CSE 572: Data Mining
Final Project Literature Review

Project Title: IDENTIFICATION OF CRITICAL RELATED SUB-COMPONENTS IN ONLINE CONVERSATIONS

Team members:

Full name	ASU ID
Dhruva Patel	1230583512
Priyesha Jethi	1229734729
Sagar Rajesh Kumar Sinha	1229611034
Swati Kashyap	1225466398

Step 1: Summary of relevant work

[1] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. arXiv:1909.04251v1 [cs.CL] (Sep. 2019).

Summary

- The paper introduces the novel task of generative hate speech intervention, aimed at automatically generating responses to counteract the subproblem in online conversations through the utilization of several Natural Language Processing (NLP) techniques.

- Two large-scale, fully-labeled datasets collected from Gab and Reddit were presented in the study. The former source comprised 11, 825 conversations with 33, 766 posts and the latter included 5, 020 conversations with 22, 324 comments.

- Experiments conducted assessed the performance of various NLP models, including Seq2Seq, Variational Auto-Encoder (VAE), and Reinforcement Learning in generating effective intervention responses, with findings based on both automated evaluation metrics and human judgment.

Strengths

- According to Most current content moderation policies primarily target the identification and flagging of hateful posts and comments, with extreme cases resulting in the banning of users who create or promote such content. However, a significant drawback of these measures is that users often migrate to alternative platforms with moderation policies aligning with their ideologies or thought processes. The authors specifically emphasize that such actions discourage free speech. Instead, a more effective strategic response to

hate speech is to promote more speech. Intervention aims at encouraging individuals to alter their thoughts rather than simply modifying their actions.

- A significant advancement highlighted in the work is the incorporation of conversational context into the intervention of the aforementioned subproblem. By treating online posts not as individual instances but as parts of broader conversations, the proposed method offers a more nuanced and effective approach to identifying and responding to hate speech.

- The study not only establishes a benchmark for future research in generative hate speech intervention through detailed dataset analysis and evaluation of common automatic response generation methods, but also showcases a comprehensive approach by applying and assessing a variety of NLP models, including Seq2Seq, Variational Auto-Encoder (VAE) and Reinforcement Learning. This dual focus on benchmarking and diversity provides a solid foundation and a wide array of tools for advancing the field.

Limitations

- The authors exclusively focussed on datasets from Reddit and Gab. Moreover, the dataset collection timeframe is limited, since it was specifically stated that conversations from Gab were solely extracted from October 2018. Thus, the temporal range is too short, which doesn't capture the evolution of diverse human interactions over time. If they could have incorporated data from a longer timespan, it would have probably provided a more comprehensive understanding of hate speech trends over the years.

- A notable limitation of the study is the limited conversational depth of the collected datasets. This particular filtering may overlook the complex engagement dynamics amongst users in longer conversational threads. Consequently, the effectiveness of these interventions might not fully account for the nuances in longer dialogues.

- While AI assisted hate speech moderation represents a promising area of research in a real-time scenario, the study's findings point that human interventions are still superior in terms of effectiveness and appropriateness. Hence, this presents the need for more research into AI models which can more closely mimic human levels of understanding and empathy.

[2] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. In Social Media + Society, October-December 2021, 1-14. https://doi.org/10.1177/20563051211052906

Summary

- The research delved into the absence of a comprehensive empirical study to establish the prevalence of hate speech in altright fringe communities, and worse, the majority of previous studies used automated methods searching for highly apparent “hate terms” such as “chink” and “nigger”, not sensitive to searches for more subtle forms of hate. The authors instead proposed that such comments had to be grouped together and studied as clusters for gaining insights into their more covert representations.

- The study analyzed user comments from Reddit, 4chan and 8chan to calculate the percentage of user comments which contained explicit or implicit hate speech, the method of expression of these tweets, potential victims of such comments, and also analyzed the topical structure of coded user comments.

Strengths

- By analyzing the subproblem across multiple platforms (Reddit, 4chan and 8chan), the study provided a comparative view of how those communities might differ in terms of the sub-component’s prevalence and nature, thereby enhancing understanding of the digital ecosystem in which these channels operate.

- The study could also aid in detection of microaggressions, which are quite subtle forms of hate mostly expressed through non-verbal cues, tone of voice or indirect language.

- The study identified significant gaps in the research landscape, such as the need for analyzing visual hate speech and the effects of hate speech normalization. By doing so, it set a clear agenda for future research, emphasizing areas that are crucial for a more thorough understanding of online hate speech.

