Mapping Study of Hate speech detection

Dyass Khalid,Syed Mohammad Arham Noman

20100004@lums.edu.pk , 20100059@lums.edu.pk

April 2, 2020

Abstract

BACKGROUND: Hate speech detection has been a growing area of interest as the user base of social media websites has grown. Human content moderators can only do so much, hence automatic detection has become crucial. This paper explores the topic with a particular focus on Roman Urdu texts as these have not been explored.

CONTEXT: Recent advances in deep learning (DL), soft computing (SC) has made tremendous progress in the last three years leading to usage in detection of hate speech systems and classification systems on the web. These advances have led researchers to focus on hybrid techniques for the detection and classification of hate speech present on the social media platforms in bulk amounts.

OBJECTIVE: We aim to conduct a systematic mapping in the area of detecting hate speech on social media platforms particularly Twitter. The aim is to identify, analyze and classify the existing literature to provide an overview of the area.

METHODOLOGY: We followed the guidelines set by the research community in systematic mapping to develop a systematic protocol to identify and review the existing literature. We formulate sets of research questions for the classification and detection of hate-speech techniques.

RESULTS: We selected 6 papers out of 100 papers based on our selection criterial up to March 2020. We analyze the classification approaches, the type of dataset used, the type of classifiers, themes in which hate speech detection is being applied and the metrics being used in language modelling. We also report the quality of the work using the established assessment criteria.

CONCLUSIONS: 1. We present an overview of the area by answering several research questions. With the increase in the amount of data and with the development of new deep learning systems that can learn deep representation of system, the trend is rapidly shifting towards deep learning-based approaches.

2. The important venues are particularly ACM conferences and journals where several techniques are being employed particularly feature engineering using genetic programming techniques. Tuning of ML hyperparameters and weight update strategy using genetic algorithm in small neural networks.

3. The trend is shifting towards auto ML for feature extraction and deep learning very rapidly.

Keywords: Systematic Mapping Study, Hybrid Models, Machine Learning, Deep Learning,Genetic Algorithms,Fuzzy Rules Systems

1. Introduction

Due to the recent advancement in the field of ML particularly due to advent of deep learning networks, NLP has seen a humongous increase in research papers and new methods coming every month. These advances have led to the adoption of NLP in automated speech recognition, question-answering systems, and more sophisticated nlp systems such as Siri and Alexa. However, recent applications such as exact voice copying with the voice generating hate-speech without even the people whose voice is being copied knowing is a big concern. To alleviate such accidents, there is a growing interest in hate-speech classification.

However, classifying hate speech accurately poses new challenges for the NLP community. First, due to multilingual nature of social media platforms, it become very difficult for the moderators to identify hate-speech. Furthermore, hate-speech for one community is not hate-speech for the other and vice-versa.

In this paper, we conduct a Systematic Mapping by including the papers published particularly in SM for hate-speech classification. The following are the main contributions of this study:

- Different themes in which classification schemes are applied.
- Identification of frequently used techniques used in hate-speech classification.
- Identification of different metrics used in hate-speech evaluation.

The rest of the paper is organized as follows. Section 2 present the summary of the related work. Section 3 discussed the methodology used and search process to perform this study. Section 4 provides the results and discussion. Section 5 discusses threats to validity. To end, section 6 concludes the study.

2. Related Work

In this section, we discussed several studies related to our work. In general, one papers is selected which is a survey. Surveys are typically not conducted in systematic and unbiased fashion, and thus are considered by some researchers to have little scientific value[**Bar09**][1].

Also, they often suffer from selection bias and their results are not repeatable due to lack of an explicitly defined protocol and search strategy[2]. Although, systematic reviews follow a well-defined protocol and search strategy which reduces the selection bias, we are unable to find any systematic study to the best of our knowledge in the field hate-speech detection.

[3]defines what hate-speech is for different social media networks. The definitions are shown in the table.

