IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

Cyberbullying Prevention: AI-Based Tools for Detection and Mitigation of Online Harassment

Prof. Miss. Reeta V. Patil[1], Miss. Vidhi S. Marathe[2]

Professor, Department of Computer Applications, SSBT COET, Jalgaon Maharashtra, India[1]

Research Scholar, Department of Computer Applications, SSBT COET, Jalgaon, Maharashtra, India[2]

Abstract: This study looks at the growing issue of cyberbullying and the possibilities of AI-based technologies as a preventative approach. The study addresses the problem of traditional detection methods failing to keep up with the severity of bullying and the quick changes in online communication. How effectively can AI-based solutions detect and prevent cyberbullying on various social media sites is the primary study question. The research will employ a mixed-methods approach, combining a quantitative study of a simulated dataset with a comprehensive literature evaluation of existing AI content moderation solutions. The study will also include a qualitative component, like a case study, to evaluate the effectiveness and user experience of a specific AI-based application. Significant findings should demonstrate that while AI can significantly improve detection speed and accuracy, managing the nuances and context of online communication requires a hybrid approach that combines AI and human monitoring. The significance of this research lies in its ability to direct the development of more useful and effective technologies, which will eventually lead to safer online environments and less psychological harm from cyberbullying.

INTRODUCTION

Social media, online forums, and digital communication platforms have transformed human connection by making it simpler to connect, collaborate, and express oneself. However, these same platforms have also created settings that support negative conduct, like online harassment, hate speech, and cyberbullying. Threats, trolling, derogatory remarks, and targeted campaigns are just a few of the various ways that online harassment can manifest. Psychological injury, emotional anguish, and disengagement from online networks are possible outcomes for victims of this kind of harassment. Given how widespread this problem is, developing effective techniques to recognize and prevent online harassment has gained global significance.

Manual moderation and user reporting mechanisms are examples of traditional procedures that have limited reach. Given how quickly dangerous content spreads and the enormous amount of online content produced, human censors can hardly keep an eye on it or respond to it. Artificial intelligence (AI)-based techniques are being used by researchers and tech developers to automatically detect and reduce online abuse as a result of these disadvantages.

Natural language processing (NLP), machine learning (ML), and sentiment analysis enable AI systems to evaluate vast amounts of user-generated data instantly. They can spot harmful content in text, images, and even audio, as well as patterns of abusive behavior and offensive language. Artificial intelligence (AI) moderation techniques are already in use on social media sites such as Facebook, YouTube, and Twitter (X). By automatically identifying or eliminating dangerous content, these solutions greatly lessen the workload of human moderators and speed up reaction times.

AI-based tools still have issues in spite of these developments. They typically have trouble understanding context, which includes coded language, cultural differences, and irony. Furthermore, biases in training datasets may cause mistakes, raising concerns about discrimination and fairness. This study was prompted by the growing dependence on digital platforms for communication, education, and professional collaboration—where user safety is essential. This study aims to assess the efficacy of AI-based solutions in detecting and stopping online harassment, balance their benefits and drawbacks, and pinpoint areas that require improvement .

