A Literature Survey on Hate Speech Detection

Manohar Gowdru Shridhara, 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 – Beginners' guide to hate speech detection. The purpose of this paper is to study hate speech detection basics and various methods for detecting hate speech. It also aims to find out, how to evaluate a hate speech detection system. It will also help you to understand datasets that are currently available, previous research, methods, results, algorithms, and features that are being used. Various forms of hate speech are available including videos, images, texts, and real-time streaming, among other challenges, such as different languages, mixed content with many words that are constantly changing, and various ways to insult with constantly changing words. By identifying and removing hate speech automatically, you can identify phrases with multiple dimensions, hidden meanings beneath the words, and integrate them into the blacklist. Due to these concerns, identifying hate speech has become more complex, so we prepare a literature review to understand the approaches and results by researchers to help choose the most appropriate datasets for research and an in-depth, comprehensive, and organized understanding of automatic hate speech detection NLP and ML, DL scientists' researchers who an introduction to the field of hate speech detection.

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

I. Introduction

There is no doubt that speech is a very important part of life; every animal on the planet uses it to communicate. Communication is made possible through speech between animals that can communicate one to one and many to many. Speaking to mankind is a way of understanding intentions, emotions, actions, and more. Speech plays a major role in communication, which can cause good or bad emotions in people and can lead to better or hateful statements. With the growing diversity in the world, more people are using social media and sharing information, which has provided many benefits to humanity. However, there are some challenges associated with spreading hate speech and messages. Speech is a very important part of life; it has become widely used by animals on all continents. HATE speech detection prevents the propagation and spread of hateful content and crimes of hate speech. Language plays an important role in communicating, and it can create an emotion in the listener, which can lead to more or less effective communication. With the growing world population and the diversity of people, social media and sharing information have greatly benefitted humanity. On the other hand, spreading hate speech presents some challenges. Preventing hate speech and crimes can be accomplished by detecting hate speech. A post or any content that contains hate speech can be considered hate speech. Whether it's language, image, video, or audio, it can be anything.

II. HATE Speech Detection

Hate Speech is a speech that attacks targets an individual person or a community group on the basis of attributes parameters such as religion culture, religion, age, ethnic origin national origin nationality, color, gender, disability, community, and group based on race, religion, age, status, sexual, orientation, culture, and 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 the internet and social media for blogging, writing articles and sharing media files and content, reacting to the posts, and sharing their opinions [3].

The definition of hate speech is a 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 [4].

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 interests, 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 faces 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.

Survey which was conducted by an organization called US Anti-Defamation League, survey of more than a thousand Americans in this time period in 2018, about 10 days in 2018. Here are some very interesting facts that they found out right after this survey. Who have experienced harassment online around 53 percent of them have actually experienced some sort of harassment, 41 percent of them have experienced name-calling but possible embarrassment, physical traits, even sexual harassment online [5].

Protected class, gender, physical appearance, political views, ethnicity, religion, racist behaviour, insecure people, sensitive people, physical here you are, obese looking, even beautiful people and most of the case hate speech going on social networks. 70 percent increase in hate speech among teens and kids online, toxicity levels in gaming community has been increasing, protecting kids also very important.

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 method used as deep learning methods for hate speech detection.

The first method followed traditional, very simple, deep learning methods like CNN's on LSD and fast paced kind of classifiers to do hate speech detection. And to apply a CNN's or Elysium's and prospects, some word embedding or words being represented using some victory presentations were used. While evolving through this path, it was learnt to rather than classifying it to incoherent, just to learn to get the features out of those the neural mechanism and then learn the gradient statistics to trees learned about them.

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, and Twitter, 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].

The method followed traditional, very simple, deep learning methods like CNN's on LSD and fast paced kind of classifiers to do hate speech detection. And to apply a CNN's or Elysium's and prospects, some word embedding’s or words being represented using some victory presentations were used. While evolving through this path, it was learnt to rather than classifying it to incoherent, just to learn to get the features out of those the neural mechanism and then learn the gradient statistics to trees learned about them.

This work was done on top of dataset, which talks about racism and sexism, and it was recognized that the approach, which uses LSD, comes along with random and buildings, Incredible state decision trees shows 93 percent effort, which was good compared to the state-of-the-art methods, which are using traditional machine learning baselines in that sense.