Source	Definition
Code of conduct between EU and companies	"All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic" .
ILGA	"Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups."
Facebook	"Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)."
YouTube	"Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and YouTube sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity."
Twitter	"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease."

Table 1: . Hate Speech Definitions by different authorities

[3]also defines the type of hate speech analyzed in the papers as shown in the following table:

Hate Type	Frequency
General hate speech	26
Racism	18
Sexism	6
Religion	4
Anti-Semitism	1
Nationality	1
Other	1
Physical/Mental Handicap	1
Politics	1
Sectarianism	1
Socio-economic status	1

Table 2: . Type of Hate Speech Analyzed in the papers from "Computer Science and Engineering"

[3]also shows the type of algorithms used when they performed the survey. As we will see in the upcoming sections, the focused has now shifted towards deep learning methods.

Algorithms	Frequencies
SVM	10
Random Forests	5
Decision Trees	4
Logistic Regression	4
Naïve Bayes	3
Deep Learning	1
DNN	1
Ensemble	1
GDBT	1
LSTM	1
Non-supervised	1
One-class classifiers	1
Skip-bigram model	1

Table 3: . Algorithms Used in the Papers from “Computer Science

[4] also has a table for the type of features and algorithms used in the techniques.

3. **Methodology**

The protocol we will be using is shown in the following flowchart:

