Related Works:

Group: Aditya Mohan Gupta Ashkan Vedadi Gargary

Centralized Hate-Speech Recognition:

- *Papers [abusive contents]:*

1. Paper Name: **Abusive Language Detection in Online User Content [most important]**

Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. InProceedings of the 25th international conference on world wide web 2016 Apr 11 (pp. 145-153).
[https://dl.acm.org/doi/pdf/10.1145/2872427.2883062]

Used Algorithm:

- NLP - Supervised Classification Method
- Features can be divided into four classes: N-grams, Linguistic, Syntatctic, and Distrbutional Semantics.
-

Assumption: N/A
Code: N/A
Privacy/Security Considerations: N/A

Summary:

1. Develop a new classification methods: We develop a supervised classification methodology with NLP features to outperform a deep learning approach. We use and adapt several of the features used in prior art in an effort to see how they perform on the same data set.
2. We make public a new dataset of 1000 user comments from different domains.
3. This is the first longitudinal study of a computational approach to abusive language detection.

Datasets:

1. Comments posted on Sampled from Yahoo! Finanace and News between October 2012 and January 2014. Reported as “Abusive” and “clean” by users/yahoo.

2. Temporal Dataset

3. WWW2015 Data Set

4. Amazon Turk Experimen

Table 1: Primary Data Set Statistics

Finance data		News data	
Clean	705,886	Clean	1,162,655
Abusive	53,516	Abusive	228,119
Total	759,402	Total	1,390,774

Table 2: Temporal Data Set Statistics

Finance data		News data	
Clean	433,255	Clean	655,732
Abusive	15,181	Abusive	70,311
Total	448,436	Total	726,073

Table 3: WWW2015 Statistics

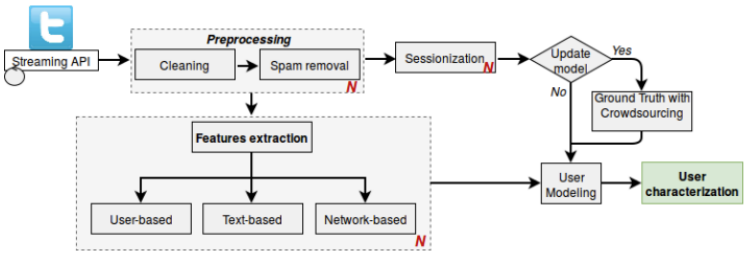
Clean	895,456
Abusive	56,280
Total	951,736

2. Paper Name: **Mean Birds: Detecting Aggression and Bullying on Twitter**

Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A. Mean birds: Detecting aggression and bullying on twitter. InProceedings of the 2017 ACM on web science conference 2017 Jun 25 (pp. 13-22).
[https://arxiv.org/pdf/1702.06877]

Used Algorithm: Random Forrest after Spam Removal and Feature Extraction
Assumption: N/A
Code: N/A
Privacy/Security Considerations: N/A
Summary:

1. we present a principled and scalable approach to detect bullying and aggressive behavior on Twitter.
2. We propose a robust methodology for extracting text, user, and network-based attributes, studying the properties of bullies and aggressors, and what features distinguish them from regular users.



3. Focus on Cyberbulling and user with hateful prompts behaviour analysis

Dataset:

1. 1.6M tweets posted over 3 months, available via request

3. Paper Name: **One-step and Two-step Classification for Abusive Language Detection on Twitter**

Park JH, Fung P. One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206. 2017 Jun 5.

Used Algorithm: CharCNN, WordCNN, HybridCNN [one-step] and Logistic Regression [two-steps]

Assumption: N/A

Code: N/A

Privacy/Security Considerations: N/A

Summary:

- In this research, we aim to experiment with a two-step approach of detecting abusive language first and then classifying it into specific types and compare it with a one-step approach of doing one multiclass classification on sexist and racist language.
- These experiments aimed to see whether dividing the problem space into two steps makes the detection more effective.