Limitations

- The analysis faced several challenges in data collection, leading to a varying base of comments across platforms, which could potentially influence the comprehensiveness and comparability of findings.

- The analysis was limited to textual content, neglecting the potential impact and prevalence of visual hate speech in the form of images, memes and videos, which are significant components of online discourse in the studied communities.

- The data was collected in a specific timeframe (April 2019), which would potentially limit the findings’ applicability over time or under different sociopolitical contexts.

[3] Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. Semantic Web 1, 0 (October 2018), 1–5. arXiv:1803.03662v2 [cs.CL].

Summary

- The study probed the performance disparity in accurately determining hate vs non-hate speech, due to the former often lacking unique, discriminative features, making it difficult to detect as it resides in the dataset's long tail.

- Deep Neural Networks (DNNs) were highlighted as potent tools for feature extraction, which could capture the complex semantics of the subproblem, thereby suggesting an advanced approach to overcoming traditional detection technologies.

- User comments from Reddit, 4chan and 8chan were analyzed to calculate the percentage of explicit or implicit hate-based user tweets, method of expression of those tweets, potential victims, and the topical structure of coded user comments.

Strengths

- The newly proposed DNN models, especially CNN + sCNN (Skipped CNNs), significantly outperformed state-of-the-art methods in hate speech detection, demonstrating their effectiveness in identifying challenging "long tail" content with low uniqueness scores. The analysis showed the models' robustness with different word embeddings and highlighted ongoing challenges in hate speech detection, such as interpreting context and managing non-discriminative features. The results underscored the potential of advanced DNN structures in improving the accuracy and reliability of automated hate speech identification on social media platforms.

- The authors conducted an in-depth analysis to illuminate the unbalanced nature and feature scarcity in hate speech datasets, addressing critical challenges in the field.

- The study extensively validated these methods across the largest compendium of Twitter datasets in this field, thereby proving their effectiveness in identifying and classifying hate speech content. Also, it established new benchmarks for research, contributing significantly to the advancement of the particular sub-component detection technologies.

Limitations

- The research relied heavily on pre-trained word embeddings (eg:, Word2Vec, GloVe, Twitter) for semantic feature extraction. While effective, this approach might limit the model's adaptability to constantly evolving language on different social media platforms. An area of improvement could be developing dynamic embedding techniques that evolve with new linguistic patterns and slang emerging on social media, ensuring the model remains effective over time.

- The models, especially in detecting subtle or implied hate speech, might struggle with tweets requiring contextual or background knowledge for accurate classification. Improving the models' ability to consider the broader context of a conversation or the cultural and societal nuances behind certain phrases could enhance their accuracy in identifying complex hate speech instances.

- While the models were tested on a substantial collection of Twitter datasets for hate speech, their generalizability across different social media platforms or languages remains an open question. Expanding the evaluation to include diverse datasets from various platforms and languages could help assess the models' robustness and adaptability, guiding further refinement for broader applicability.

[4] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, San Francisco, CA, USA, 855-864. DOI: http://dx.doi.org/10.1145/2939672.2939754

Summary

- The paper presented a scalable algorithm for feature learning in networks, optimizing a novel objective to preserve network neighborhoods in a low-dimensional feature space. To do so, it used a biased random walk procedure to explore diverse neighborhoods efficiently.

- It provided a framework that could adapt to capture both homophily (community-based) and structural equivalence (role-based) similarities among nodes, crucial for understanding the dynamics of online conversations.

Strengths

- Node2Vec's ability to capture diverse neighborhood structures through its biased random walk can be particularly beneficial for analyzing online conversations, where interactions and roles may vary widely.

- The scalable nature of Node2Vec, combined with its efficient random walk sampling and optimization process, makes it able to handle potentially large and complex datasets derived from social platforms.

- The algorithm's flexibility at modeling different types of node relationships (homophily and structural equivalence) could be used to extract several implicit features which otherwise would have been quite difficult across different subproblems.

Limitations

- The algorithm requires tuning of hyperparameters (eg., return parameter p, in-out parameter q) for optimal performance, which could be challenging without sufficient domain knowledge or labeled data for specific contexts like online discussions.

- While efficient at exploring network neighborhoods, Node2Vec's performance is inherently tied to the underlying network structure. On online platforms, the dynamic nature of interactions and the presence of noise (e.g., irrelevant or misleading comments) may affect the algorithm's effectiveness.