Training models can be cross validation method to evaluate the models by using text preprocessing and Cross-validation. Preprocessing which consider to remove certain tokens type like space hashtags special tags and Cross validation uses to avoid overlap and class distribution training data to the model and labeled high probability instance.

Deep Learning Methods approached build for text classification in general, Long Short Term Memory (LSTM) layers (CNN-LSTM) and BERT transformer, RoERTa also a famous transfer which extractor for machine learning methods and which are member of transformer family [8].

V. EVALUATE A HATE-SPEECH DETECTION SYSTEM AND ARCHITECTURES

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.

Deep learning for a Language Processing. Here we understanding why is hate speech detection important and hate speech dataset reference on different kinds of hate speech datasets that are available across different kinds and different forms of hate speech, a feature-based approaches, natural language processing features that have been used for hate speech detection and deep learning methods which have been specifically adapted or proposed for hate speech detection tasks like, Hate speech fiction has traditionally focused on basically using the text part of the post for social media, post of a comment or an article and recently leveraging both the text as well as the image part one like liberal hate speech prediction, that is why we talk about multimodal hate speech detection as well analysts of hate speech prediction results, what are the what is the sort of interpretability which is supported by these hate speech prediction methods to talk and hate speech dataset, different kinds of hate speech datasets that are available, challenges and limitations, existing mechanisms have some limitations.

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 [9].

It was found that the random embedding work better than love buildings, and while investigating the reason for this, it was identified that the glove and bindings don't take care of the hate speech around the particular words which are being used intended.

The outcome of it was that the Globe ratings initialization, doesn't work that well. There by to start off with random ratings and in the process, to learn some sort of hate speech-oriented embedding's, apart from CNN Celestials, many researchers have also experimented with other kind of traditional machine learning traditional deep learning methods like typical CNN's analyst teams on other datasets.

Considering an example on Apple's analyzer, which is basically about analyzing abuse on Gab, right anti-Semitic and making some related abuse, setting it up as a multiclass classification task. So just learning kind of a setup, where the three tasks were to figure out if there is abuse or not and if there is abuse, what is the severity of abuse? Again, three different classes out there. And then to know, if there's abuse or the target of abuse. So, whether it is individual second percentage or third person or a group against which some abuse is being done, they so much as learning has also been played out across other kinds of data sets like using less DMs are also using CNN Ottoman architecture.

In one of the use case where learning was done using LSD, we pass the sentence through an LSD, and then the output actually goes to not just a typical hate speech classifier loss, but also an auxiliary language module. Now, oftentimes, these hateful datasets will have multi label kind of a setup because hate speech classes are many times overlapping in nature.

Considering one dataset, where the labels are around spitting, groping and commenting, there is a whole bunch of overlap across those labels, therefore it was considerate to start off with character embedding's. For this CNN's to come up with bold embedding's sites that each sentence essentially has a fixed size representation. And on top of that, they are each word basically says representation, then on top of that, they have a lesbian, bisexual audience running it and then the output of the biotech slot in an essentially connected to the final upwards of backslid, which basically, predicts one of those three categories, or multiple of those three categories, like a multilevel setup. So, these are all basic, deep learning architectures, whether it is difficulty in and out of steam.

Beyond this, there are also spatial architecture designs to handle hate speech specifically, which talks about text-based hate speech detection using deep learning mechanisms. Skip to CNN's skip to Engrams and Skip CNN's. The idea is that the best practice can be used on text data as well, where we first represent the word using word embedding's, and that it's all for every sentence [10].

We get a matrix on which we could use filters which extend to the victor of the embedding, but then the depth or rather the height could be variable later. For example, a precise filter will have type of pool or try to do convolution or three words at a time. Typically, people have filters which are contiguous in nature, but then it could also have a gap, essentially for an instance the filter of size or with the two of those ignored, so it is more like saying that, Hey, I'll take the first word, take the photo, and then I'll sort of not care about the embedding's for the two words, but just to convolutional the first in the forward and come up with the output. And then we could basically do convolution with multiple such characters and then combine sort of pooling across all those speakers.