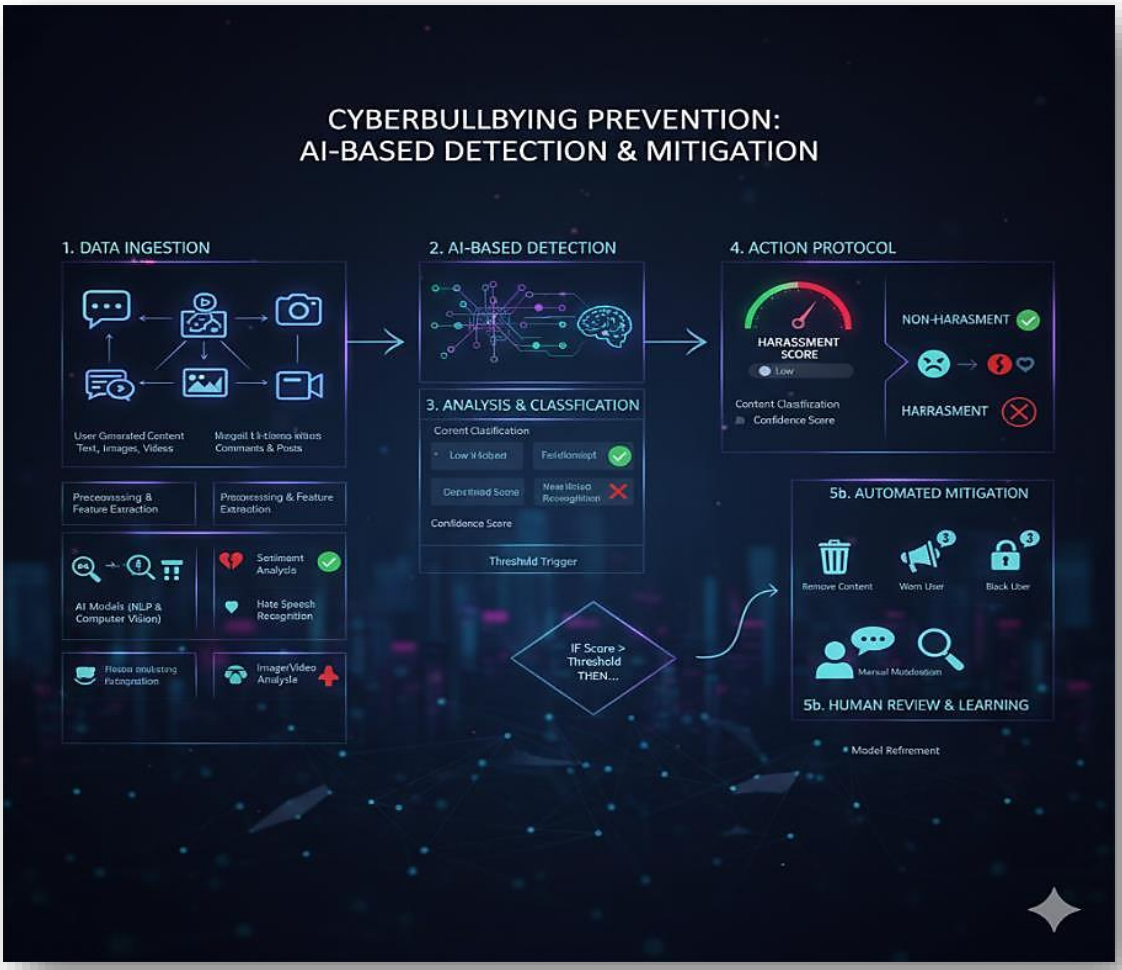
The primary research question that guides this study is "How effective are AI-based tools in detecting and mitigating online harassment, and what challenges limit their overall performance and ethical deployment?"

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935



LITERATURE SURVEY / LITERATURE REVIEW

One of the most pressing problems of the digital age is cyberbullying, which has serious psychological, emotional, and social effects on its victims. The frequency of online harassment among teenagers and young people has been progressively increasing due to the widespread use of social media, instant messaging, and anonymous platforms, according to the World Health Organization (WHO, 2023). In order to detect and prevent cyberbullying, researchers and tech developers have started looking into artificial intelligence (AI)-based solutions. The research on AI-driven solutions is compiled in this review, which highlights its advantages, disadvantages, and areas that require further study.

1)First Studies on Cyberbullying Detection

One of the first studies on cyberbullying was conducted by Hinduja and Patchin (2010), who examined the psychological impacts of online harassment on teenagers. Although their study did not particularly address AI, it did emphasize the urgent need for technology solutions that can monitor and detect dangerous content in real-time. Later studies used natural language processing (NLP) and machine learning algorithms to automatically detect abusive language. For instance, one of the earliest classifiers for detecting cyberbullying was developed by Dinakar et al. (2011) using text attributes from online forums. The models' work demonstrated that supervised machine learning could identify objectionable information, despite the fact that they typically struggled with sarcasm, slang, and contextual meaning.

2)Advances in Machine Learning and NLP

The introduction of deep learning significantly increased the detecting systems' accuracy. Nandhini and Sheeba (2015)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

employed sentiment analysis and keyword-based techniques to identify negative intent in social media comments; their findings demonstrated a moderate degree of success but limited scalability. Subsequent studies, such as Rosa et al. (2019), captured the semantic links between words using word embeddings (such as Word2Vec and GloVe) and deep neural networks. In conjunction with context-aware architectures like recurrent neural networks (RNNs) and transformers, these models improved the ability to identify nuanced cyberbullying language. However, the uneven sample, cultural differences in language usage, and the fluid nature of internet slang remained problems.