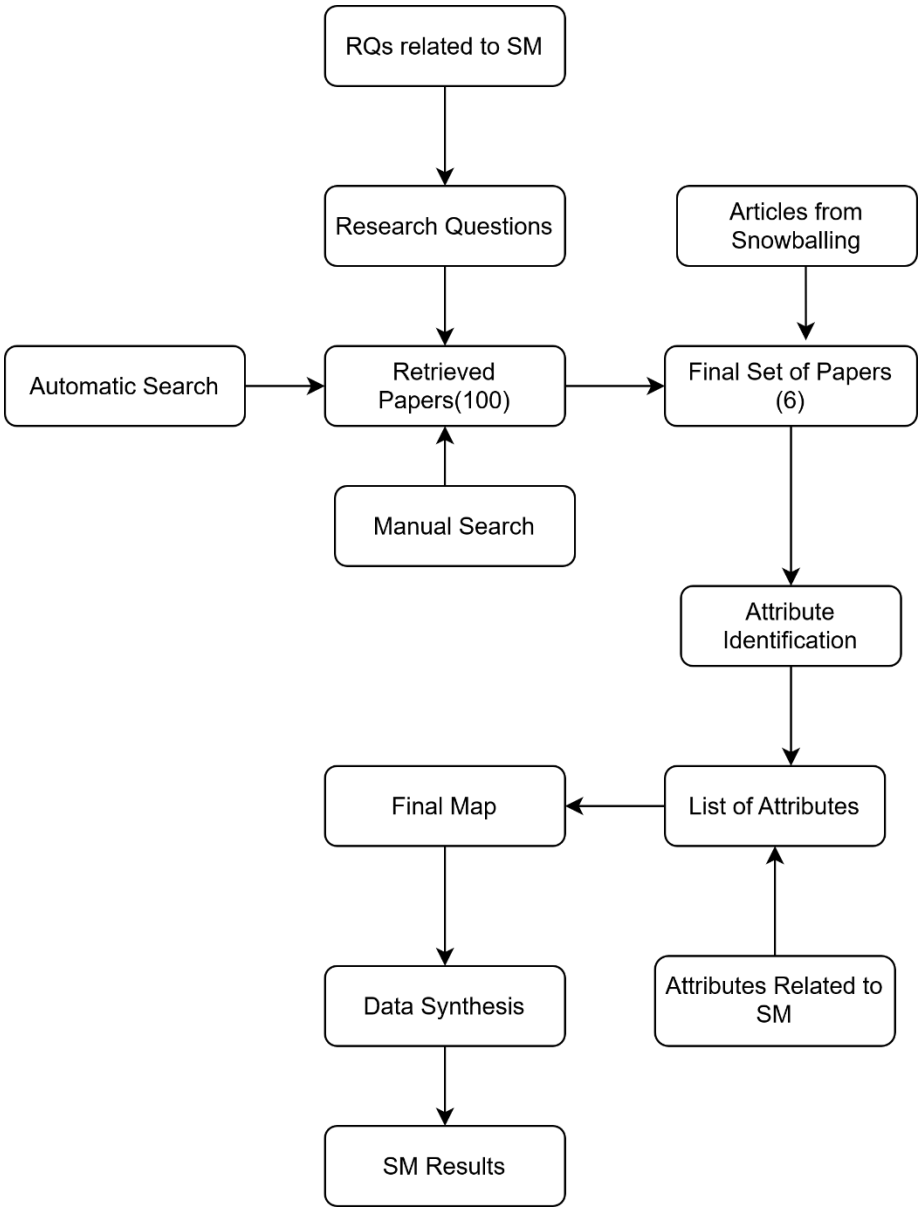

Figure 1: . Protocol Used in the Mapping Study

3.1. Goals and Research questions

The goal of this study is to identify and analyze the work published in the area of hate-speech classification using hybrid systems which can include combination of neural, fuzzy, genetic, and deep learning systems. Based on our goal, we formulate research questions in a single category i.e. systematic mapping.

3.1.1. Systematic Mapping

This category is focused on identifying the type of research in the area of hate-speech identification and classification. The research questions are the following:

RQ1: How popular are hybrid models?

This question will help us identify the ongoing trend used for the purpose of hate speech identification.

RQ2: What are the most important venues, research groups, year of publications of hybridization models?

This question will help us identify the applications in which hybridization models are used and how can we extend those models.

RQ3: What are the most common methods that have been implemented in the last five years in hate-speech identification?

This question will help us in identifying where the research trend is currently shifting towards and what are the hot areas in hate-speech to research on.

RQ4: What are the gaps and review of the papers we have included in the study?

This question will help us to look around the methodology being used and why a certain method is used rather than other method and what are limitations of following this approach and why this approach should be used and why this should not be used here.

RQ5: What is an innovative approach that could be implemented in hate-speech detection?

This question will help us to come up with new and novel ideas and maybe combine ideas of two papers and implement them.

3.2. Search Process

We conducted manual search for identification of primary papers.

3.2.1. Manual Search

Manual search is conducted by manually searching the venues and journals mentioned in the requirements process. Based on that we identified 21 related papers after of which their venues are listed in the following table

Table 4: Venues of selected Papers

Source	**Acronym**	**Type**	**Year**
Proceedings of the second workshop on Language in social media	LSM	Workshop	2012
World Wide Web	WWW	Conference	2015,2017,2019
Proceedings of the 2nd International Conference on compute and data analysis	ICCDA	Conference	2018
Advances in Social Networks Analysis and Mining	ASONAM	Conference	2018

continued on next page

continued from previous page

Hypertext and Social Media	HT	Conference	
ACM Magazine for Students	XRDS	Mazagine	2017
ACM Transactions on IT	ACM-IT	Article	2020
ACM Computing Surveys	ACM-CS	Article	2018
Security and Privacy Analytics	IWSPA	Conference	2020
Technological Ecosystems for enhancing multiculturality	TEEM	Conference	2019
ACM SIGR conference on Research and Development in Information Retrieval	SIGR	Conference	2019
Data and Application Security and Privacy	CODASPY	Conference	2020
Artificial Intelligence and Security	AISec	Workshop	2018
Information and Knowledge Management	CIKM	Conference	2019
Web Science	WebSci	Conference	2017,2019
ACM Transactions on the Web	ACM-W	Article	2019

3.3. Study Selection

We initially obtained 21 papers through our search. We define the following inclusion and exclusion criteria for thorough selection of related papers.

