A Significant Analysis Machine Learning Method for Detection of Hate Speech on Social Media

Bishnu Gupta[1], Chetan Agrawal[2], Pawan Meena[3]
Dept. of CSE, Radharaman Institute of Technology & Science, Bhopal, India[1, 2, 3]
bishnugupta2k12@gmail.com[1], chetan.agrawal12@gmail.com[2], pawan191423@gmail.com[3]

Abstract: *Social networking is a low-cost way to communicate with millions of individuals. Because of this, anyone may produce anything on these platforms, and everyone can acquire it, which is a revolutionary revolution in our society. Social media platforms have great potential, but they can allow harmful discourses. This problem includes bullying, insulting content, and hate speech. Many countries quickly understand that hate speech is an issue. It's tough to erect barriers on the internet to prevent the spread of hate between countries or among ethnicities. This is the first systematic, large-scale research of hate speech in online social media. Our goal is to study the prevalence of hate speech in online social media, the most common manifestations of hate, the impact of anonymity on hate speech, and the most loathed groups in various geographic areas. This survey describes the area's state. It presents a methodical examination of past attempts, encompassing fundamental algorithms, methodology, and essential qualities.*

Keywords: *Hate Speech, Social media, Machine Learning, CNN, BERT.*

1. Introduction

Online social media sites today allow users to freely communicate at nearly marginal costs. Increasingly users leverage these platforms not only to interact with each other but also to share the news. While the open platforms provided by these systems allow users to express themselves, there is also a dark side of these systems. Particularly, these social media sites have become a fertile ground for inflamed discussions, that usually polarize 'us' against 'them', resulting in many cases of insulting and offensive language usage.

Another important aspect that favors such behavior is the level of anonymity that some social media platforms grant to users. For example, "Secret" was created, in part, to promote free and anonymous speech but became a means for people to defame others while remaining anonymous. The secret was banned in Brazil for this very reason and shut down in 2015 1. There are reports of cases of hateful messages in many other social media independently of the level in which the online identity is bonded to an offline identity – e.g., in Whisper, Twitter, Instagram, and Facebook.

With this context, it is not surprising that most existing efforts are motivated by the impulse to detect and eliminate hateful messages or hate speech [1, 2]. These efforts mostly focus on specific manifestations of hate, like racism [3]. While these efforts are quite important, they do not attempt to provide a big picture of the problem of hate speech in the current popular social media systems. Specifically providing a broad understanding of the root causes of online hate speech was not the main focus of these prior works. Consequently, these prior works also refrain from suggesting broad techniques to deal with the generic offline hate underlying online hate speech.

In this paper, we take the first step towards a better understanding of online hate speech. Our effort consists of characterizing how hate speech is spread in common social media, focusing on understanding how hate speech manifests itself under different dimensions such as its targets, the identity of the haters, geographic aspects of hate contexts. Particularly, we focus on the following research questions.

What is hate speech about? We want to understand not only which the most common hated groups of people are, but also what are the high-level categories of hate targets in online hate speech is.

What role does anonymity play on hate speech? Is anonymity a feature that exacerbates hate speech or is social media users not worried about expressing their hate under their real names? What fraction of haters uses their names in social media?

How does hate speech vary across geography? Does hate speech targets vary across countries? And, within states of a country like the USA? Are there categories of hate speech that are uniformly hated and others that are hated only in specific regions?

Answering these questions is crucial to help authorities (including social media sites) for proposing interventions and effectively deal with hate speech. To find answers, we gathered one-year data from two social media sites: Whisper and Twitter. Then, we propose and validate a simple yet effective method to detect hate speech using sentence structure and using this method to construct our hate speech datasets. Using this data, we conduct the first of a kind characterization study of hate speech along multiple different dimensions: hate targets, the identity of haters, geographic aspects of hate, and hate context. Our results unveil a set of important patterns, providing not only a broader understanding of hate speech but also offering directions for detection and prevention approaches.