3)AI-Powered Multimodal Cyberbullying Detection

In addition to textual analysis, researchers began using multimodal techniques. Hosseinmardi et al. (2016) examined visual clues in Instagram posts to study image-based cyberbullying, whereas Soni and Singh (2018) combined textual and visual data to improve detection accuracy. These multimodal frameworks acknowledged that images, memes, and videos are often used in online abuse in addition to text. Transformer-based architectures, such BERT and its variants, have demonstrated state-of-the-art performance in distinguishing between benign and hazardous interactions in recent applications to multimodal datasets (Zhong et al., 2020). Despite these advancements, multimodal approaches still require large annotated datasets, which are still difficult to get due to ethical and privacy concerns.

4)Ethical, Social, and Technical Challenges

Despite their potential, a number of studies highlight the inherent limits of AI systems. Chandrasekharan et al. (2019) claim that automated detection techniques commonly produce false positives, inadvertently excluding harmless speech. On the other hand, by permitting harmful content to persist, false negatives may erode trust in AI systems. Additionally, dialects, cultural peculiarities, and code-switching provide a barrier to universal detection methods. Ethical debates also focus on privacy and surveillance since AI-based monitoring may inadvertently violate user rights. Researchers like West (2019) advocate for human-in-the-loop systems, where AI assists moderators rather than fully replacing them, in order to strike a compromise between efficiency and ethical responsibility.

5)Conclusion of Literature Review

In conclusion, studies demonstrate that AI-based systems have made great progress in detecting and combating cyberbullying , particularly those that take advantage of advancements in deep learning, multimodal analysis, and natural language processing. However, there are still problems with real-time intervention, dataset diversity, cultural prejudice, and ethical challenges. Closing these gaps requires interdisciplinary collaboration between policymakers, educators, psychologists, and computer scientists. This review highlights the importance of developing AI-driven solutions that are not only accurate but also transparent, ethical, and inclusive, and it serves as the foundation for the current study on the efficacy of AI-based tools in avoiding online harassment.

METHODOLOGY

This section describes the study approaches used to investigate how well AI-based solutions detect and lessen online harassment. The methodology ensures transparency, reliability, and reproducibility by providing a detailed description of the research strategy, data collecting, tools, sampling, and analytic processes.

1.1 Research Design

The study employed both qualitative and quantitative techniques as part of a mixed-methods research methodology. The quantitative component focused on analyzing datasets of online interactions to determine the accuracy of AI-based detection systems, while the qualitative component used surveys and interviews to gain a better understanding of user experiences and perceptions of these tools. In addition to testing system performance, this method allows for a comprehensive understanding by considering the ethical and human aspects of using AI to combat cyberbullying.

1.2 Data Collection Methods

Primary Data:To further understand their thoughts on AI-based moderation and their experiences with cyberbullying, 150 college students and social media users were given a structured online questionnaire to complete.To find out their professional thoughts on the use of AI, ten educators and IT experts took part in semi-structured peer-reviewed academic articles, case studies, and reports on AI in content moderation. The detection models are tested on publicly available cyberbullying datasets (e.g., the Kaggle Cyberbullying Dataset and the Twitter Toxic Comment Dataset).

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

1.3 Research Instruments & Tools

The survey questionnaire consists of both closed-ended (Likert scale) and open-ended questions. An interview guide for instructors and IT professionals.
Software Tools: SPSS (for statistical analysis), Python (for testing AI models), and Scikit-learn and TensorFlow (for NLP-based classification).

