A Literature Survey on Hate Speech Detection

Manohar GS, Jozef Juhár, Daniel Hladek
Department of Electronics and Multimedia Telecommunications
Technical University of Kosice, Slovakia
Manohar.gowdru.shridhara@tuke.sk, Jozef.Juhar@tuke.sk, daniel.hladek@tuke.sk

Abstract – The understanding of hate speech detection for beginners. This scientific paper deals with hate speech detection basics and to know where, why, and how it can be used to detection of hate speech and brief idea about methods which are used for hate speech detection and how can we evaluate a hate speech detection system and to know about datasets which are available, problems and previous research, methods, results, algorithms, features are used. Hate speech is available in many different contents like video images texts in real-time streaming and many other problems and challenges, like a different language, mixed content with many words which are keep updating day by day, new ways to insult with different words. Phrases with contexts that are in different dimensions, hiding many other meanings behind the words, adding them to the black list, and understanding them to prepare learning data models are also problem by using automatic hate speech detection and removing contents. Identifying hate speech is become complex due to the above concerns, to this we are preparing a literature survey to understand the approaches and results by the researcher to find out the best type of datasets for upcoming research activities and an "extensive and organized understanding of automatic hate speech detection, focused at NLP and ML scientist who needs an introduction to the field of hate speech detection."

Keywords— natural language processing; hate-speech; artificial intelligence.

I. Introduction

Speech is a very important part of life; it is become widely used by any animal on the planet. Speech helps communicate between two animals that can understand each other or one to one or many to many. Considering to mankind, speech is a mode of communication to understand the intention, emotion, action, and more. Using speech plays a major role and speech is able to create an emotion in the person which is good and bad communication which leads to better or hate speech. The growing world due to a huge diversity of people are started using social media and sharing information has given major benefits to humanity. Similarly facing some challenges in terms of sharing hate speech and messages. Hate speech detection overcomes the propagation of hateful content.

II. HATE Speech Detection

Hate Speech is a speech that targets an individual or a community group on the basis of parameters such as culture, religion, age, nationality, color, gender, status, sexual, gender identity, and disability [1].

Social media platforms like Facebook, Twitter have raised concerns about an emerging dubious activity such as the intensity of hate, abusive and offensive behavior [2]. One of the breakthroughs on the internet is social media and blogging. Mankind uses internet and social media for blogging, writing articles and sharing media files and content, reacting to the posts and sharing their opinions [4].

The definition of hate speech is post, content, language, the image on social media. With malicious intentions of spreading hate, being derogatory, encouraging violence, or aiming to dehumanize (comparing people to non-human things e.g., animals) insult, promote or justify, hatred, discrimination, or hostility [5].

III. WHERE AND WHY YOU CAN USE HATE-SPEECH DETECTION

A. *where you can use hate-speech detection*

Generally, more often using is social media and internet websites and media clips and recently world biggest countries had elections, political interest, gender, race, religion, disability, cultures, and many countries in case of any incidents which are going separated wrong message to the society than shut the internet to avoid hate speech content and some hate speech content leads to crimes and society facing many challenges.

B. *why you can use hate-speech detection*

In the case of hate speech due to access to the internet spreading of the hate content is so high to avoid these issues. Hate speech detection plays a major role and its importance. The solution is to detect the hate speech at the earliest and report it. The solution must be automated and due to the concerns manual detection facing challenges and widespread hate speech content on the internet.

IV. State-of-the-art methods for hate speech detection

A wide range of methodologies have been evolved for automated hate speech detection content online. Considering the existing definitions and complex methods the ideology is to build a new method to detect the content automatically. Different state-of-the-art methods are required in some cases.

The verity of data content over internet exists with range of data with different language and context, different styles which consist of some proportion of hate speech detection required. Considering the text-based classification approaches goes beyond its capacity to capture like Encyclopedia, Facebook, Twitter, Davidson et al Language group, de Gilbert et al. direct attacking a specific group, Fortuna et al. violence or hate against based or characteristics physical appearance religion, etc. and in some case real-world fact verification of sentence proposed keywords are really hate speech due to the context of meaning which is created with respect to the situation which related group of the population required further complicating to detect the hate speech[6].

Over the years of the research, the focus of data sources systematically updates to date and organized and significantly the systematic survey consists helping and identifying the gaps in the current research and further investigation and exploring the robust framework for improving the research on hate speech detection [7].