Rest of the paper is organized as follow: in section II we explained related work done previously in the field of Hate speech detection on social media like Twitter etc., in section III we discussed about various approaches used in detection of Hate Speech, section IV presents some challenges and opportunities in the field of hate speech detection, finally we conclude this paper with future research directions in section V followed by references used in this paper.

2. Related Work

In the past few years, research on hate speech content in multimedia formats has started to appear, but the body of previous work is still rather small. In the year 2020, the Facebook challenge on vile memes received a great amount of attention from scholars. The top three winning teams utilized pre-trained multimodal transformer models to successfully merge the visual characteristics of the image with the textual characteristics of the caption [4], [5], and [6]. [4], [5], and [6]. A far more recent effort was also undertaken in the realm of video hate speech identification [7]. [Show citation] [Show citation] On the other hand, it is solely concerned with the text component of the movie and ignores any extra functions that the multimedia data can supply. Another piece of study with the objective of identifying offensive video content compiled and published a dataset in Portuguese [8]. In order to identify inappropriate content, it examined social network properties like tags and titles, in addition to transcripts. In spite of this, such models are reliant on the capabilities that become available after the content has been disseminated to a wider audience, which causes harm to the group that was originally intended to benefit from it. As a result of this, there is a pressing requirement for a more advanced approach of detecting hate speech in multimedia data.

In order to do feature extraction on text, we first train a language model for the detection of hate speech. Over the course of the last ten years, the NLP techniques for identifying hate speech have progressed through a number of stages. The first attempts at author profiling utilized TFIDF [5], bag-of-words (BOW) or n-gram [4], user-specific features such as age, and social media features such as shares, retweets, and reports [6], [9]. In recent years, the majority of research attention has been concentrated on various neural architectures due to the maturation of various deep learning methodologies. Recurrent neural networks (RNNs) and convolution neural networks (CNNs) were originally used for the purpose of detecting hate speech in tweets by Badjatiya et al. [10] and Gamback et al. [11], respectively. The most cutting-edge models now available can fine-tune pre-trained transformers such as BERT and ALBERT. There was significant variance in the parameters and pre-processing that BERT uses [12, 13], but seven out of the top 10 offensive language detection task teams used it in 2019. Again in the year 2020 [14], the top ten teams employed various combinations of BERT, RoBERTa, or XML-RoBERT, and the team that ended up victorious used ALBERT [15]. There has been a significant amount of research conducted in the field of Speech Emotion Recognition (SER), which has a wide number of applications including Human-Computer Interaction [16], Sentiment Analysis, and Enhancing cinema sound design [17]. Research can be divided into two categories based on whether it classifies the speech into an emotional state (such as happiness, sadness, anger, fear, disgust, or boredom) [18] or whether it predicts the emotional attributes, such as valence, arousal, and dominance [19]. [18] Research can be divided into two categories based on whether it classifies the speech into an emotional state (such as happiness, sadness, anger, fear, disgust, or boredom). Zisad et al. [20] developed a CNN model based on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data. This model was merged with some locally generated problem specific data that was collected by them in order to classify the speech emotion as either happy, surprised, angry, fearful, disgusted, neutral, or sad. In this study, features for the model include things like MFCCs and other tonal aspects. An attention-based CNN model that was trained on the speech spectrogram as the input rather than acoustic or statistical data was proposed by Zhang et al. [21]. This model was thought to be capable of producing better outcomes. It is important to highlight that another focus of the research is the archetypal classification of different emotions. On a related point, Weiser et al. [22] examines the differences and similarities between an end-to-end learning network that was trained on raw audio data and a feature based network. In their call center-based SER system, Bojanic et al. [23] emphasis on archetype emotions in addition to the emotional characteristics of speech. According to the findings of the research conducted by Parthasarathy and Busso [24], the emotional characteristics are interrelated; as a result, using a unified learning framework to predict these characteristics will provide us more accurate results. In addition to using models created with deep neural networks, they employ multitask learning.