Dataset	**One-step**			**Two-step-1**		**Two-step-2**	
Label	None	Racism	Sexism	None	Abusive	Sexism	Racism
#	12,427	2,059	3,864	12,427	5,923	2,059	3,864

Table 1: Dataset Segmentation

Dataset:

- 20 thousand public english tweets

4. Paper Name: **Do you really want to hurt me? Predicting Abusive Swearing in Social Media**

Pamungkas EW, Basile V, Patti V. Do you really want to hurt me? predicting abusive swearing in social media. InProceedings of the Twelfth Language Resources and Evaluation Conference 2020 May (pp. 6237-6246).

Used Algorithm: LSVC, RF, LR, BERT
Assumption: N/A
Code: N/A
Privacy/Security Considerations: N/A

Summary:

- Taking the possibility of predicting the abusiveness of a **swear word** in a tweet context as the main investigation perspective
- we adapt the BERT Transformer-based architecture (Devlin et al., 2019) with the pre-trained model for English bert-base-cased
-

Dataset:

- corpus of tweets selected from the training set of Offensive Language Identification Dataset (OLID)
- The corpus is available for research purpose at the following URL: https://github.com/dadangewp/SWAD

- *Papers [hate speech]:*

5. Paper Name: **Hate me, hate me not: Hate speech detection on Facebook [lots of cite]**

Del Vigna12 F, Cimino23 A, Dell'Orletta F, Petrocchi M, Tesconi M. Hate me, hate me not: Hate speech detection on facebook. InProceedings of the first Italian conference on cybersecurity (ITASEC17) 2017 Jan (pp. 86-95).

Used Algorithm:

- Support Vector Machines (SVM)
- Recurrent Neural Network named Long Short Term Memory (LSTM)

Assumption: Italian Language
Code: N/A

Summary:

- First hate speech classifier for Italian texts: Hate Speech Recognition for the Italian Language
- First, propose a variety of hate categories to distinguish the kind of hate.

Dataset:

- built a corpus of comments retrieved from the Facebook public pages of Italian newspapers, politicians, artists, and groups by *Crawler*.
- 17567 Facebook comments

6. Paper Name: **Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter**

Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. InProceedings of the NAACL student research workshop 2016 Jun (pp. 88-93).
[https://github.com/zeeraktalat/hatespeech]

Used Algorithm: n-grams
Assumption: N/A
Code: .csv avaialable
Privacy/Security Considerations: N/A

Summary:

- a data set of 16k tweets annotated for hate speech.

Dataset:

- a data set of 16ktweets annotated for hate speech.

7. Paper Name: **HateCheck: Functional tests for hate speech detection models**

Röttger P, Vidgen B, Nguyen D, Waseem Z, Margetts H, Pierrehumbert JB. HateCheck: Functional tests for hate speech detection models. arXiv preprint arXiv:2012.15606. 2020 Dec 31.

Used Algorithm:

- *Weight BERT*
- *Google Jigsaw's Perspective [testing]*
- *Two Hat's SiftNinja [testing]*

Assumption: N/A
Code: N/A
Privacy/Security Considerations: N/A

Summary:

- Functional Test: targeted diagnostic insights into model functionalities
- 29 Functional Test:
    - Interviews 21 NGO workers whose work directly relates ro online hate
    - Review of Previous Research
- Their results [models weakness]:
    - Academic models have biased target coverage
    - Models are overly sensitive to certain key phrases
- Model Weakness can create concrete harms

- *penalising communities* that are most often targeted by online hate to begin with,
- *undermining positive efforts* to fight hate speech, and
- *reinforcing biases* in how different communities are protected in online spaces.

Dataset:

- **Automated hate speech detection and the problem of offensive language:** 24,783 tweets annotated as either hateful, offensive or neither
- Large
- **scale crowdsourcing and characterization of Twitter abusive behavior:** 99,996 tweets annotated as hateful, abusive, spam and normal

8. Paper Name: **BERT-based ensemble learning for multi-aspect hate speech detection**

Mnassri K, Rajapaksha P, Farahbakhsh R, Crespi N. Bert-based ensemble approaches for hate speech detection. InGLOBECOM 2022-2022 IEEE Global Communications Conference 2022 Dec 4 (pp. 4649-4654). IEEE.
[https://arxiv.org/pdf/2209.06505]