This contribution leads to proposed in this field to prepare benchmark datasets and how data extracted and simplified with multi-languages and flow of evidence in some potential sources are overlooked and finding out the areas of interest which is including and excluding the criteria must have the border and key points for this hate speech datasets and hate speech focus on characteristics computational linguistics details helps build solid framework [6, 7].

V. Evaluate a hate-speech detection system

There are a range of hate speech detection methodologies that exists to determine with respect to the distinct type of data. Assessing the performance of hate speech detection system building the model and classifiers for the hate speech data and choosing the model in machine, deep learning [8] used for the hate and non-hate detection. Twitter datasets contains wide range of characters, emotions and special tags that can be validated using convolution Neural Network.

Adding the classifiers remove the code-mixed text [8] which is the main challenge to handle the weight of hate and non-hate data thus weighting the data leads the model to be trained more accurately.

The machine learning classifiers are, namely a support K-Nearest Neighbors (KNN), vector machine (SVM), multinomial n Bayes (MNB), and a decision tree (DT). Term frequency (TF) used for character-level CNN model character level model gives a better performance than all other classifiers.

Corpus Creation and Annotation in social media data best known for code-mixing which is from Facebook, Twitter social media. These datasets helps to identify the hate speech automatically.

There are few models used for text mining and labeling with the help of SVM Support Vector Machine and Logistic Regression, Machine Learning classifiers which predicts scores features by the help python ML packages.

Neural Ensemble helps combine average scores, Fast Tex Text from the Facebook fast categorization of text. BERT additional pre-trained model embedding as output layer achieves a state of the art performance in text classification and language inference. C-GRU deep neural network CNN model.

Corpus Creation and Annotation in social media data best known for code-mixing which is from Facebook, Twitter social media. This kind of data and methods can lead to an invaluable resource for understanding as well as automatically identifying hate speech. As hate speech detection is reasonable accuracy in traditional approaches to building automatic detection of hate speech system to benefit to improve interpretability and in this way evaluating both technical and practical matters.

A. List of methods

here carried out on the two directions one is what are the methods are used to detect automatic hate speech detection and what kind of datasets are been widely used.

Widely used methods are Paragraph2Vec, Term Frequency–Inverse Document Frequency, Long Short-Term Memory, Gradient Boosted Decision Trees, FastText, Convolutional Neural Network, Random Forest Decision Tree, Support Vector Machine, Logistic Regression J48 Graft, Naive Bayes, XGBoost.

Simple Surface features: Paragraph2Vec method contains component called Continuous Bag of Words (CBOW) model which analysis the type of word and takes the central words to detect the hate speech content or not. The results of this method are promising and considered as one of the best predictive method and these features are often used [8].

Character-level n-gram features is more predictive and effective approach as this method reduces the effect of spelling variation issues faced with online datasets generated by the social media comments by the users [9].

Word- and character-based features deal with URL, Uppercase lowercase, English dictionaries and numbers, non-alphanumeric characters, etc.

Word Generalization: Bag of words features as an efficient classifier in the hate speech detection using predictive words in test data and training data and this method works efficiently on less data and short label texts. It can deal with word embedding’s with domain-specific.

Sentiment Analysis: Hate speech can be correlated to Sentiment Analysis as it can be assumed that negative sentiments often concerns to a hate speech, this method characterizes the amount of negative, neutral and positive words on a data set according to sentiment lexicon

Lexical Resources: In general hate speech words which leads to negative meanings which are hateful texts. To get and filter these kind of information contains predictive structures.

Linguistic Features: basically customization for the particular tasks for detection of hate speech.

Data and Annotation: corpus is standard data which is used by researchers store and append labels and customize with own data and able to explore on hate speech detention to access the labelled data from corpora.

Meta-Information: stored and explored by the social media datasets and able to consume by APIs which wide range of different format of the data can be explored and it can be easily accessed.

Multimodal Information: today’s world contains social media content in different way of presentation, like videos, texts, and images, audio. These kind of the data leads to multi domination data and predictive features.

Classification Methods: Classifiers are most commonly used for hate speech detection which are used for Machine leaning process and methods. recent methods are deep leaning, Convolutional Neural Networks CNN, Deep Neural Networks DNN, Long short Term memory LSTM and many other methods are used to build improve the detection of hate speech and combining methods can also enhance the results and accuracy[12]. The researchers are more indent towards data and datasets compared to model [13].