3. Approaches in Hate Speech Detection

In this section, we analyze features described in the papers focusing on algorithms for hate speech detection, and also other studies focusing on related concepts (e.g., Cyber bullying). Finding the right features for a classification problem can be one of the more demanding tasks when using machine learning. Therefore, we allocate this specific section to describe the features already used by other authors. We divide the features into two categories: general features used in text mining, which are common in other text mining fields; and the specific hate speech detection features, which we found in hate speech detection documents and are intrinsically related to the characteristics of this problem. We present our analysis in this section.

General Features Used in **Text Mining:** The majority of the papers we found to try to adapt strategies already known in text mining to the specific problem of automatic detection of hate speech. We define general features as the features commonly used in text mining. We start by the most simplistic approaches that use dictionaries and lexicons.

Dictionaries: One strategy in text mining is the use of dictionaries. This approach consists of making a list of words (the dictionary) that are searched and counted in the text. These frequencies can be used directly as features or to compute scores. In the case of hate speech detection, this has been conducted using:

Content words (such as insults and swear words, reaction words, and personal pronouns) collected from www.noswearing.com [25].

The number of profane words in the text, with a dictionary that consists of 414 words, including acronyms and abbreviations, where the majority is adjectives and nouns.

Label Specific Features that consisted of using frequently used forms of verbal abuse as well as widely used stereotypical utterances.

Ortony Lexicon was also used for negative affect detection; the Ortony lexicon contains a list of words denoting a negative connotation and can be useful, because not every rude comment necessarily contains profanity and can be equally harmful.

This methodology can be used with an additional step of normalization, by considering the total number of words in each comment. Besides, it is also possible to use this kind of approach with regular expressions [26].

Distance Metric. Some studies have pointed out that in text messages the offensive words may be obscured with an intentional misspelling, often a single character substitution. Examples of these terms are "@ss," "sh1t", "nagger," or homophones, such as "Joo". The Levenshtein distance, i.e., the minimum number of edits necessary to transform one string into another can be used for this purpose. The distance metric can be used to complement dictionary-based approaches.

Bag-of-words (BOW). Another model similar to dictionaries is bag-of-words. In this case, a corpus is created based on the words that are in the training data, instead of a pre-defined set of words, as in the dictionaries. After collecting all the words, the frequency of each one is used as a feature for training a classifier. The disadvantages of this kind of approach are that the word sequence is ignored, and also it's syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation N-grams can be adopted.

N-grams. N-grams are one of the most used techniques in hate speech automatic detection and related tasks. The most common N-grams approach consists in combining sequential words into lists with size N. In this case, the goal is to enumerate all the expressions of size N and count all occurrences. This allows improving classifiers' performance because it incorporates at some degree the context of each word. Instead of using words, it is also possible to use N-grams with characters or syllables. This approach is not so susceptible to spelling variations as for when words are used. Character N-gram features proved to be more predictive than token N-gram features, for the specific problem of abusive language detection.

However, using N-grams also have disadvantages. One disadvantage is that related words can have a high distance in a sentence and a solution for this problem, such as increasing the N value, slows down the processing speed. Also, studies point out that higher N values (5) perform better than lower values (unigrams and trigrams). In a survey researchers report that N-grams features are often reported to be highly predictive in the problem of hate speech automatic detection, but perform better when combined with others.

Profanity Windows: Profanity windows are a mixture of a dictionary approach and N-grams. The goal is to check if a second person pronoun is followed by a profane word within the size of a window and then create a Boolean feature with this information.

TF-IDF: The TF-IDF (term frequency-inverse document frequency) was also used in this kind of classification problems. TF-IDF is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times that a word appears in the document. However, it is distinct from a bag of words, or N-grams, because the frequency of the term is off-settled by the frequency of the word in the corpus, which compensates the fact that some words appear more frequently in general (e.g., stop words).