This concept is very much like the Graham's kind of concept, Captain Graham's kind of feature that people have traditionally used in traditional entropy mechanisms. For instance, the sentence could be Muslim refugees are troublemakers or something of that kind, Muslim refugees, the troublemakers are still a reasonable feature and to say non-Muslim

troublemakers are also a reasonable feature in that sense. So, it makes sense, and it is exactly the spirit behind the design of the window kind of filters and applying them for hate speech addiction.

There's also very much related to the concept of aerosol dilated convolutions that is very popular in the computer vision community. So essentially, once we have these kinds of filters, we would apply multiple of them and then do pulling across these filters and then finally come up with the output, which is connected to the outputs of maxilla.

Not with just limited to text, we can leverage a whole bunch of material that exists along with the text, so therefore, in their overall model, they have a true power network. One is a text that is Democratic Dark Tower and in the text part, they basically are using Ottoman with like 128 Ottoman units and they're Mexicans. It has been said that 30 tokens are what it can take. It's good as a traditional text tower, but then in the majority side, they basically have a multilayer perceptron. They have an MLP with like six different dense layers. So, there are like five dense layers of that sizes and then there are six layers, which essentially ensures that the embedding size of the material apart is the same as the building size from the text box 128 activations finally, which are then concatenated and passed to the final output classification live [11].

The kind of features in this metadata input are of three different types in this paper.

- Tweet based features
- User based features
- Network based features

These features basically include features like number of hashtags, mentions of user's emoticons, how many words with uppercase, the amount of audience included tweet sentiments, those emotions, president offensiveness scores and so on.

We are passing on the tweet separately for not only is to extract semantics out of it but can also pass on some features in the mixed-up artifact metadata part. Then there are also user-based features like, for example, number of followers of friends of this user and the network number of posted on the proportion favorite retweet, subscribe lists, age of their account and so on. oftentimes in hate speech, people just create new accounts and just start putting out hate speech based on those new accounts, those are suspicious accounts in the incorporated that other feature the age of like, then the alternate will be happy just, for example, number of followers and friends. That ratio of such measures, like followers divide by friends, end up, in some senses.

By extent to which a user tends to reciprocate the follower connections, he receives power difference between a user and his mentions. The user's position in the network is that a user or data, various kinds of network based statistical measures.
The idea is that the two of the networks is trained in multiple different ways, multiple interesting ways that the two outputs come up with the final classification. But there are multiple ways to train.

The first way of training is train the entire network, training data, essentially train the entire network at once, training it end to end for once.

The second way of training is to train each of Gustavo's separately. Having labelled data and those text separately and then train two different, totally different multilayer networks separately, one of them is auto and the other is an MLP. To train them, separate, we must train them separately and then transfer, learn those weights or take those between weights and use them to initialize the combination.

In the third kind of a method, we don't freeze any of the tower weights, we learn the tower weights also while fine tuning the overall combination. It's the same as the second rule, Combined the network and then the combined network is sort of fine.

The fourth method is called interleaving, the idea is that for this entire competition network, keep two copies of the competition. The tower, there are two towers and copy it, so it will be initiated with a combination of Network B. So, we initialize the two article models in B, the text part of it is that non-credible are frozen and the metadata part of B is frozen. it is going to specialize in learning the weights for the data that B is going to specialize in learning the weights for the text part.
For every batch, pass the Bachelor, one by one, both the networks and b one by one, metro networks it and train the Model B after one batch of training is done.

While training a model year, the next part is frozen, which means we are not going to update the weights of the text box. But we are going to still make use of those which says to come up with the final prediction after a batch of four for the network, but the network is done. copy most of its from network, which were updated due to this batch and then copy them to be able to be better kept frozen, they are actually used for computing the final bridge. That's how, interleaving training is done even batch because there is batch code to network. And finally, we end up with two networks, we could use whichever network as a final access network [11].

This method was experimented with four different hate speech, cyberbullying, offensive hit and then sarcasm dataset. And it was observed across all those four datasets that overall interleaving method gives the best results. The interleaving method gives the best results across these four different methods and clearly, the combined network gives the better results compared to the competitor. It's also observed that in

the measured-up part, we could just use network features only or treat features only or other features only or combine everything. And the observation was combining everything basically gives the best results.