Inclusion Criteria

- Papers that proposed technique, method, tool, framework or an experiment on testing of hate speech classification
- Papers which are available freely on the Internet
- Papers available in full text
- Papers focused on hybrid techniques
- Papers that uses novel datasets

Exclusion Criteria

- Off topic papers
- PhD dissertations
- Papers not available in full text
- Papers not written in English

After the application of inclusion and exclusion criteria, only 6 papers were left. The search process is illustrated in the following figure:

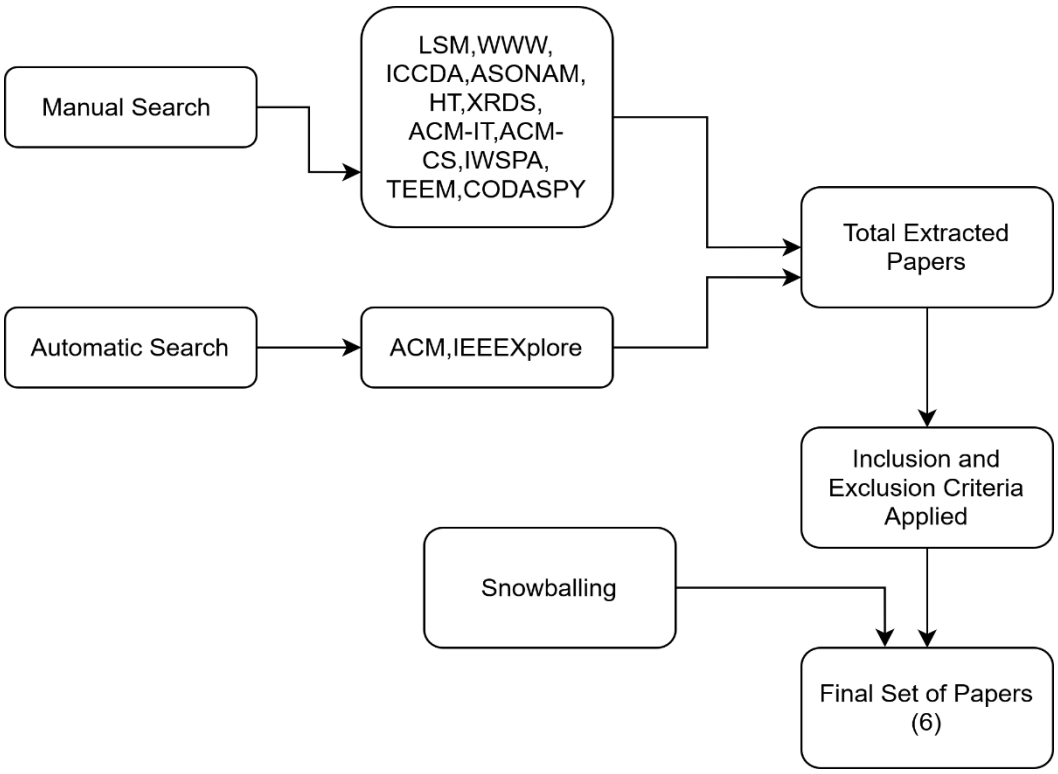

Figure 2: . An illustration of the search process

3.14. Data Extraction Strategy

The final set of 6 papers are taken for detailed analysis to answer our research questions. First, we identified attributes corresponding to each research question which were to be extracted from the included papers. Table 5 provides the detailed listing of these attributes. The first column shows the unique identifiers for attributes listed in the second column. Third column shows the corresponding RQ whereas the last column shows the possible values extracted from papers for each attribute.

ID	Attribute	Corresponding RQ	Possible Values
ID1	Hybrid Model	RQ1	• Neuro-Genetic • Neuro-Fuzzy • Genetic-Fuzzy • Neuro-Genetic-Fuzzy • Deep Neural Architectures
ID2	Year	RQ2	• Year of Publication
ID3	Venues	RQ2	• Conference • Workshop
ID4	Method	RQ3	• Supervised • Unsupervised • Semi-Supervised • Structured
ID5	Gaps	RQ4	• See section 4
ID6	Novel Approach	RQ5	• See section 4

Table 5: List of data attributes with the corresponding RQs and possible values

4. Results and Discussion

In this section, we discuss the results of the study and presents the synthesis of data extracted in the previous section to answer the research questions in detail.