Part-of-speech: Part-of-speech (POS) approaches to make it possible to improve the importance of the context and detect the role of the word in the context of a sentence. These approaches consist in detecting the category of the word, for instance, personal pronoun (PRP), Verb non-3rd person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part- of-speech has also been used in hate speech detection problems. With these features, it was possible to identify frequent bigram pairs, namely PRP_VBP, JJ_DT, and VB_PRP, which would map as “you are”. It was also used to detect sentences such as “send them home,” “get them out,” or “should be hung”. However, POS proved to confuse the class identification, when used as features.

Lexical Syntactic Feature-based (LSF): In a study, the natural language processing parser, proposed by Stanford Natural Language Processing Group was used to capture the grammatical dependencies within a sentence. The features obtained are pairs of words in the form “(governor, dependent)”, where the dependent is appositional of the governor (e.g., “You, by any means, an idiot.” means that “idiot,” the dependent, is a modifier of the pronoun “you,” the governor). These features are also being used in hate speech detection.

Rule-Based Approaches: Some rule-based approaches have been used in the context of text mining. A class association rule-based approach, more than frequencies, is enriched by linguistic knowledge. Rule-based methods do not involve learning and typically rely on a pre-compiled list or dictionary of subjectivity clues. For instance, rule-based approaches were used to classify antagonistic and tense content on Twitter using associational terms as features. They also included accusational and attributional terms targeted at only one or several persons following a socially disruptive event as features, to capture the context of the terms used.

Participant - Vocabulary Consistency (PVC): In a study about cyber bullying, this method is used to characterize the tendency of each user to harass or to be harassed, and the tendency of a key phrase to be indicative of harassment. For applying this method it is necessary a set of messages from the same user. In this problem, for each user, it is assigned a bully score (b) and a victim score (v). For each feature (e.g., N-grams) a feature-indicator score (w) is used. It represents how much the feature is an indicator of a bullying interaction. Learning is then an optimization problem over parameters b, v, and w.

Template Based Strategy: The basic idea of this strategy is to build a corpus of words, and for each word in the corpus, collect K words that occur around. This information can be used as a context. This strategy has been used for feature extraction in the problem of hate speech detection as well. In this case, a corpus of words and a template for each word was listed, as in “W-1: go W+0: back W+1: to.” This is an example of a template for a two-word window on the word “back.”

Word Sense Disambiguation Techniques: This problem consists of identifying the sense of a word in the context of a sentence when it can have multiple meanings. In a study, the stereotyped sense of the words was considered, to understand if the text is anti-Semitic or not.

Typed Dependencies: Typed dependencies were also used in hate speech related studies. First, to understand the type of features that we can obtain with this, the Stanford typed dependencies representation describes the grammatical relationships in a sentence that can be used by people without linguistic expertise. These were used for extracting Theme-based Grammatical Patterns and also for detecting hate speech specific other language that we will present within the specific hate speech detection features. Some studies report significant performance improvements in hate speech automatic detection based on this feature.

Topic Classification: With these features, the aim is to discover the abstract topic that occurs in a document. In a particular study, topic modeling linguistic features was used to identify posts belonging to a defined topic (Race or Religion).

Sentiment: Bearing in mind that hate speech has a negative polarity, authors have been computing the sentiment as a feature for hate speech detection Different approaches have been considered (e.g., multi-step, single-step) Authors usually use this feature in combination with others that proved to improve results.

Word Embeddings: Some authors use a paragraph2vec approach to classify language on user comments as abusive or clean and also to predict the central word in the message. Fast Text is also being used. A problem that is referred to in hate speech detection is that sentences must be classified and not words. Averaging the vectors of all words in a sentence can be a solution; however, this method has limited effectiveness. Alternatively, other authors propose comment embeddings to solve this problem.

Deep Learning: Deep learning techniques are also recently being used in text classification and sentiment analysis, with high accuracy.