Used Algorithm: BERT
Assumption: N/A
Code: N/A
Privacy/Security Considerations: N/A

Summary:

- *This paper focuses on classifying hate speech in social media using multiple deep models that are implemented by integrating recent transformerbased language models such as BERT, and neural networks.*
-

Dataset:

- available Twitter datasets (Davidson, HatEval2019, OLID) [DHO]

9. Paper Name: **Hate Speech: Detection, Mitigation and Beyond**

Saha P, Das M, Mathew B, Mukherjee A. Hate Speech: Detection, Mitigation and Beyond. InProceedings of the Sixteenth ACM International Conference on Web Search and Data Mining 2023 Feb 27 (pp. 1232-1235).

Tutorial: https://hate-alert.github.io/talk/aaai_tutorial/Tutorial_AAAI_2022.pdf
Code: https://github.com/hate-alert/Tutorial-Resources/tree/main?tab=readme-ov-file

Used Algorithm: BERT
Assumption: N/A
Code: Yes
Privacy/Security Considerations: N/A

Summary:

- Current Methods: Bert-based Methods, Re-training BERT with banned subreddit data → HateBERT
-

Datasets:

1. Different datasets have different **taxonomies**.
2. Different datasets have different **sources**. Twitter is one of the major sources
    a. **Automated hate speech detection and the problem of offensive language:** 24,783 tweets annotated as either hateful, offensive or neither Large
    b. **scale crowdsourcing and characterization of Twitter abusive behavior:** 99,996 tweets annotated as hateful, abusive, spam and normal
3. Different datasets have different **languages**, English being the prominent one.

- *Centralized Remaing Papers to Read:*

1. **Not All Counterhate Tweets Elicit the Same Replies: A Fine-Grained Analysis**
2. **Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies**
3. **Large-Scale Hate Speech Detection with Cross-Domain Transfer**
4. **Deep learning for hate speech detection in tweets**
5. **Hatexplain: A benchmark dataset for explainable hate speech detection**
6. **Deep learning models for multilingual hate speech detection**
7. **A multilingual evaluation for online hate speech detection**

- *Centralized Projects + Codes:*

Learn Hate Speech Recognition By Implementing NLTK Module - Projects Using Python | AISciences.io [code]

- https://www.youtube.com/watch?v=11koTpwmgzs

Detoxify

- https://github.com/unitaryai/detoxify/tree/master

Deep Learning for Hate Speech Detection in Social Media Comments:

- Over youtube and reddit comments
- The fine-tuned BERT model
- https://github.com/JensBender/hate-speech-detection?tab=readme-ov-file

Kaggle - Hate Speech and Offensive Language Detection:

- https://www.kaggle.com/code/kirollosashraf/hate-speech-and-offensive-language-detection/notebook

Kaggle - Hate Speech Deetection curated Dataset:

- https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset

Kaggle - Twitter Hate-Speech Detection (Different Model)

- RF, Decision Tree, Naive Bayes, KNN, Logistic Regression
- https://www.kaggle.com/code/subhajeetdas/twitter-hate-speech-detection-different-model/notebook

Kaggle - Work with davidson:

- https://www.kaggle.com/code/eldrich/intro-hate-and-offensive-language-classification

IRE-project-hatEcal2019

- https://github.com/ash0904/IRE-Project-hatEval-2019/tree/master?tab=readme-ov-file

ConvAbuse

- https://github.com/amandacurry/convabuse/tree/main

- *Centralized Surveys:*

1. **A survey on hate speech detection and sentiment analysis using machine learning and deep learning models [2023]**
2. **Resources and benchmark corpora for hate speech detection: a systematic review [survey]**
3. **A systematic review of Hate Speech automatic detection using Natural Language Processing [2023 - NLP]**
4. **Abusive content detection in online user-generated data: a survey [survey]**
5. **Hate speech detection: Challenges and solutions [survey]**
6. **Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources [survey]**
7. **Challenges and frontiers in abusive content detection [workshop]**

Federated Hate-Speech Recognition:

10. Paper Name: **A Federated Learning Approach to Privacy Preserving Offensive Language Identification**

Zampieri M, Premasiri D, Ranasinghe T. A Federated Learning Approach to Privacy Preserving Offensive Language Identification. arXiv preprint arXiv:2404.11470. 2024 Apr 17.

