Trolling Detection System Using Natural Language Processing (NLP)

Harsh Sahu[1]; Gaurav Narkhede[2]; Bhanupratap Gangboir[3]; Ayush Mendke[4]

G H Raisoni College of Engineering Nagpur, India

Publication Date: 2025/05/13

Abstract; **The rise of social media has led to an increase in online trolling, which negatively impacts users' mental health and disrupts digital communities. Detecting and mitigating trolling behavior is a significant challenge due to the evolving nature of language, sarcasm, and contextual variations. This research explores the application of Natural Language Processing (NLP) in developing an automated trolling detection system. By leveraging sentiment analysis, text classification, and deep learning techniques, NLP-based models can identify trolling content with high accuracy. This paper examines various approaches, challenges, and future prospects in NLP-based trolling detection systems.**

How to Cite: Harsh Sahu; Gaurav Narkhede; Bhanupratap Gangboir; Ayush Mendke (2025). Trolling Detection System Using Natural Language Processing (NLP). *International Journal of Innovative Science and Research Technology*, 10(4), 3408-3412. https://doi.org/10.38124/ijisrt/25apr2049

I. INTRODUCTION

Trolling, a form of online harassment, involves posting inflammatory, offensive, or disruptive comments to provoke reactions or create discord. As social media platforms struggle with moderating such content, NLP emerges as a powerful tool for automated trolling detection. Traditional keyword-based moderation is ineffective due to sarcasm, context, and evolving slang. This research focuses on how NLP techniques, including sentiment analysis, transformer models, and linguistic pattern recognition, can be used to detect and mitigate trolling behavior in online conversations.

II. APPLICATIONS OF NLP IN TROLLING DETECTION

NLP enables automated systems to detect trollingbehavior in various online environments:

➢ *Sentiment Analysis:*

Identifies aggressive, hateful, or toxic speech patterns in comments and posts.

➢ *Sarcasm Detection:*

Uses deep learning models to differentiate between sarcasm and genuine negative sentiment.

➢ *Text Classification:*

Categorizes comments as trolling, offensive, or neutral using supervised and unsupervised machine learning models.

➢ *Contextual Understanding:*

NLP models analyze conversations to detect subtle forms of trolling that rely on context.

➢ *Real-Time Moderation:*

Implements NLP-based filters for real-time detection and removal of trolling content.

➢ *User Behavior Analysis:*

Detects repeated trolling patterns by analyzing posting history and engagement metrics.

III. NLP TECHNIQUES FOR TROLLING DETECTION

Various NLP approaches are used to detect trolling behavior:

➢ *Lexicon-Based Analysis:*

Uses predefined dictionaries of offensive words and phrases to identify trolling comments.

➢ *Machine Learning Models:*

Uses classifiers such as Support Vector Machines (SVM), Random Forest, and Naïve Bayes to detect trolling patterns.

➢ *Deep Learning Approaches:*

Implements transformer-based models like BERT, GPT, and LSTM networks to understand complex linguistic patterns.

➢ *Embedding Techniques:*

Utilizes word embeddings (Word2Vec, GloVe, FastText) to capture semantic relationships in text.

➢ *Named Entity Recognition (NER):*

Identifies user mentions, targets of trolling, and contextual references to detect abusive language.

➢ *Emotion and Tone Analysis:*

Detects emotions such as anger, frustration, and sarcasm in textual conversations.

IV. CHALLENGES IN TROLLING DETECTION

Despite advances in NLP, trolling detection faces multiple challenges:

➢ Evolving Language & Slang:

Trolls constantly create new slang and coded language to evade detection.

➢ Contextual Ambiguity:

Some comments appear neutral or humorous but may have an underlying offensive intent.

➢ *Sarcasm & Irony:*

Standard sentiment analysis struggles to differentiate sarcasm from genuine expressions.

➢ *Multilingual Trolling:*

Trolling occurs in multiple languages, requiring robust multilingual NLP models.

➢ *False Positives & Negatives:*

Overly strict detection models may flag non-trolling comments, while lenient models may miss actual trolling.