Other Features: Other features used in this classification task were based in techniques such as Named Entity Recognition (NER), Topic Extraction, Word Sense Disambiguation Techniques to check Polarity, frequencies of personal pronouns in the first and second person, the presence of emoticons and capital letters. Before the feature extraction process, some studies have also used stemming and removed stop-words. Characteristics of the message were also considered such as hash tags, mentions, retweets, URLs, number of tags, terms used in the tags, number of notes (re-blog and like count), and link to multimedia content, such as image, video, or audio attached to the post.

4. Research Challenges and Opportunities

Hate speech is a complex phenomenon and its detection problematic. Some challenges and difficulties were highlighted by the authors of the surveyed papers:

- Low agreement in hate speech classification by humans, indicating that this classification would be harder for machines.
- The task requires expertise in culture and social structure.
- The evolution of social phenomena and language makes it difficult to track all racial and minority insults.
- Language evolves quickly, in particular among young populations that communicate frequently in social networks.
- Despite the offensive nature of hate speech, an abusive language may be very fluent and grammatically correct, can cross sentence boundaries, and the use of sarcasm in it is also common.
- Finally, hate speech detection is more than simple keyword spotting.

We find it relevant to present those difficulties so that we bear in mind the kind of challenges that researchers face in their work.

5. Conclusion

In this survey, we offered a critical review of how the automatic detection of hate speech in text has grown over the past few years. This was done in response to a survey that was conducted by the National Center for Hate Studies. To begin, we investigated the meaning of hate speech in a variety of settings, ranging from the platforms of social networks to those of other organizations. On the basis of our research, we proposed a unified and more precise definition of this idea, which we believe can assist in the construction of a model for the automatic identification of hate speech. In addition, we provided examples and rules for classification that were identified in the literature, together with the arguments in support of or in opposition to those principles. Our critical view highlighted the fact that our definition of hate speech is both more inclusive and general than other perspectives found in the literature. [Citation needed] [Citation needed] This is the situation as a result of our suggestion that more covert forms of prejudice on the internet and in online social networks should also be identified and addressed. As a result of our investigation, we came to the realization that it would be useful to examine hate speech alongside other problematic phenomena such as cyber bullying, abusive language, discrimination, toxicity, flame, extremism, and radicalization. Through this comparison, we were able to see how hate speech is separate from these other similar notions, which assisted us in comprehending the bounds of its definition as well as its nuances.

After doing an in-depth analysis of the relevant published material, we came to the conclusion that, from a computing and informatics point of view, there have not been very many studies or articles published on the topic of automatic hate speech detection. The majority of the efforts that have been done on the topic view the issue as a classification assignment for machine learning. In this area of study, researchers typically begin their work by collecting and annotating fresh communications; however, these datasets are typically kept confidential. Because of this, the progress of the research is slowed down because there is fewer data available, and as a result, it is more difficult to compare the results of other investigations. Despite this, we were able to locate three datasets, two of which were written in English and one in German. In addition, we assessed the relative merits of the several studies that made use of various algorithms to identify instances of hate speech and ranked them accordingly. Our objective was to identify the strategies that proved to be the most successful so that we could draw some conclusions. On the other hand, and in part because there aren't any standard datasets, we discover that the various publications don't point to a single strategy that has been shown to get superior results than the others.

Reference

[1] Swati Agarwal and Ashish Sureka. 2015. Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter. In Proceedings of The 11th International Conference on Distributed Computing and Internet Technology (ICDCIT'15).

[2] J. Bartlett, J. Reffin, N. Rumball, and S. Williamson. 2014. Anti-social media. DEMOS

[3] Irfan Chaudhry. 2015. #Hashtagging hate: Using Twitter to track racism online. First Monday 20, 2 (2015).

[4] R. Zhu, "Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution," 2020.

[5] N. Muennighoff, "Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes," 2020.

[6] R. Velioglu and J. Rose, "Detecting Hate Speech in Memes Using Mul- timodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge," 2020.