VI. DATASETS FOR HATE-SPEECH DETECTION

The datasets could pertain to social networks because most people put up hate speech comments on social networks, and comments on websites and non-social networks are also studied. Among the oldest datasets, Waseems 2016, there are 67,000 tweets, out of which 30,383 are sexist, racist, and neutral.

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-labelled datasets collected and annotated at different times by different people can be helpful for datasets. HateBase is a corpus of automated content that detects hateful and offensive content, labels it, and categorizes it into three classes, namely hate speech and offensive language.

The majority of the data from Twitter comes from social media platforms. Hate-related characteristic datasets include Hatebase Twitter, WasseemA, WaseemB from Twitter in English, Stormfront from an online forum in English along with TRAC from Facebook, HatEval, Kaggle, and German Twitter from Twitter. The type datasets consist of different type labels and languages [12].

Twitter Datasets which are popular datasets, the dataset to dataset is different, which indicates toxicity or harassment, offensive language or hate speech at different levels, but the important fact is that not every offensive language is hate speech.

Several datasets are collected across multiple social media platforms, such as Facebook and WhatsApp, and labelled across numerous hostility dimensions, such as hate speech and fake news, offensive posts, and defamation posts. In 2020 Andrea, Anti-Semitism Datasets were created by students by using social media labelled 7600 posts and then exploring the datasets to find the data containing anti-Semitic Semitism prospective. We should label these posts in three different categories: biased attitude, bias and discrimination, violence, and genocide, and also the target of the abusive behaviour, whether it is an individual or a group, so that we know if this is hate speech or not and what degree of it or its severity and type of victim.

Researchers are looking into social media datasets such as Instagram bullying and Instagram and wine datasets for 2020. Then bullying becomes a multi-modal dataset and again on cyberbullying, considering the image as to whether or not it is prone to a cyberbullying conversation. The wine datasets are basically datasets one or more that are multidimensional or essential multi-modal datasets from Facebook. Kaggle's toxic comment datasets classification is beyond social networks such as Wikipedia, which contains six types of toxic comment datasets: toxic or hatred, toxic severe, toxic obscene, type kit insults, and identity hate [13].

A Whisper dataset is similar to a Twitter dataset, which consists of large amounts of blacklists and is very useful for researching hate speech. In addition, the Storefront web domain dataset contains victim accounts from websites that talk about sexism and contains multiple tags such as groping, touching and so on.

Various kinds of datasets are available. Before deep learning became popular, the hate speech community worked and leveraged a whole bunch of traditional machine learning classifiers. The first method is rule-based methods which contain words or anagrams, such as insult swear words, action words, personal pronouns, word lists, blacklist sites, hateful terms, phrases for hate speech based on race, disability, and sexual orientation are also available on Wikipedia pages, acronyms and abbreviations and slang and variations of profane words are also available and holds large lists across different languages. This method is designed for a quick and efficient manner.

Machine learning methods and traditional machine learning methods and their features that have been explored are traditional linguistic features that one could expect to write all kinds of natural language processing features like a bag of words, that is practically taken each word and derive frequency-based feature out of it and grammar, word and grounds to correct anagrams as well. The frequency, inverse document frequency of various words part of speech tagging. Use the part of speech dog on the words contained in the sentence as a feature domain, whether the sentences are hated speech or not, right. For example, a number of nouns, number of pronouns, and so on.

Feature-based topic modelling can be applied to features like sentiment, classification frequency of personal pronouns in the first and second person, and presence of emoticons in capital letters. Flex Kincaid grade level is a measure of how well grammar is used in the sentence. People believed that, for a time at least, hate speech is typically authored by people who cannot write very correct grammar and therefore use at least grade level English correctness in English and medical correctness corsages as some sort of distinguishing features. Lastly, there are binary features, binary and count indicators for hashtags and mentions. The number of characters, words, and syllables in each tweet when it comes to hate speech detection, language-based features, or other languages communities,

features are basically around analysing the contrast between different groups by examining the expressions or phrases that are used for other groups [14].