➢ *Privacy & Ethical Concerns:*

Automated moderation must balance censorship and free speech rights.

V. PROPOSED MODEL WORK-FLOW

➢ *Brief Description of Drawing:*

- **Start** – The system begins operation when a user submits a text comment or message on a platform.
- **Collect User Input (Text)** – The system captures user-generated text from a social media post, chat, or forum comment.
- **Preprocessing** – The text is cleaned and prepared for analysis.

This step includes:

✓ Tokenization – Breaking text into words or phrases.
✓ Stopword Removal – Removing common words (e.g., "the," "is," "and") that do not add significant meaning.
✓ Lemmatization/Stemming – Converting words to their root forms (e.g., "running" → "run").

- **Feature Extraction –** The system converts text into numerical features using techniques such as:

✓ TF-IDF (Term Frequency-Inverse Document Frequency) – Assigns importance to words based on their frequency in the document.
✓ Word Embeddings (Word2Vec, GloVe, BERT) – Converts words into dense vector representations to capture semantic meaning

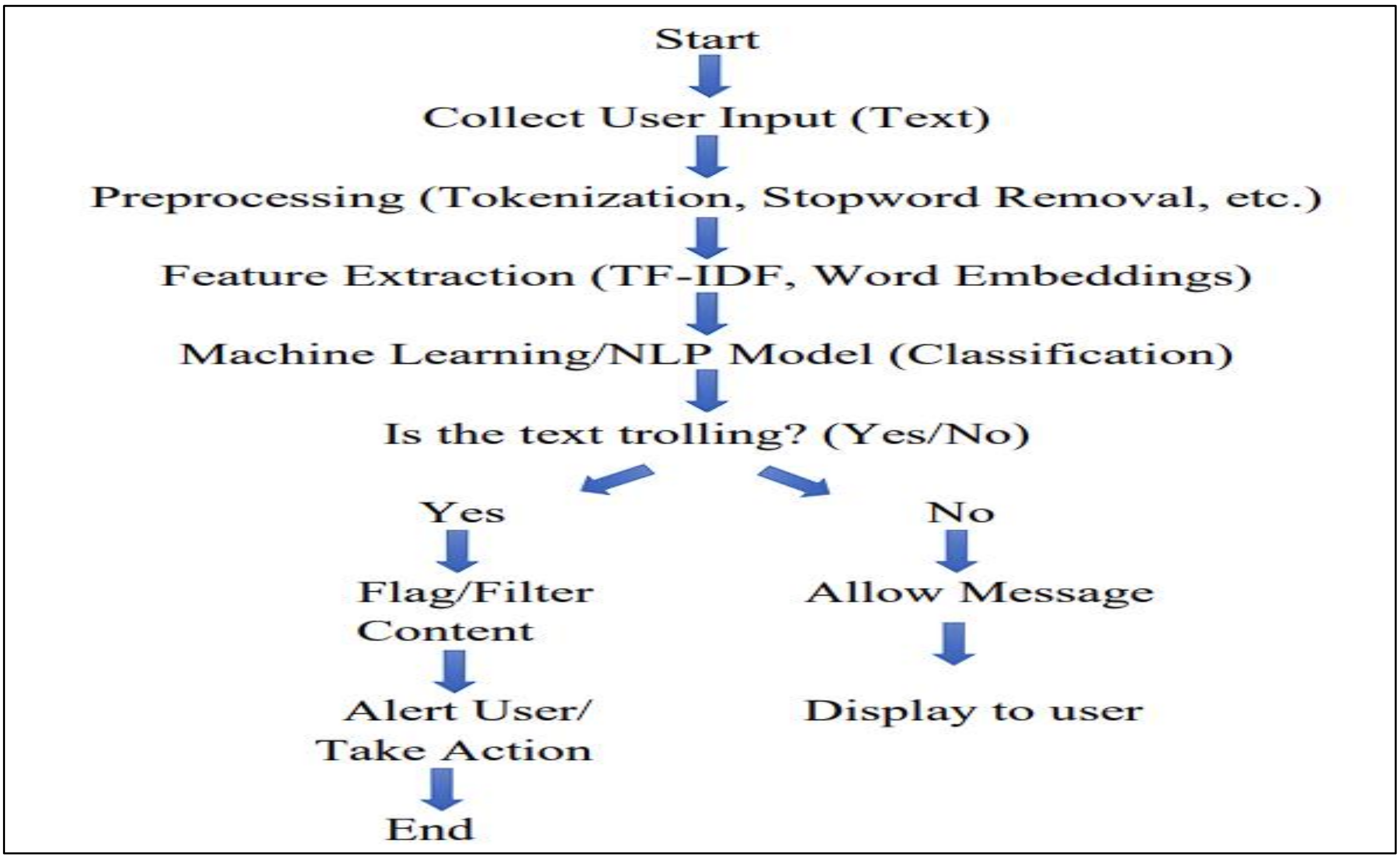

Start
Collect User Input (Text)
Preprocessing (Tokenization, Stopword Removal, etc.)
Feature Extraction (TF-IDF, Word Embeddings)
Machine Learning/NLP Model (Classification)
Is the text trolling? (Yes/No)