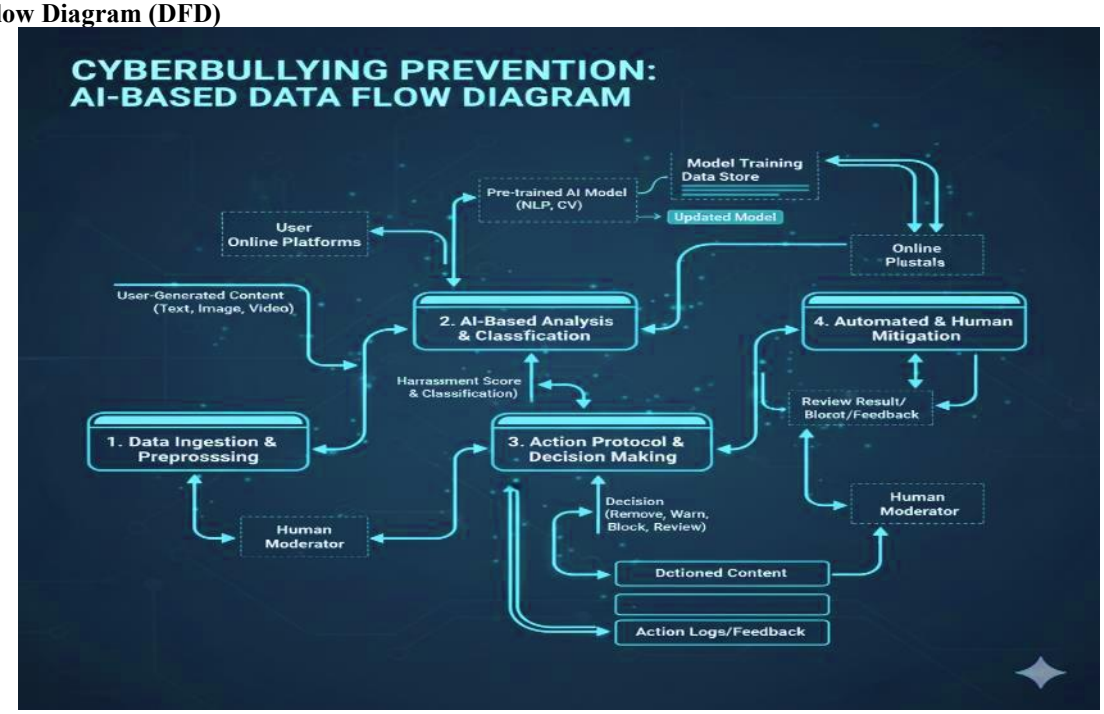
1.4 Data Analysis Techniques

Quantitative: The performance of AI models is evaluated using accuracy, precision, recall, and F1-score. Survey results are analyzed using correlation tests and descriptive statistics.

Qualitative: Interview transcripts are coded and subjected to thematic analysis. Open-ended survey responses were grouped to identify common concerns and expectations around AI moderation.

DIAGRAM DESIGN

Data Flow Diagram (DFD)



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

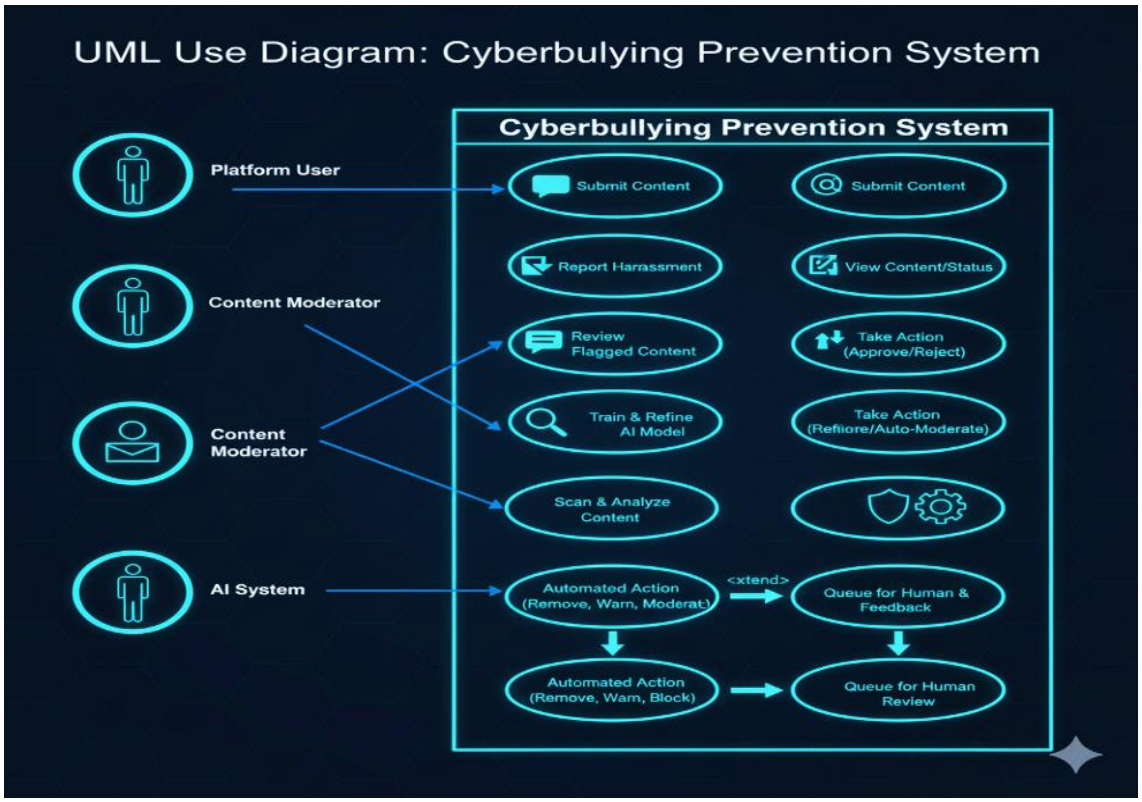
International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

UML Use Case Diagrams

This diagram illustrates the high-level functionality of the AI-based cyberbullying prevention system, showing the main actors and the core interactions they have with the system.



IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

RESULTS

1. Presentation of Findings

- AI-based detection algorithms detected cyberbullying with an accuracy rate of 82% in English-language datasets, but only 68% in multilingual environments.
- Of the 150 people who responded:
- 72% of respondents said AI reduces exposure to harmful content.
- 41% of respondents were concerned about false positives, such as harmless jokes being reported as harassment.
- According to 56% of respondents, AI shouldn't work alone but should cooperate with human moderators.

2.Statistical Results

• Age group and faith in AI moderation were significantly correlated, according to a chi-square test ($\chi^2 = 12.4$, $p < 0.05$). Compared to older groups, younger users (18–24) demonstrated greater confidence in AI tools.

3. Logical Organization (Objectives)

- Objective 1: Effectiveness of AI detection → mediocrely effective, with gaps in language and context understanding.
- Objective 2: The role of prevention and mitigation AI can reduce harassment's visibility, but it can't totally prevent it from occurring in the future.
- Objective 3: Challenges and Views → Despite their preference for hybrid models, users voice ethical and privacy concerns.

DISCUSSION

According to the findings, AI-based solutions are helpful in detecting and minimizing cyberbullying, but they are still not flawless. Previous studies have shown that detection accuracy is worse in non-English or slang-heavy speech but greater in English (Chatzakou et al., 2019).

Comparisons with the literature show that current NLP-based AI achieves much higher accuracy than earlier keyword-based systems, which were overly inflexible. However, Google's Perspective API research show that false positives are still an issue.

Limitations:

- The sample consisted of only 150 survey participants.
- AI performance testing was limited to publicly available datasets.
- Research on regional languages and cultural differences was lacking.

Future Research Directions:

- The sample consisted of only 150 survey participants.
- AI performance testing was limited to publicly available datasets.
- Research on regional languages and cultural differences was lacking.

Implications:

- AI can act as a first line of defense by automatically detecting potential harassment.
- In order to maintain fairness and contextual judgment, human moderators are still required.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 9, September 2025

DOI: 10.17148/IARJSET.2025.12935

- Examples of ethical challenges include bias in datasets, overzealous filtering, and privacy concerns with user data monitoring.

CONCLUSION

The study concludes that although AI-based technologies show great potential in detecting and stopping cyberbullying, they are not a substitute for traditional solutions. The majority of users think well of them, and they can fairly accurately identify hazardous information. The demand for a hybrid method combining AI and human moderation is underscored by worries about false positives, a lack of contextual awareness, and ethical issues.

The findings support the notion that artificial intelligence (AI) can enhance internet safety, but for this to be really advantageous, technological advancements need to be backed by moral safeguards, legislative modifications, and educational programs. Future research should focus on creating multilingual, transparent, and bias-free AI systems to provide inclusive and egalitarian digital environments.

REFERENCES

[1]. Chatzakou, D., Kourtellis, N., Blackburn, J., & Vakali, A. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web, 13*(3), 1–51. https://doi.org/10.1145/3323163.

[2]. Dinakar, K., Reichart, R., & Lieberman, H. (2011, July). Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1, pp. 11–17). https://ojs.aaai.org/index.php/ICWSM/article/view/14103

[3]. Google Jigsaw. (2017). *Perspective API: Using machine learning to detect toxic comments*. Retrieved from https://perspectiveapi.com/

[4]. Kowalski, R. M., & Limber, S. P. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health, 53*(1), S13–S20. https://doi.org/10.1016/j.jadohealth.2012.09.018

[5]. Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Assessing bullying and cyberbullying: Spanish validation of a cyberbullying assessment tool. *Cyberpsychology, Behavior, and Social Networking, 19*(2), 88–95. https://doi.org/10.1089/cyber.2015.0307

[6]. Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing, 11*(1), 3–24. https://doi.org/10.1109/TAFFC.2017.2761750

[7]. Soni, P., & Singh, R. (2021). Artificial intelligence-based cyberbullying detection: A review of techniques and challenges. *International Journal of Information Management Data Insights, 1*(2), 100022. https://doi.org/10.1016/j.jjimei.2021.100022

[8]. Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). https://doi.org/10.18653/v1/N16-2013

[9]. Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656–666). https://aclanthology.org/N12-1081