Additionally, users can specify their gender and geographic location, which are popular features. Because languages are subjective and objective, hate speech is sometimes related to more subjective communication and therefore, if the sentence looks more formal, the likelihood of hate speech is lower. Scientifically speaking, anti-Semitic hate speech often refers to money, banking, and the media [15].

Basically, a short summary of the various features that have been used for hate speech detection and the traditional machine learning models that have been used to leverage these features and solve the hate speech detection issue.

Datasets are used to train algorithms to solve various types of problems: 2 label datasets, 3 label datasets, and balanced, imbalanced, and number of counts datasets. Most researchers use imbalanced datasets. By removing the repeated words that a sampling method requires to build balanced datasets, better results can be achieved.

Using data annotation, the datasets will be prepared by sampling the data and training the model using classifiers to detect hate speech. Data from the online system to be compared correlates with meanings and opinions with data annotations need to be prepared beforehand. Datasets are typically built by twitter on a large scale, and hashtags are used to select the data. Unfiltered datasets like hashtags have many issues including ambiguity and stability down the line for this kind of issue, data annotation is crucial along with setting up the benchmarks process. By implementing these policies, we can retrieve data from the start of hashtags and hateful content keywords according to the guidelines of the data annotations. Data content includes all domain-specific content associated with specific attributes, and annotator role includes more or fewer keywords, like language choices, topics, sensitive words, and sensitive attributes. For research on these datasets, online data must be accompanied by privacy policies governed by GDPR [16].

There are several challenges associated with data annotations, including images, videos, content, and text in social media platforms, and some data are tricky since they can disappear after some time. Annotating data to automatic detections has some challenges Multilingualism, multimodality (combining images, videos, and texts), detection in context and platform-specific improve the classification results Merging the data with different platforms improves the classification results [16].

Augmenting vision with data is much easier and more popular than ever. Although data augmentation is challenging in LP, NPIs text is discrete in nature, while images are continuous. This makes augmentation for text difficult in general.
As far as hate speech detection for augmentation is concerned, there are a variety of methods that have been proposed. Generalized, it could be used to increase the size of hate speech due to hate speech, and I would say it's usually smaller.

Second, it affects the number of instances for the hate class. The hate class is typically small. In that case, does it lead to an imbalance between minorities, the hate classes, and the non-hate classes, As a result, the simplest first method is simple oversampling. The idea is to sample all hate speech incidents that are of minority status. Copy what minority class data points appear more than once.

The second kind of method is the ideal method, which is a combination of four different methods.

- The first one is a synonym replacement from Fortnite. In the given hate speech syntax, take a word from the document or from this post, and then replace it with a synonym from Fortnite.

- The second one is essentially a random substitution of a synonym, just insert a synonym in that sentence.

- The third step involves selecting two random words from a sentence, then just sweeping them aside.

- Fourth, we can delete a random word in the sentence, hoping that we won't end up deleting the main target hate speech word.

It is also possible to use a word-based method instead of ediyor, provided that while replacing the word with a random synonym from word, it must also be ensured that the replacement words have the same sense as the original word.

We, therefore, perform word sense disambiguation on the original hate speech sentence in order to understand the sense of each hate speech word, and then find a replacement from the word note that matches that sense.

A fourth method is a paraphrase database-based method. PPD is a very popular paraphrase database, which contains a phrase and then equivalent phrases which are semantically similar in meaning. The idea is to control the grammatical context within which those phrases appear so that we will actually replace the phrases. In order to ensure that the part of speech tag of the original word and the deepest word remains the same in a single-word replacement, we are interested in the part of speech.

Because a replacement is a multiple-word paraphrase, it is important to consider the syntactic category in the PDB training corpus after the original phrase, and that is that the

next method that is embedding will never replace anything. By taking a hate speech sentence, we take a random word, and then, by using cosine similarity, we find the top end nearest embedding neighbours, and we replace the original word with these neighbours. Replace one of the words in the sentence with or stop the neighbours to lead us to 10 other hate speech instances. In the present, embedding's can be used in a variety of ways, and this paper could be used for kinds of embedding's between buildings, beddings, and pieces of sentences. There are two embedding's that have used the next method, which is a majority class sentence addition.