Yes
No
Flag/Filter Content
Allow Message
Alert User/ Take Action
Display to user
End

Fig 1 Proposed Model Work-Flow

VI. APPLICATION OF THE SYSTEM

- ➢ Social Media Platforms: Twitter, Facebook, Instagram, Reddit – to monitor toxic comments.
- ➢ Online Forums & Communities: Detect trolls in discussion boards.
- ➢ Gaming & Live Streaming Platforms: YouTube, Twitch, Discord – filter harmful live chat messages.
- ➢ Educational & Workplace Communication Apps: Prevent cyber harassment in digital classrooms and professional workspaces.
- ➢ Customer Support & Review Sections: Identify spam/trolling in customer feedback.

VII. FUTURE PROSPECTS AND RECOMMENDATIONS

➢ *To enhance the effectiveness of NLP-Based Trolling Detection, Future Research Should Focus On:*

- Hybrid AI Models: Combining NLP with behavioral analytics to improve accuracy.
- Explainable AI (XAI): Enhancing model transparency to reduce false positives.
- Cross-Platform Analysis: Developing unified trolling detection across multiple social media platforms.
- Multilingual NLP Models: Expanding detection capabilities to include diverse languages.
- Real-Time Adaptation: Implementing self-learning AI models that adapt to new trolling trends.
- User-Feedback Integration: Using crowdsourced reports to refine detection algorithms.