4.1. RQ1: How popular are hybrid models?

The popularity of hybrid models in the domain of hate-speech classification is growing rapidly particularly in the area of deep learning where hybrid models are not only being used for feature extraction but also used

for classification purposes. However, the trend has started to shift towards method of unsupervised-deep learning.

RQ2: What are the most important venues, research group, year of publications of hybridization models?

The important venues for the publication of hybrid models along with year of publications are shown in Table 4.

RQ3: What are the most common methods that have been implemented in the last 5 years in the area of hate-speech classification?

The most common methods in the papers we have got after applying the inclusion and exclusion criteria are shown in the following table

Table 6: Methods used by researchers

ID	**Methods**
ID1	Linear Regression
ID2	Logistic Regression
ID3	Support Vector Machines (Different kernels)
ID4	Naïve Bayes
ID5	KNN
ID6	Feed-Forward Neural Network
ID7	Convolution Neural Network with different type of pooling layers
ID8	Recurrent Neural Network
ID9	LSTM
ID10	BI-LSTM
ID11	GRU
ID12	BI-GRU
ID13	Decision Trees
ID14	Gradient Boosted Decision Trees
ID15	Fast Text
ID16	Combination of ID1-ID15
ID17	Ensemble Learning
ID18	Non-supervised Method

This is also illustrated in the following diagram:

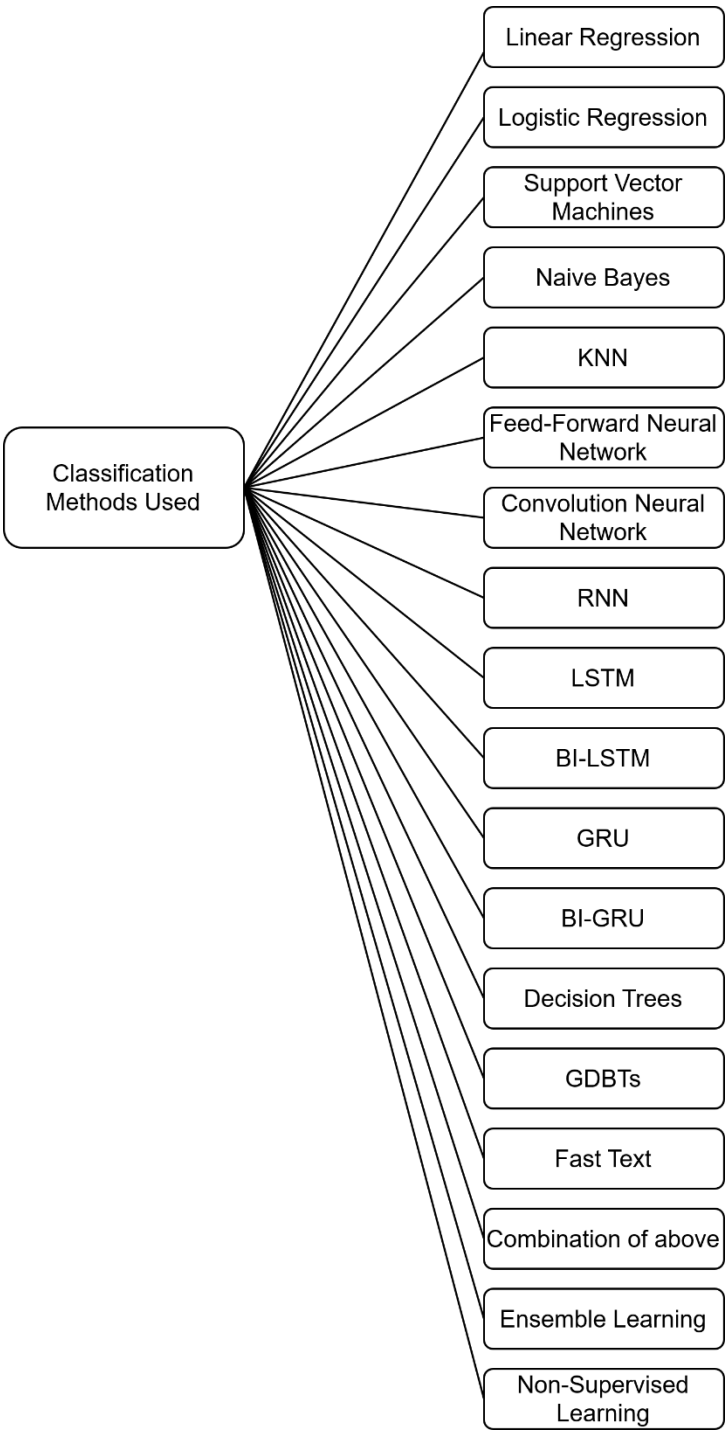

Figure 3: . Classification Methods used

RQ4: **Identify the gaps and write a critical review on the papers.**

Review of Paper [4]:

1) Language modelling for stereotypes is built separately for each stereotype.

2) Inter-labeler agreement metric is Fleiss Kappa.

3) Template based feature extraction strategy is used which will not be able to extract the intra dependency of sentences along with POS tagging with a window size.

4) SVM light with a linear kernel is used as a classifier foe each type of feature template with total of six classifiers.

Short Comings of Paper [4]:

1) Unigram model is only used which cannot capture the dependencies.

2) Feature extraction criteria can be made more robust by using ridge and lasso regression strategies.

3) Only a single classifier is being used. The bias and variance tradeoff are never mentioned in the paper.

4) Ensemble classifiers can be used to improve the scores because in the worst case they will perform better than the SVM classifier being used.

5) For feature extraction, we can use encoder-decoder network which will significantly reduce the size of the feature vector.

6) TF-IDF and word embeddings can also be used as a feature vector.

Review of Paper [5]

1) Paragraph to vector embeddings using continuous bag of world model is being obtained.

2) These embeddings are then trained using logistic regression classifiers using 5-fold cross validation strategy.

3) Metrics used are precision,recall.

Short Comings of Paper [5]:

1) Data distribution is not mentioned. Neither is the source of the data

2) Glove or Elmo embeddings can also be used instead of paragraph to vector embeddings.

3) Bleu score can be used as a metric for the evaluation of model.

4) Deep learning networks can be used to obtain better metric scores.

5) Macro precision and recall should be used since we do not know if the classes used are imbalance or not.

Review of Paper [6]:

1) Base line methods for feature extraction are character level n-gram, TF-IDF and Glove embeddings

2) Base line classifiers are Logistic Regression, Balanced SVM, Gradient Boosted Decision Trees.

3) For testing purposes these classifiers are CNN's, LSTM's and Fast text classifier used with random embeddings or Glove embeddings.

4) Further, they combine the classifiers mentioned in (b) and (c) to outperform the models in (b) and (c).

Short Comings of Paper [6]:

1) Same embeddings are used for a word irrespective of its context. We can use Elmo embedding which learn separate word level embeddings for the words depending on the context.
2) We can also use AUC score as a metric.
3) We can also use pre-trained encoders to reduce the time for feature extraction.
4) We can also use bidirectional LSTM to preserve information about both past and future.

Review of Paper [7]:

1) Following features are used alone or their combination:
    a) Word Embeddings
    b) Emoji Embeddings
    c) N-grams
    d) Social Network Specific Features
    e) Emotion Lexica
2) Fast text embeddings are also used.
3) Standard pre-processing steps are used
4) Recurrent Neural Network is used with a standard neural network.