Hate2Vec is methods where to detect offensive language in hate speech detection datasets. Twitter comments which is annotated labeled data and Hate2Vec average F1-Score of 0.93.

HateDefender which is novel based system on deep LSTM neural network and accuracy is high that is F1-Scores 90.82%. The datasets consists Twitter labeled as offence language and hate speech detection [16].

Social networks datasets Methods LSTM and BiLSTM, are embedding data F1-0.90 which contains datasets [17].

Crowdflower is datasets which is hateful offensive language and accuracy is 78.4% on tweet datasets [18].

Tweets using ML approach from South African datasets from twitter and Machine learning approach method us used SVM and Logistic Regression and character n-gram achieved the best positive rate is 0.894 [19].

Offensive language detection in different languages, LogitBoost, which is based on the AdaBoost procedure that trains the model on weighted samples and F1-score of 99.2% on Roman Urdu dataset using character tri-gram [20]

The method used by scientist are Naive Bayes, SVM, and Random Forest Decision Tree (RFDT). The result shows that Naive Bayes is better than SVM and RDFT with an F1-Score of 86.43% using word unigram feature extraction [21].

The mixed content Hindi-English language and The methods that were used are SVM, SVM-Radial Basis Function(SVM-RBF), and Random Forest. The result shows that SVM-RBF combined with FastText gives an F1-Score of 85.81%, higher compared to SVM-RBF combined with word2vec gives an F1-score of 75.11% [22].

Cyber hate speech on Social media like Twitter. It uses CrowdFlower to annotate the tweets that contain hate speech about disability, race, sexual orientation, or none. The classification was done using SVM and RFDT combined with BoWV. The result showed that the overall F1-Score is 0.96 [23].

Cyberspace approached considering micro data and Meta data. The dataset contains 9,488 annotated tweets. The method for classification uses RFDT. This research focuses more on the metadata rather than text variables [24].

The datasets created by multiple social media platform short cut names, the datasets keep grows from one researcher to another researcher. The result showed that XGBoost was the best classifier combined with BERT as the best feature representation with the F1-Score of 0.916 [25].

VI. DATASETS FOR HATE-SPEECH DETECTION

A set of keywords was set up and used to browse search engines and repositories. English keywords since English is used worldwide as a working language among scholars;

however, restrict the search to works based on English data alone, instead of including as many languages as possible.

HASOC 2019, OLID labeled datasets, collected and annotated at various times by various people can be helpful for datasets. HateBase is a corpus that is automated detection of hateful and offensive content and labeled data and annotated into three classes namely hate speech and offensive languages.

Social media platforms are a hotbed for hare speech most of the data from Twitter. Datasets hate-related characteristics HatebaseTwitter, WasseemA, waseemB which from Twitter in English, and Stormfront which is from online forum with English along with that TRAC from Facebook, HatEval, Kaggle, German Twitter from Twitter. These type datasets consist of different type labels and languages [6].

The different types of datasets training to solve various types of problems 2 label datasets, 3 label datasets and Balanced, Imbalanced, Number of counts datasets most of the researchers are using imbalanced datasets. Removing the repeated words that the sampling method is required to build balanced datasets helps drive for better results.

Datasets are going to prepare by collecting sampling the data by using data annotation to train the model using classifiers to detect hate speech. Preparing the datasets from the online system which are to be compared correlates with meanings and opinions with data annotations is important. Typically, datasets are built by twitter on a large scale, selecting the data by using hashtags is also a key issue of chosen data. We find many issues with unfiltered datasets like hashtags must be unambiguous and stable down the line for this kind of issues data annotation is the key solution and setting up the benchmarks process. This process can set some policies to retrieve the data from the start of hashtags and hateful content keywords with respect to the guidelines of the data annotations using data sources. The data content is any domain-specific associated with certain attributes and annotator role is more or fewer keywords like language choices, topics, and sensitive words. Online data must contain privacy policies GDPR rules are considered to start researching on these datasets [26].

Data annotations process consists of data analysis and challenges like image, video, content, and text in social media platforms, few data are difficult since data can go disappear after some time. Data annotations to automatic detections have some challenges Multilingualism, multimodality (combination of images videos, and texts), detection in context and platform-specific improve the classification results merging the data with different platforms improves classification results [26].

VII. CONCLUSION

The aim of this work was to propose a basics of hate speech detection overview for beginner on automatic detection of hate speech based on NLP, Deep learning and Machine learning. We have gone thru with different methods and datasets, its results, approaches to understand the hate speech detection. Automatic detection of hate speech and listed available methods, algorithms and its results which gives new insight for the researchers who needs an introduction to the field of hate speech detection.