One last example is a digitally-based conditional generation, here the idea is to use a pretend deputy to one hundred and ten million perimeter models, fine-tune those models on hate speech documents in the current corpus, and then for each of those hate speech documents, essentially generate and subtract one Novell document using the LGBTQ model.

Since no model is required for the first three methods, no parameters are needed. Then there are these methods that are nearly dependent on embedding test parameters. Additionally, there are methods that require a lot of parameters.

Several large models are available for generating instances for data augmentation, the four models are character-based representation, curriculum-based logistic regression, word-based logistic regression, and CNN purchased models; the parameters in these models differ, as do their levels of accuracy. The table shows the precision recall and if one that can be obtained based on a model like this, in the case of hate speech in 2020, when it was in the case of 20 excitations. It appears that some models show good precision when they are augmented with some data, while when augmented with the other method they show good recall. Using deep learning with augmentation methods to improve hate speech detection was the first step.

VII. CHALLENGES AND CONCLUSION

The hate speech domain has seen a lot of work in the past, but annotation still remains a major challenge.

In general, humans have a low agreement on the classification of hate speech. An opinion that offends anything offensive is considered hate speech. Of course, many people end up using offensive language in their speech. However, ever, if they are doing so in a jocular manner as part of a joke, then it doesn't mean they are being hateful. This is because there might not be any result, offensive words offensive keywords plus malicious intention constitute hate speech.

Moreover, hate speech may not be confined to individuals but can be directed to a group, which means it may also be culturally sensitive. Therefore, their task requires expertise about the culture and social structure in order to understand it.

Another challenge is the annotation tools. Even if it is just text-based hate speech, there are many tools that can indicate whether it is speech or not, or which phrase in this input document indicates that this is hate speech. Labelling in the image domain is difficult, however. For example, identifying which part of the image corresponds to hate speech, or even labelling videos and stating that this part of the video is hateful, is difficult to achieve.

There is this evolution of social phenomena and language, which makes it hard to say whether something is hateful or not. In five years or ten years from now, something that is considered hateful today might not be considered hateful. Language evolved in Wales, and especially in the language of young people. A lot of social network phenomena are centred on young people. Young people's language evolves quickly, their slang changes, and something that is an insult today may not be considered so after 5-10 years. Hence, the models need to be updated and the annotations need to be updated.

People used to believe that abusive language is used by people who don't know correct English grammar. Therefore, such language may not be very fluent or grammatically correct. Those clues to determine if it is hate speech or not. Nevertheless, in recent years people have found that there are in fact published papers that prove anti-Semitic theories and so forth. So there are scientific literature, scientifically sound, very principled, grammatically correct sentences, which are hate speech. As long as the grammar is correct, the diction is simpler than keyword spotting.

Since character adversarial attacks are common, people can come up with such obligations to evade hate speech moderation mechanisms, which also adds another challenge for hate speech detection mechanisms.

Another challenge is interpretability. The idea is that people expect systems that automatically detect when a person's speech might require a manual appeal.

Based on hate speech prediction, Facebook labels a bunch and demands a bunch of content. Now, it is very difficult to solve this interpretability challenge by allowing a manual appeal process for each. It is evident that the problem can be solved to some extent with the methods, but some of the deep learning-based hate speech mechanisms are clearly circumventions.

There are design mechanisms that are sometimes multimodal and sometimes just look at a text and judge whether it is hate speech or not. People have been posting content as images that contain the text rather than the text itself. Those are also limitations of the current mechanisms, which will need to be addressed by more advanced deep learning models in the future.

ACKNOWLEDGMENT

This work has been supported by the Slovak Research and Development Agency projects APVV-SK-TW-21-0002 \& APVV-15-0517, by Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under the research projects VEGA 1/0753/20 \& VEGA 2/0165/21 and by Cultural and Educational Grant Agency of the Slovak Republic grant Nos. KEGA 009TUKE-4/2019 \& KEGA 048TUKE-4/2022, funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic.

REFERENCES

[1] Nanlir sallau mullah 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] Claudia Zaghi "Automatic detection of hate speech in social media", University of Malta 2018

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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.

[15] 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

[16] 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.