Short Comings of Paper [7]:

1) Ensemble learning can be used to further improve the metrics

Review of Paper [8]:

1) LSTM is used as neural network
2) Salience is used to identify which words are responsible for hate-speech within a tweet.

Short Comings of Paper [8]:

1) Bi-directional GRU or bi-directional LSTMs can be used to better capture information.
2) Embeddings from Fast-text or Elmo can also be used.
3) Ensemble learning can be further used to improve the accuracy of the model.

Review of Paper [9]:

1) Following baseline models are used:
    a) SVM and RF combined with BoW
    b) Logistic Regression with Paragraph2Vec
    c) Vowpal Wabbit regression model with Paragraph 2 Vector

d) Gradient Boosted Decision Trees in combination with LSTM model.
e) CNN combination with Word2Vec embedding.
f) Modified CNN with a GRU layer.
g) LSTM classifier with random embedding

2) Precision, Recall, F-score and AUC are used as metric.
3) Othering theory is what inspired the authors
4) For features, a semantic algorithm and standard nlp features are used.

Short Comings of Paper [9]:

1) Pre-trained embeddings can be used for example fast-text or Elmo

RQ5: Based on your experience suggest an innovative approach that could be implemented in your application domain

Proposed Approach:

Features:

- Create features such as unigram, bigram, trigram
- Create POS features
- Create tokens for users and links so that their dependencies are captured
- Create linguistic based features based on othering theory

Feature Selection:

- Use ridge and lasso regression for feature selection
- Use encoder for deep feature selection
- Use pre-trained embeddings for deep feature selection

Classifier Selection:

- Genetic approach to select a combination of classifiers
- Train deep learning architectures mentioned in Table 6
- Create an ensemble and use different features for the relevant metrics.

Papers further excluded out of 21 selected:

1) **Detecting Hate Speech within the Terrorist Argument: A Greek Case.[10]**
2) **Identifying and categorizing profane words in hate speech.[11]**
3) **A measurement study of hate speech in social media.[12]**
4) **Identifying hate speech in Social Media.[13]**
5) **Fuzzy Multi-Task learning for hate speech type identification.[14]**
6) **A survey on Automatic Detection of Hate Speech in Text.[15]**

7) **Hate speech against Central American immigrants in Mexico: Analysis of xenophobia and racism in politicians, media and citizens [16]**

8) **Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations.[17]**

9) **Spread and reception of fake news promoting hate speech against migrants and refugees in social media Research Plan for the Doctoral Programme Education in the Knowledge Society.[18]**

10) **Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation[19]**

11) **On the Impact of Word Representation in Hate Speech and Offensive Language Detection and Explanation[20]**

12) **All You Need is "Love" : Evading Hate Speech Detection [21]**

13) **Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification.[22]**

14) **Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere.[23]**

15) **The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection[24]**

Papers selected for the mapping study after applying all the protocols:

1) **Detecting hate speech on the world wide web.[4]**

2) **Hate speech detection with comment embeddings.[5]**

3) **Deep learning for hate speech detection in tweets.[6]**

4) **A multilingual evaluation for online hate speech detection.[7]**

5) **Towards Automatic Detection and Explanation of Hate Speech and Offensive Language.[8]**

6) **"The Enemy Among Us" : Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings.[9]**

References

[1] Ludwik Kuzniarz Kai Petersen Sairam Vakkalanka. "Guidelines for conducting systematic mapping". In: (2015), pp. 1–18.

[2] Wohlin Claes. "Guidelines for snowballing in systematic literature studies and a replication in software". In: *ACM* (2014), p. 8.

[3] Paula Fortuna Sérgio Nunes. "A Survey on Automatic Detection of Hate Speech in Text." In: *ACM Comput. Surv.* (2018), p. 30.

[4] William Warner and Julia Hirschberg. "Detecting Hate Speech on the World Wide Web". In: *Proceedings of the Second Workshop on Language in Social Media*. LSM '12. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 19–26.

