analysis(e.g., Kennedy et al. (2020)) most effectively? Would that introduce more bias or reduce bias in content moderation? Would such effects differ across different hate expressions? What would the impact be on the users of the platform? To answer these questions, interdisciplinary collaboration is needed.

4. Conclusion

Existing hate speech detection models generalise poorly on new, unseen datasets. Cross-dataset testing is a useful tool to more realistically evaluate model generalisation performance, but the problem of generalisability does not stop there. Reasons why generalisable hate speech detection is hard come from limits of existing NLP methods, dataset building, and the nature of online hate speech, and are often intertwined. The behaviour of social media users and especially haters poses extra challenge to established NLP methods. Small datasets make deep learning models prone to overfitting, and biases in datasets transfer to models. While some biases come from different sampling methods or definitions, others merely reflect long-standing biases in our society. Hate speech evolves with time and context, and thus has a lot of variation in expression. Existing attempts to address these challenges span across adapting state-of-the-art in other NLP tasks, refining data collection and annotation, and drawing inspirations from domain knowledge of hate speech. More work can be done in these directions to increase generalisability in two main directions: data and models. At the same time, wider context and impact should be carefully considered. Open-sourcing and multilingual research are also important.

References

[1] Agrawal S, Awekar A. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In: European conference on information retrieval. Grenoble, France. Springer. 141-153

[2] Al-Hassan A, Al-Dossari H. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In: Computer Science & Information Technology (CS & IT). Chennai, India. AIRCC Publishing Corporation. 83-100

[3] Alatawi HS, Alhothali AM, Moria KM. 2020. Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT.

[4] Alorainy W, Burnap P, Liu H, Williams ML. 2019. The enemy among us: detecting cyber hate speech with threats-based othering language embeddings. ACM Transactions on the Web 13(3):1-26

[5] Arango A, Pérez J, Poblete B. 2020. Hate speech detection is not as easy as you may think: a closer look at model validation (extended version) Information Systems Epub ahead of print 2020 30 June

[6] Badjatiya P, Gupta M, Varma V. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: Liu L, White RW, Mantrach A, Silvestri F, McAuley JJ, Baeza-Yates R, Zia L, eds. The World Wide Web conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019. ACM. 49-59

[7] 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. 759-760

[8] Banko M, MacKeen B, Ray L. 2020. A unified taxonomy of harmful content. In: Proceedings of the fourth workshop on online abuse and harms. Association for Computational Linguistics. 125-137

[9] Basile V. 2020. It's the end of the gold standard as we know itOn the impact of pre-aggregation on the evaluation of highly subjective tasks. In: CEUR workshop proceedings. 10

[10] Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M. 2019. SemEval-2019 Task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th international workshop on semantic evaluation. Minneapolis, Minnesota, USA. Association for Computational Linguistics. 54-63

[11] Baziotis C, Pelekis N, Doulkeridis C. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). Vancouver, Canada. Association for Computational Linguistics. 747-754

[12] Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3(Jan):993-1022

[13] Blodgett SL, Green L, OConnor B. 2016. Demographic dialectal variation in social media: a case study of African-American English. In: Proceedings of the 2016 conference on empirical methods in natural language processing. 1119-1130

[14] Blodgett SL, O'Connor B. 2017. Racial disparity in natural language processing: a case study of social media African-American English.

[15] Bodapati S, Gella S, Bhattacharjee K, Al-Onaizan Y. 2019. Neural word decomposition models for abusive language detection. In: Proceedings of the third workshop on abusive language online. Florence, Italy. Association for Computational Linguistics. 135-145

[16] Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5:135-146

[17] Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai AT. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R, eds. Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016, December 5-10, 2016, Barcelona, Spain. 4349-4357

[18] Breitfeller L, Ahn E, Jurgens D, Tsvetkov Y. 2019. Finding microaggressions in the wild: a case for locating elusive phenomena in social media posts. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong, China. Association for Computational Linguistics. 1664-1674

[19] Buolamwini J, Gebru T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. 77-91

[20] Cao R, Lee RK.-W, Hoang T.-A. 2020. DeepHate: hate speech detection via multi-faceted text representations. In: 12th ACM conference on web science, WebSci '20. New York, NY, USA. Association for Computing Machinery. 11-20

[21] Caruana R. 1997. Multitask learning. Machine Learning 28(1):41-75

[22] Caselli T, Basile V, Mitrović J, Kartoziya I, Granitzer M. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: Proceedings of the 12th language resources and evaluation conference. Marseille, France. European Language Resources Association. 6193-6202

[23] Caselli T, Basile V, Mitrovi J, Granitzer M. 2021. HateBERT: retraining BERT for abusive language detection in english.

[24] Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, StJohn R, Constant N, Guajardo-Cespedes M, Yuan S, Tar C+2 more. 2018. Universal sentence encoder for english. In: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. Brussels, Belgium. Association for Computational Linguistics. 169-174

[25] Chen H, McKeever S, Delany SJ. 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In: Staab S, Koltsova O, Ignatov DI, eds. Social informatics, Lecture notes in computer science. Cham: Springer International Publishing. 117-133

[26] Chen H, McKeever S, Delany SJ. 2019. The use of deep learning distributed representations in the identification of abusive text. Proceedings of the International AAAI Conference on Web and Social Media 13:125-133

[27] Chung Y-L, Kuzmenko E, Tekiroglu SS, Guerini M. 2019. CONAN - Counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy. Association for Computational Linguistics. 2819-2829