ACKNOWLEDGMENT

The research in this paper was supported by the Department of Electronics and Multimedia Telecommunications
Technical University of Kosice, Slovakia.

REFERENCES

[1] Nanlir sallau mullah 1,2, (member, ieee), and Wan mohd nazmee wan zainon "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review", date of current version June 28, 2021.

[2] Sindhu Abro, Sarang Shaikh, Zafar Ali, Sajid Khan Mujtaba,"Automatic Hate Speech Detection using Machine Learning: A Comparative Study" International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020

[3] Mansi Dhawan, Dr. M.L. Sharma, "HATE SPEECH DETECTION AND SENTIMENT ANALYSIS" *Volume: 07 Issue: 05 | May 2020*.

[4] Claudia Zaghi "Automatic detection of hate speech in social media", University of Malta 2018

[5] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language" [Submitted on 11 Mar 2017]

[6] Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder Hate speech detection: Challenges and solutions". Published: August 20, 2019

[7] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco & Viviana Patti. "Resources and benchmark corpora for hate speech detection: a systematic review". Published: 30 September 2020.

[8] Priya.rani,shardul.Suryawanshi,koustava.goswami,bharathi.raja, theodorus.fransen,john.mccrae. "Comparative Study of Different State-of-the-Art Hate Speech Detection Methods for Hindi-English Code-Mixed Data". 11–16 May 2020.

[9] Calvin Erico Rudy Salim, Derwin Suhartono A Systematic Literature Review of Different Machine Learning Methods on Hate Speech Detection, VOL 4 (2020).

[10] Anna Schmidt & Michael Wiegand, A Survey on Hate Speech Detection using Natural Language Processing, Proceedings of the Fifth International Workshop on Natural Language Processing for social media, Valencia, Spain, April 3-7, 2017. c 2017 Association for Computational Linguistics.

[11] Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N. Hate speech detection with comment embeddings. InProceedings of the 24th international conference on world wide web 2015 May 18 (pp. 29-30).

[12] Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. InProceedings of the 26th International Conference on World Wide Web Companion 2017 Apr 3 (pp. 759- 760).

[13] Gröndahl T, Pajola L, Juuti M, Conti M, Asokan N. All You Need is" Love" Evading Hate Speech Detection. InProceedings

of the 11th ACM Workshop on Artificial Intelligence and Security 2018 Jan 15 (pp. 2-12).

[14] Pelle R, Alcântara C, Moreira VP. A classifier ensemble for offensive text detection. InProceedings of the 24th Brazilian Symposium on Multimedia and the Web 2018 Oct 16 (pp. 237-243).

[15] Aulia N, Budi I. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. InProceedings of the 2019 5th International Conference on Computing and Artificial Intelligence 2019 Apr 19 (pp. 164-169).

[16] Dorris W, Hu R, Vishwamitra N, Luo F, Costello M. Towards Automatic Detection and Explanation of Hate Speech and Offensive Language. InProceedings of the Sixth International Workshop on Security and Privacy Analytics 2020 Mar 16 (pp. 23- 29).

[17] Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT). 2020 Mar 14;20(2):1- 22.

[18] Watanabe H, Bouazizi M, Ohtsuki T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access. 2018 Feb 15;6:13825- 35.

[19] Oriola O, Kotzé E. Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. IEEE Access. 2020 Jan 20;8:21496-509. 218

[20] Akhter MP, Jiangbin Z, Naqvi IR, Abdelmajeed M, Sadiq MT. Automatic Detection of Offensive Language for Urdu and Roman Urdu. IEEE Access. 2020 May 15;8:91213-26.

[21] Ibrohim MO, Budi I. A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science. 2018 Jan 1;135:222-9.

[22] Sreelakshmi K, Premjith B, Soman KP. Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science. 2020 Jan 1;171:737-44.

[23] Burnap P, Williams ML. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data science. 2016 Dec 1;5(1):11.

[24] Miró-Llinares F, Moneva A, Esteve M. Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Science. 2018 Dec 1;7(1):15.

[25] Salminen J, Hopf M, Chowdhury SA, Jung SG, Almerekhi H, Jansen BJ. Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences. 2020 Dec 1;10(1):1

[26] Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, Serena Villata. A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis. In Proceedings of the 3rd Workshop on Abusive Language Online, 2019.