[5] Nemanja Djuric et al. "Hate Speech Detection with Comment Embeddings". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: Association for Computing Machinery, 2015, pp. 29–30.

[6] Pinkesh Badjatiya et al. “Deep Learning for Hate Speech Detection in Tweets”. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW ’17 Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[7] Michele Corazza et al. “A Multilingual Evaluation for Online Hate Speech Detection”. In: *ACM Trans. Internet Technol.* 20.2 (Mar. 2020).

[8] Wyatt Dorris et al. “Towards Automatic Detection and Explanation of Hate Speech and Offensive Language”. In: *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*. IWSPA ’20. New Orleans, LA, USA: Association for Computing Machinery, 2020, pp. 23–29.

[9] Wafa Alorainy et al. ““The Enemy Among Us”: Detecting Cyber Hate Speech with Threats-Based Othering Language Embeddings”. In: *ACM Trans. Web* 13.3 (July 2019).

[10] Ioanna K. Lekea and Panagiotis Karampelas. “Detecting Hate Speech within the Terrorist Argument: A Greek Case”. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’18. Barcelona, Spain: IEEE Press, 2018, pp. 1084–1091.

[11] Phoey Lee Teh, Chi-Bin Cheng, and Weng Mun Chee. “Identifying and Categorising Profane Words in Hate Speech”. In: *Proceedings of the 2nd International Conference on Compute and Data Analysis*. ICCDA 2018. DeKalb, IL, USA: Association for Computing Machinery, 2018, pp. 65–69.

[12] Mainack Mondal, Leandro Araújo Silva, and Fabrıcio Benevenuto. “A Measurement Study of Hate Speech in Social Media”. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT ’17. Prague, Czech Republic: Association for Computing Machinery, 2017, pp. 85–94.

[13] Alexandra Schofield and Thomas Davidson. “Identifying Hate Speech in Social Media”. In: *XRDS* 24.2 (Dec. 2017), pp. 56–59.

[14] Han Liu et al. “Fuzzy Multi-Task Learning for Hate Speech Type Identification”. In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 3006–3012.

[15] Paula Fortuna and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text”. In: *ACM Comput. Surv.* 51.4 (July 2018).

[16] Maximiliano Frıas-Vázquez and Carlos Arcila. “Hate Speech against Central American Immigrants in Mexico: Analysis of Xenophobia and Racism in Politicians, Media and Citizens”. In: *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM’19. León, Spain: Association for Computing Machinery, 2019, pp. 956–960.

[17] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. “Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-Based Generalizations”. In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 49–59.

[18] David Blanco-Herrero and Carlos Arcila Calderón. “Spread and Reception of Fake News Promoting Hate Speech against Migrants and Refugees in Social Media: Research Plan for the Doctoral Programme Education in the Knowledge Society”. In: *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM’19. León, Spain: Association for Computing Machinery, 2019, pp. 949–955.

[19] Aymé Arango, Jorge Pérez, and Barbara Poblete. “Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR’19. Paris, France: Association for Computing Machinery, 2019, pp. 45–54.

[20] Ruijia (Roger) Hu et al. “On the Impact of Word Representation in Hate Speech and Offensive Language Detection and Explanation”. In: *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. CODASPY ’20. New Orleans, LA, USA: Association for Computing Machinery, 2020, pp. 171–173.

[21] Tommi Gröndahl et al. “All You Need is “Love”: Evading Hate Speech Detection”. In: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. AISec ’18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 2–12.

[22] Georgios Rizos, Konstantin Hemker, and Björn Schuller. "Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Beijing, China: Association for Computing Machinery, 2019, pp. 991–1000.

[23] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. "Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere". In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '18. Barcelona, Spain: IEEE Press, 2018, pp. 69–76.

[24] Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. "The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection". In: *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17. Troy, New York, USA: Association for Computing Machinery, 2017, pp. 405–406.