Published at: Accepted to TRAC 2024 (Fourth Workshop on Threat, Aggression and Cyberbullying) at LREC-COLING 2024

Summary: This paper likely focuses on several key areas:

1. Offensive Language Detection: This section explores existing research on methods and techniques for identifying offensive language in social media. It covers various approaches, including rule-based systems, machine learning models, and deep learning architectures.
2. Privacy Concerns in Social Media: Here, they delve into literature addressing the privacy implications of offensive language detection methods. It discusses how current approaches often involve centralized data storage and the associated risks to user privacy.
3. Federated Learning (FL): This part reviews studies related to FL, emphasizing its role in decentralized model training without compromising user data privacy. It discusses how FL enables collaborative learning across distributed devices while keeping data local.
4. Model Fusion Techniques: The literature survey explores existing techniques for aggregating models in FL setups. It examines approaches for combining locally trained models to achieve improved performance while maintaining privacy.
5. Benchmark Datasets and Evaluation Metrics: This section covers publicly available datasets commonly used for evaluating offensive language detection models, such as AHSD, HASOC, HateXplain, and OLID. It also discusses standard evaluation metrics used in this domain.
6. Cross-Lingual Experiments: The survey touches upon research concerning cross-lingual offensive language detection, including studies that explore model transferability and performance across different languages, such as English and Spanish.

Overall, this paper provides a comprehensive overview of existing research in offensive language detection, privacy-preserving techniques, and the application of FL in this context while highlighting gaps and opportunities for further exploration.

11. Paper Name: **Pars-HaO: Hate and Offensive Language Detection on Persian Tweets Using Machine Learning and Deep Learning**

Sheykhlan MK, Shafi J, Kosari S, Abdoljabbar SK, Karimpour J. Pars-HaO: Hate and Offensive Language Detection on Persian Tweets Using Machine Learning and Deep Learning. Authorea Preprints. 2023 Dec 7.

Summary : This paper covers several key areas:

1. Offensive Language and Hate Speech Detection in Social Media: This section reviews existing research on automatic detection of offensive language and hate speech, particularly in the context of social media platforms. It discusses various methods and techniques employed, including machine learning models, deep learning architectures, and the challenges associated with detecting such content.
2. Research and Datasets in Persian Text: This part explores the existing literature and datasets available for identifying hate speech and offensive language in Persian text. It would highlight the current gaps in research and the scarcity of datasets in this domain.
3. Introduction of Pars-HaO Dataset: The paper discusses the importance of introducing the Pars-HaO dataset, which contains 8013 tweets in Persian language. It would explain the methodology used for collecting the dataset, including the sources of comments and the annotation process.
4. Machine Learning Models as Baselines: Here, they cover commonly used machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) as baselines for offensive language detection. It discusses their advantages and limitations in this context.
5. Deep Learning Models and BERT-based Techniques: This section delves into deep learning models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and their combinations with BERT embeddings. It reviews how these models have been applied to offensive language detection tasks and their performance compared to traditional machine learning approaches.
6. Experimental Results and Evaluation Metrics: The paper discusses the experimental setup, including training and testing procedures, as well as evaluation metrics used to assess model performance. It analyzes the results obtained from different models and techniques applied to the Pars-HaO dataset.

Overall, the paper provides insights into the current state of research in offensive language detection, the scarcity of resources in Persian text, the introduction of the Pars-HaO dataset, and the performance of various machine learning and deep learning models on this dataset.

12. Paper Name: **Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability**

Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics. 2015 Feb 1;32(1):89-97.