[5] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2022. BertGCN: Transductive Text Classification by Combining GCN and BERT. arXiv:2105.05727v4 [cs.CL], March 21, 2022

Summary

- BertGCN presents an innovative approach by integrating BERT's deep semantic understanding with GCN's capacity to model related context. This amalgamation could provide a nuanced perspective on complex online conversations within platforms like Reddit.

- It provides a framework that could adapt to capture both homophily (community-based) and structural equivalence (role-based) similarities among nodes, crucial for understanding the dynamics of online conversations.

Strengths

- BertGCN's utilization of BERT embeddings ensures a deep semantic understanding of textual content. This feature is critical for accurately identifying nuanced instances of hate speech, trolling, and deception, which are often context-dependent and subtle.

- The GCN component can analyze the structure of online conversations, identifying patterns that may signal the transition from collaborative to canceling discourse. This structural insight could be vital for understanding how toxic behaviors propagate within discussion threads.

Combining both of these components, this makes it well-suited to explore the multifaceted nature of graph data in general, and online discussions in particular.

Limitations

- Determining the most effective way to construct the graph (i.e., defining nodes and edges) for any specific task may require extensive experimentation. The quality of the graph significantly impacts the model's ability to capture discourse complexity

- The model, owing to its complexity, necessitates considerable computational resources, which could potentially limit the scalability of analysis across extensive datasets like the complete archive of Reddit conversations.

- BertGCN's effectiveness in identifying the nuanced categories of interest of our study (hate speech, trolls, and deception) depends on the availability of accurately labeled

data. The need for detailed annotation to train the model poses challenges given the vast and varied nature of online discourses.

[6] Anna Vartapetiance and Lee Gillam. 2014. Deception detection: dependable or defective?. *Social Network Analysis and Mining 4*. Retrieved from https://doi.org/10.1007/s13278-014-0166-8

Summary

- The paper investigates how people distinguish between truthful and deceptive statements, and whether machines can accurately identify the veracity of claims. The authors critically analyze the challenges in deception research, highlighting how the type of deception and the context impact the effectiveness of detection methods. The paper ultimately aims to provide insights into the complexities of deception detection and its practical applications.

- The author says deception includes actions done on purpose to trick others. There's a difference between doing something to mislead people and doing something without meaning to do that. The definition also covers deceiving by not doing anything, like not correcting someone's false belief so they keep believing it. The paper separates lies from deception. Lies are a type of deception. With lies, people say things they know aren't true to make others think those false statements are true facts. This shows deception has complex levels.

- The report examines indicators used to spot lies in writing. These clues are based on the length of sentences and word count. Also, specific words, complex phrasing, and impressions matter. But studies disagree on how liars' texts differ from truth-tellers. There are contradictions in how sentence length, word complexity, etc. should change with deception. The authors say there is no standard set of signs to reliably identify lies across all situations. Without consistent patterns, deception spotting methods don't work well.
    - Deception detection may rely on features like response length, talking time, unique words, generalizing terms, self-references, repetitions, negative statements, and extreme descriptions. But experts disagree whether these cues increase or decrease in deceptive text.
    - The researchers explored Pennebaker's LIWC categories like self-references, negative words, exclusive words, and motion verbs. They might indicate deception. However, the expected changes varied across studies. It's uncertain if Pennebaker proposed them as deception cues. The authors stressed detecting deception is tough without baseline values when using raw frequencies.
    - Experts like DePaulo, Pennebaker, and Burgoon suggest different ways to spot lies. But their ideas often disagree, even on what clues to look for and how they should look. This paper says simply counting words from LIWC groups may not work well to tell lies from truth.

    - The researchers did experiments with various data, like news articles, science summaries, and Enron emails. They studied cues that might show deception. Their findings showed it's hard to reliably detect deception "in the wild" using only those cues. They also looked at readability measures.

Strengths

- This paper presents an extensive overview of previously published research on detecting deception. It covers a wide range of methods, techniques, and the challenges associated with this field. By taking this comprehensive approach, the authors aim to provide readers with a deep understanding of the topic.

- The authors critically evaluate various deception detection methods, highlighting their strengths and weaknesses. This critical analysis allows readers to assess the reliability and efficacy of these methods in real-world scenarios.

- The paper explores practical applications of deception detection techniques in various fields, including law enforcement, intelligence, and interpersonal communication. The paper provides practical and policy-relevant insights by examining real-world applications.

Limitation

- While the paper provides a thorough review of existing methods, it does not include a rigorous empirical evaluation or comparative analysis of the techniques discussed, making it difficult to determine their relative strengths and weaknesses.

- The authors acknowledge the possibility of bias in deception detection methods, but they do not provide in-depth analysis or strategies to mitigate these biases, which are critical for ensuring fair and equitable applications.