VIII. RESULT AND OUTPUT

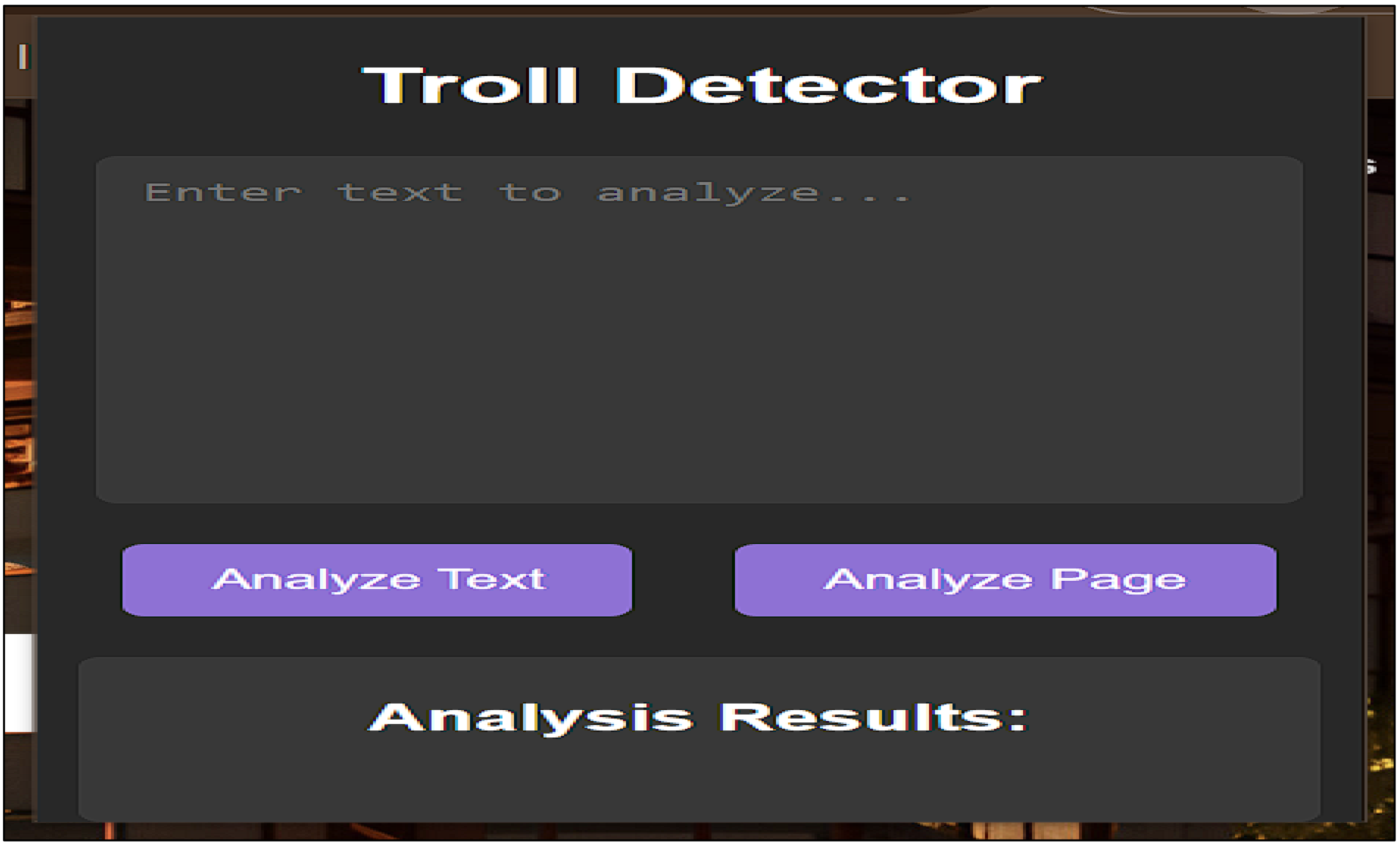


Fig 2 Troll Detector UI

The figure titled **"Troll Detector"** presents the user interface of a sentiment and intent classification tool developed to identify potential trolling behavior in textual input. This tool is a part of the broader system designed to enhance online discourse by automatically analyzing content and flagging text that could be malicious, provocative, or disruptive.

➢ *Interface Overview*

The graphical interface is minimalistic and user-friendly, composed of four primary sections:

- *Input Text Area:*

- ✓ Located at the center top of the interface, this area is labeled with the placeholder *"Enter text to analyze..."*.
- ✓ It serves as the main input field where users can manually enter or paste a piece of text suspected of containing troll-like behavior.
- ✓ The dark background with faint placeholder text ensures readability and encourages input without visual strain.

- *Control Buttons:*

✓ Beneath the input area are two buttons: **"Analyze Text"** and **"Analyze Page"**, both styled in a consistent purple color, signifying action triggers.
✓ The **"Analyze Text"** button initiates the analysis of the content directly entered into the input box.
✓ The **"Analyze Page"** button likely extends the analysis to an entire webpage's content (possibly using DOM parsing or content scraping in implementation).