Summary: The paper encompasses the following key areas:

1. Automatic Cyberbullying Detection Systems: This section reviews existing research on automatic cyberbullying detection systems on social networking sites. It discusses various methods and technologies for monitoring and identifying cyberbullying behaviors, including natural language processing, machine learning, and deep learning techniques.
2. Desirability and Requirements of Automatic Monitoring: This paper explores the literature discussing the desirability of automatic monitoring of cyberbullying and its associated requirements. It covers studies investigating the effectiveness, ethical considerations, and societal implications of implementing such systems.
3. Qualitative Studies and Expert Opinions: This part delves into qualitative studies and expert opinions regarding the desirability and requirements of automatic cyberbullying monitoring. It discusses methodologies for soliciting and analyzing expert feedback, including open-ended questions and qualitative content analysis.
4. Expert Perspectives on Automatic Monitoring: The survey discusses the findings from the study where 179 experts in the field of cyberbullying were contacted, and 50 (28%) responded. It summarizes the opinions expressed by these experts, including their support for automatic monitoring, conditions for implementation, and concerns regarding privacy and feasibility.
5. Conditions and Priorities for Implementation: This section highlights the conditions specified by experts for implementing automatic monitoring systems, such as effective follow-up strategies, protection of adolescents' privacy, and empowerment of the involved parties. It discusses the priorities identified by experts for detection, such as threats and the misuse of pictures.
6. Doubts and Challenges: The paper covers doubts and challenges expressed by some experts regarding the desirability and feasibility of automatic monitoring. It explores concerns related to the effectiveness of follow-up strategies, severity-based intervention, and the need for further research.
7. Future Research Directions: Finally, the paper discusses recommendations for future research, including the importance of incorporating the perspectives of adolescents, parents, and social network providers. It suggests areas for further investigation, such as user desirability, prioritization of cyberbullying detection, and the effectiveness of follow-up strategies.

Overall, the paper will provide insights into the current understanding of automatic cyberbullying detection, expert opinions on its desirability and requirements, and avenues for future research in this domain.

13. Paper Name: **Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives**

Chopra S, Sawhney R, Mathur P, Shah RR. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. InProceedings of the AAAI conference on artificial intelligence 2020 Apr 3 (Vol. 34, No. 01, pp. 386-393).

Summary : This paper covers several key areas:

1. Code-Switching and Hate Speech Detection:
    - Review of existing literature on code-switching in linguistically diverse, low resource languages.
    - Exploration of methodologies used for hate speech detection in code-switched text.
    - Identification of challenges in detecting hate speech in such contexts due to semantic complexity.
2. Existing Approaches and Methodologies:
    - Discussion of current methodologies used for hate speech detection, including profanity modeling, deep learning, and author profiling.
    - Overview of techniques applied to real-world data for hate speech detection.
    - Evaluation of existing methodologies and their limitations in detecting hate speech in code-switched languages.
3. Proposed Three-Tier Pipeline:
    - Explanation of the three-tier pipeline proposed in the paper, including profanity modeling, deep graph embeddings, and author profiling.
    - Description of how each tier contributes to hate speech detection in Hindi-English code-switched language (Hinglish) on social media platforms like Twitter.
    - Discussion on the novelty of the proposed pipeline and how it addresses the identified challenges.
4. Comparison Against Baselines:
    - Analysis of comparison results against several baselines on two real-world datasets.
    - Evaluation of the performance of the proposed pipeline quantitatively and qualitatively.
    - Discussion on how targeted hate embeddings combined with social network-based features outperform existing state-of-the-art approaches.
5. Expert-in-the-Loop Algorithm for Bias Elimination:
    - Introduction of an expert-in-the-loop algorithm for bias elimination in the proposed model pipeline.
    - Examination of the prevalence and performance impact of debiasing techniques.
    - Analysis of the effectiveness of the expert-in-the-loop approach in improving hate speech detection accuracy.
6. Deployment Considerations:

- Discussion on the computational, practical, ethical, and reproducibility aspects of deploying the proposed pipeline across the web.
- Examination of challenges and considerations in deploying hate speech detection systems on social media platforms.
- Exploration of ethical considerations such as user privacy, fairness, and potential unintended consequences.

Overall, the literature survey of this paper here provides a comprehensive overview of existing research in hate speech detection, methodologies for code-switched languages, the proposed three-tier pipeline, comparison against baselines, expert-in-the-loop algorithms, and deployment considerations.