- The paper discusses the challenges of generalizability in deception detection, but it provides no concrete solutions or approaches to address this issue, which is critical for developing robust and widely applicable methods.

[7] A. Zubiaga and Heng Ji. 2013. Tweet, but verify: epistemic study of information verification on Twitter. *Social Network Analysis and Mining 4*. Retrieved from https://doi.org/10.1007/s13278-014-0163-y

Summary

- The study focuses on the dynamic environment of Twitter, where real-time information flows quickly during events such as natural disasters. The primary goal is to understand how users interact and evaluate the veracity of information shared through tweets.

- The authors conducted a qualitative study by analyzing Twitter conversations about the 2013 Hurricane Sandy. The study aims to understand how Twitter users assess the veracity of information, identify rumors and misinformation, and collaborate to verify or debunk claims. The authors identified several verification strategies used by users, such as seeking authoritative sources, cross-validating information across multiple sources, evaluating source credibility and trustworthiness, fact-checking and debunking rumors or misinformation, and using multimedia evidence (e.g., images, videos) to corroborate or refute claims.

- The study takes in Fallis identifies four critical features that aid humans in making accurate assessments:
    - Authority - The paper emphasizes the importance of using authoritative sources on Twitter, such as official accounts, news agencies, and trusted organizations, to verify information during high-impact events. Users trust these authoritative sources to provide accurate information. Finally, it was determined that disinformation on Twitter shares characteristics with both propaganda and mainstream news, making it difficult to detect. Furthermore, relying solely on authority cues (e.g., verified accounts) may not provide accurate verification.
    - plausibility and support - The study discovered that Twitter users frequently cross-validate information from multiple sources to determine its plausibility and support. They look for supporting evidence from multiple sources to increase the likelihood that the information is correct. Finally, it was discovered that disinformation incorporates linguistic cues from both fake news and legitimate sources, blurring the line. Users must evaluate content critically, looking beyond plausibility and taking into account additional features.
    - Independent corroboration - The use of multimedia evidence, such as images and videos, was identified as a key verification strategy among Twitter users. Independent corroboration through visual or multimedia content assists users in validating or refuting claims and rumors. Finally, the authors concluded that cross-validation of multiple tweets improves credibility. Encouraging users to seek out independent sources can help improve verification accuracy.
    - Presentation - While not explicitly discussed in the paper, the presentation of information on Twitter, such as the source credibility, language used, and the way it is framed or presented, can influence users' perceptions of its accuracy and reliability. Finally, it was determined that disinformation headlines frequently

display "clickbaiting," which attracts attention. Users can assess the veracity of tweets by taking into account their headline format and language.

Strengths

- The qualitative approach provides rich insights into user behavior and strategies through in-depth analysis of real-world Twitter conversations. It captures real-world behavior and trust dynamics by polling users on their perceived credibility.

- Paper provides evaluating features from epistemology which provides a theoretical grounding for verification. Authors have also provided a framework and understanding for the development of automated verification systems or tools. Comparing user perceptions with professional evaluations ensures robustness.

- Authors have ensured to address the critical issue of misinformation and rumor spreading on social media during high-impact events. The study focuses on information verification during fast-paced events, which is crucial for accurate reporting.

Limitations

- The study focused on two specific crisis events, which may limit the findings' applicability to other types of events or contexts. The study focuses on tweets about Hurricane Sandy, which limits its applicability to other events.

- Because this is a qualitative study, data analysis and interpretation may be influenced by researcher bias or subjectivity. Furthermore, some features (such as author details) may not be easily accessible or discernible on Twitter. Behavioral patterns discovered in tweets may not fully capture the complexity of user verification behavior.

- Social media platforms and user behaviors are constantly evolving, which may render some findings or implications obsolete over time.

[8] Luis Gerardo Mojica de la Vega and Vincent Ng. 2018. Modeling trolling in social media conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC '18), May 7-12, 2018, Miyazaki, Japan. ACM, New York, NY, USA, 123-130.

Summary

- The paper presents an innovative method for automatically detecting trolling behavior in textual comments, addressing the growing concern of online disruption and maintaining constructive online discourse. The authors' approach incorporates linguistic features and sentiment analysis techniques, allowing for a multifaceted analysis of comment content to identify trolling behavior accurately.

- The proposed methodology also takes into account contextual cues within comments, enabling a deeper understanding of the intended meaning behind comments and facilitating more precise trolling detection.