- *Analysis Output Section:*

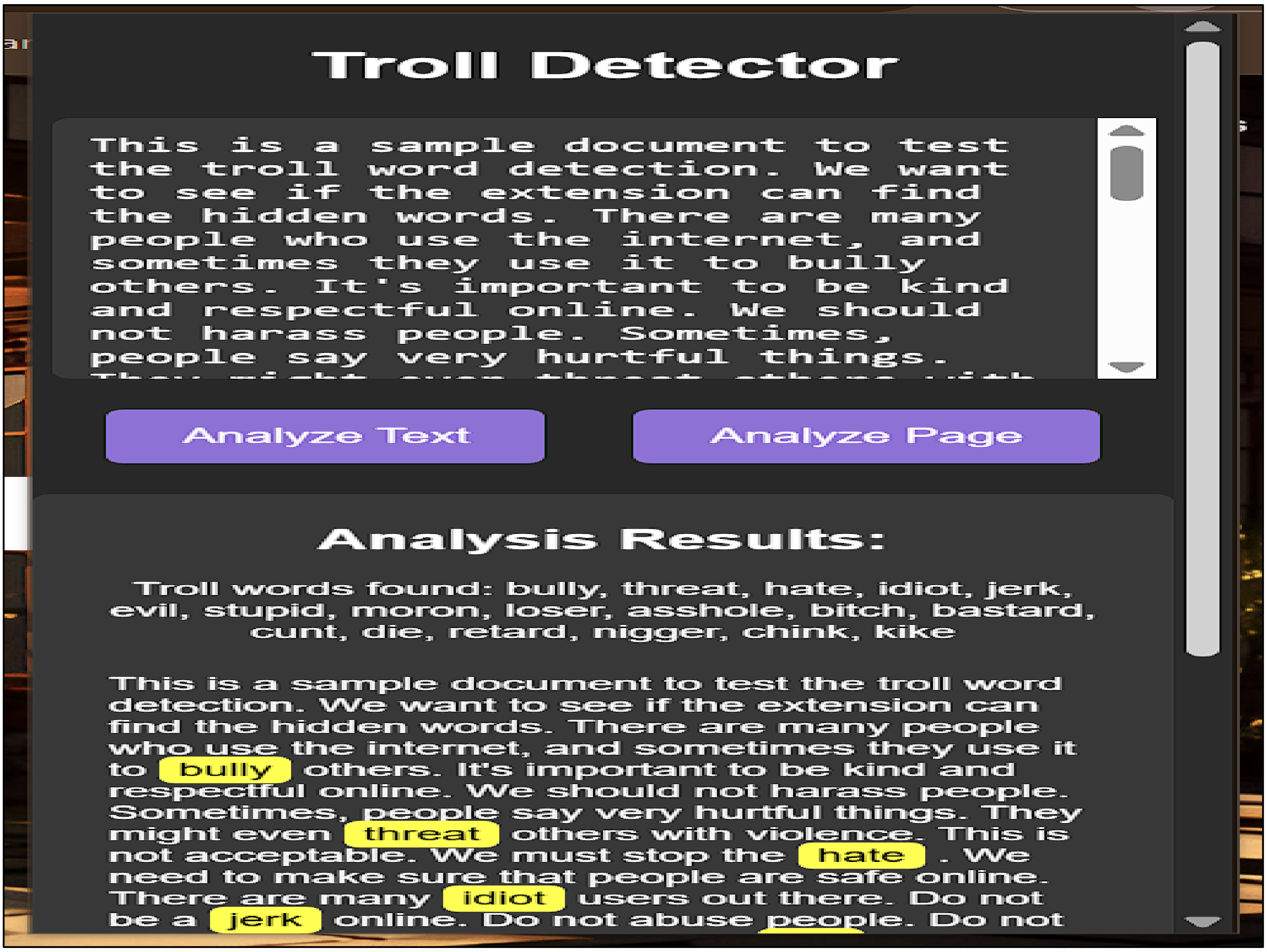
✓ Located at the bottom of the interface is the **"Analysis Results:"** panel.
✓ This section displays the output after the analysis has been performed. The results might include classifications such as *"Troll Detected"*, *"Neutral"*, or *"Non-Troll"*, depending on the model's assessment of the language used.
✓ Although no output is shown in the figure, this area is clearly designated to present the conclusions drawn by the system's backend model.

➢ *Functionality and Application*

This tool can be used in real-time by moderators, forum administrators, or users to:

- Pre-screen comments or posts before publishing.
- Analyze existing posts across a page using the "Analyze Page" feature.
- Contribute to a safer and more respectful online community by reducing the spread of inflammatory or disruptive comments.