[7] C. S. Wu and U. Bhandary, "Detection of Hate Speech in Videos Using Machine Learning," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020, pp. 585–590. [Online]. Available: 10.1109/CSCI51800.2020.00104

[8] C. Alcântara, V. Moreira, and D. Feijo, “Offensive Video Detection: Dataset and Baseline Results,” in Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 4309–4319.

[9] P. Mishra, M. D. Tredici, H. Yannakoudakis, and E. Shutova, “Author Profiling for Abuse Detection,” in Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, August 2018, pp. 1088–1098.

[10] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep Learning for Hate Speech Detection in Tweets,” 06 2017. [Online]. Available: 10.1145/3041021.3054223

[11] B. Gambäck and U. K. Sikdar, “Using Convolutional Neural Networks to Classify Hate-Speech,” in Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, August 2017, pp. 85–90. [Online]. Available: 10.18653/v1/W17-3013

[12] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval),” in Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 75–86. [Online]. Available: 10.18653/v1/S19-2010

[13] P. Liu, W. Li, and L. Zou, “NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers,” in Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 87–91. [Online]. Available: 10.18653/v1/S19-2011

[14] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020),” in Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona (online): International Committee for Computational Linguistics, December 2020, pp. 1425–1447. [Online]. Available: 10.18653/v1/2020.semeval-1.188

[15] G. Wiedemann, S. Yimam, and C. Biemann, “UHH-LT at SemEval-2020 Task 12: Fine-tuning of pre-trained transformer networks for offensive language detection,” Proceedings of the International Workshop on Semantic Evaluation (SemEval), 2020.

[16] S. Ramakrishnan and I. M. E. Emary, “Speech emotion recognition approaches in human computer interaction,” Telecommun Syst, vol. 52, pp. 1467–1478, 2013. [Online]. Available: 10.1007/s11235-011-9624-z

[17] S. Cunningham, H. Ridley, and J. Weinel, “Supervised machine learning for audio emotion recognition,” Pers Ubiquit Comput, vol. 25, pp. 637–650, 2021. [Online]. Available: 10.1007/s00779-020-01389-0

[18] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, “Automatic Speech Emotion Recognition: A survey,” in 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014, pp. 341–346. [Online]. Available: 10.1109/CSCITA.2014.6839284

[19] K. Sridhar and C. Busso, “Modeling Uncertainty in Predicting Emotional Attributes from Spontaneous Speech,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8384–8388. [Online]. Available: 10.1109/ICASSP40776.2020.9054237

[20] S. N. Zisad, M. S. Hossain, and K. Andersson, “Speech emotion recog- nition in neurological disorders using Convolutional Neural Network,” Proceedings of the 13th International Conference on Brain Informatics (BI2020), pp. 287–296, 2020.

[21] Y. Zhang, J. Du, Z. Wang, J. Zhang, and tu Yanhui, “Attention Based Fully Convolutional Network for Speech Emotion Recognition,” 11 2018, pp. 1771–1775. [Online]. Available: 10.23919/APSIPA.2018. 8659587

[22] I. Wieser, P. Barros, and S. Heinrich, “Understanding auditory representations of emotional expressions with neural networks,” Neural Comput & Applic, vol. 32, pp. 1007–1022, 2020. [Online]. Available: 10.1007/s00521-018-3869-3

[23] “Call Redistribution for a Call Center Based on Speech Emotion Recognition,” Applied Sciences, vol. 10, no. 13, 2020. [Online]. Available: 10.3390/app10134653

[24] S. Parthasarathy and C. Busso, “Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning,” in Proc. Interspeech 2017, 2017, pp. 1103–1107. [Online]. Available: 10.21437/Interspeech.2017-1494

[25] Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K’15), Vol. 1. IEEE, 487–495

[26] Wilson Jeffrey Maloba. 2014. Use of Regular Expressions for Multi-lingual Detection of Hate Speech in Kenya. Ph.D. Dissertation. iLabAfrica