- Leveraging machine learning techniques, the authors develop a scalable solution for trolling detection, which has the potential for practical implementation in online platforms to enhance content moderation efforts and user experience. The paper includes a comprehensive evaluation of the proposed model, demonstrating its effectiveness across various online platforms and comment types, thereby establishing its credibility and potential for real-world application.

Strengths

- Comprehensive Methodology: The paper introduces a multifaceted approach that considers linguistic features, sentiment analysis, and contextual cues, providing a holistic framework for detecting trolling behavior in textual comments.

- Rigorous Evaluation: The authors thoroughly evaluate their model on a diverse dataset, showcasing its effectiveness across different platforms and comment types. This rigorous evaluation enhances the credibility and generalizability of their proposed method.

- Practical Applicability: Leveraging machine learning techniques, the proposed approach offers scalable and automated trolling detection solutions, which could be implemented by online platforms to improve content moderation and user experience.

Weaknesses

- Potential Bias in Dataset: The effectiveness of the model heavily relies on the quality and representativeness of the training dataset. If the dataset is biased or limited in scope, it may affect the model's performance and generalizability.

- External Validity: The paper lacks discussion on the external validity of the proposed method beyond the datasets used in the study, such as real-world implementation challenges and potential ethical implications, which limits the broader applicability and robustness of the approach.

Step 2: Organization of relevant work

Based on the summaries, strengths and limitations provided for each paper, we can organize the related work into two broad groups

- **Part A**: Data format and feature extraction for the chosen sub-components (hate speech, trolls and deception)
- **Part B**: Deep-learning based network modeling for the features extracted from above

Part A

Hate Speech

- [1] and [2]: Both papers focus on identifying and intervening in hate speech conversations, emphasizing the importance of conversational context and the subtleties of language. However, both papers are limited by their dataset's scope, either in terms of platform diversity or temporal coverage, potentially affecting the generalizability of their findings.
- [3] provides insight into how several potential discriminative features could be extracted before the data is fed into deep learning model(s) for classification of our nodes and the sub-problem. They are as follows:
    - Node embeddings for posts/comments could be generated using word embeddings like Word2Vec and Glove
    - Uniqueness score for bigrams and trigrams
    - Part-of-Speech (POS) tags
    - Term Frequency Inverse Document Frequency (TFIDF) weighted top-k keyword extraction
    - Number of syllables in each comment
    - Flesch-Kincaid Grade Level and Flesch Reading Ease Scores to measure a document's readability
    - Sentiment and polarity based features using public APIs and libraries

Trolls

- [8] proposes a comprehensive categorization scheme for modeling online trolling that considers perspectives from both the troll and the responders. It defines four aspects:
    - the troll's intention behind their comment (trolling, playing/mocking, or normal)
    - whether they disclosed or hid their real intention
    - the responder's interpretation of the troll's perceived intention,
    - the responder's strategy in reaction (engage, praise, counter-troll, play along, criticize, dismiss, or normal response).
- For prediction tasks, [8] investigates using n-gram text features (unigrams, bigrams, with POS tags) and averaged word embeddings from GloVe trained on Twitter data as initial

feature sets. The authors intend to identify challenging cases, investigate additional features from previous work on related tasks such as abusive language detection, and conduct additional analysis in the final version.

Deception

- [6] delves into the diverse definition of deception, emphasizing intentional creation of false beliefs over simple lying. The authors:
    - Examines previous research on verbal cues for detecting deception in text, revealing inconsistencies in their effectiveness.
    - Discovered that traditional deception cues, such as word counts, were less effective.
    - Discovered techniques like stopword associations, lexical features, and specific request categories produced better results.
    - Argues against the concept of universal verbal deception cues, instead advocating for domain-specific approaches to effective detection.
- [7] focuses on to verify the credibility of tweets during fast-paced events like natural disasters:
    - Fallis identifies four critical features: Authority(Details about the author), Plausibility and Support (text plausibility and picture plausibility, corroboration and presentation (Use of proper writing, spelling and grammar in the tweet).
    - The study discovered that providing more author details upfront and considering the plausibility of text/images were most helpful for accurate credibility assessments, whereas presentation quality was not. Repeated exposure to the same claim, even if inaccurate, tended to increase perceived credibility incorrectly.

Modeling

The following papers explore network-based methods to analyze and model the above mentioned features through Graph Neural Networks (GNNs)

- [4] and [5]: Introduce algorithms that leverage network structures to model the above mentioned features. [4] focuses on scalable feature learning through biased random walks, while [5] combines Bert's semantic analysis with GCN's structural modeling.