The tool is potentially powered by machine learning or NLP models trained on annotated datasets that distinguish troll content from regular user interaction. Algorithms such as sentiment analysis, keyword spotting, and contextual understanding could be part of the underlying engine.



This is a sample document to test the troll word detection. We want to see if the extension can find the hidden words. There are many people who use the internet, and sometimes they use it to bully others. It's important to be kind and respectful online. We should not harass people. Sometimes, people say very hurtful things. They might even threat others with

Analysis Results:

Troll words found: bully, threat, hate, idiot, jerk, evil, stupid, moron, loser, asshole, bitch, bastard, cunt, die, retard, nigger, chink, kike

This is a sample document to test the troll word detection. We want to see if the extension can find the hidden words. There are many people who use the internet, and sometimes they use it to bully others. It's important to be kind and respectful online. We should not harass people. Sometimes, people say very hurtful things. They might even threat others with violence. This is not acceptable. We must stop the hate . We need to make sure that people are safe online. There are many idiot users out there. Do not be a jerk online. Do not abuse people. Do not

Fig 3 Troll detector working

This figure demonstrates the practical implementation of the proposed model, highlighting both the user interaction and the output feedback mechanism. It reflects the usability focus of the application and the real-time analysis capabilities. The visual clarity and intuitive layout contribute to the system's accessibility, making it suitable for deployment across a variety of online platforms.

IX. CONCLUSION

NLP provides a powerful solution for detecting and mitigating trolling behavior online. By leveraging sentiment analysis, text classification, and deep learning, NLP-based systems can improve the accuracy and efficiency of content moderation. However, challenges such as contextual ambiguity, sarcasm detection, and multilingual analysis must be addressed to enhance detection performance. With continuous advancements in NLP, AI-driven trolling detection can create safer and more inclusive online communities while balancing free speech considerations.

REFERENCES

[1]. T. K. Das, D. P. Acharjya and M. R. Patra, "Opinion mining about a product by analyzing public tweets in Twitter", Proc. Int. Conf. Comput. Commun. Informat., pp. 1-4, Jan. 2014. [2] H. Rosa et al., “Automatic cyberbullying detection: A systematic review,” Comput. Hum. Behav., vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.

[2]. B. S. Nandhini and J. I. Sheeba, “Online Social Network Bullying Detection Using Intelligence Techniques,” Procedia Comput. Sci., vol. 45, pp. 485–492, 2015, doi: 10.1016/j.procs.2015.03.085.

[3]. A. Ioannou et al., “From risk factors to detection and intervention: A metareview and practical proposal for research on cyberbullying,” in 2017 IST-Africa Week Conference (IST-Africa), Windhoek, May 2017, pp. 1–8, doi: 10.23919/ISTAFRICA.2017.8102355. Electronic copy available at: https://ssrn.com/abstract=4340372

[4]. A. A. Mazari, “Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies,” in 2013 5th International Conference on Computer Science and Information Technology, Amman, Jordan, Mar. 2013, pp. 126–133, doi: 10.1109/CSIT.2013.6588770.

[5]. Mathew, B, Dutt R, Goyal P, Mukherjee A (2018) Spread of hate speech in online social media. In: Proceedings of the 10th ACM Conference on web science, pp 173–182, 2019.

[6]. A Sarkar, "MACHINE LEARNING TECHNIQUES FOR RECOGNIZING THE LOAN ELIGIBILITY", International Research Journal of Modernization in Engineering Technology and Science, Vol.3, Iss:12, December 2021.

[7]. Jurafsky D, Martin J H (2002) Speech and Language Processing - An Intro to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education Asia, ISBN 81-7808-594-1

[8]. Mueller E (1998) Natural Language Processing with Thought-Treasure. Erik T. Mueller, New York

[9]. Rahm E. & Hai Do Hong. 2000. Data Cleaning: Problems and current approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 2000.

[10]. Grefenstette, G. (1999). Tokenization. In: van Halteren, H. (eds) Syntactic Wordclass Tagging. Text, Speech and Language Technology, vol 9. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9273-4_9

[11]. R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1-5, doi: 10.1109/ICCIC.2014.7238427.