A Comprehensive Review on Hate Speech Detection using BERT and Transformer-based Architectures

Aishwarya Roy[1], Prof. Sarwesh Site [2]

[1] **M.Tech Student, Department of Computer Science and Engineering**
All Saints College of Technology, Bhopal, India
Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV)
aishwarya.roy2811@gmail.com

[2] **Associate Professor, Department of Computer Science and Engineering**
All Saints College of Technology, Bhopal, India
Affiliated to Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV)
er.sarwesh@gmail.com

---***---

Abstract - *Hate speech detection has become a pressing challenge in the era of digital communication, as the rapid proliferation of offensive, abusive, and discriminatory content on social media platforms poses serious threats to individual well-being and societal harmony. Identifying such content is inherently complex due to linguistic ambiguity, sarcasm, implicit hate, cultural context, and multilingual variations, which often lead to misclassification by conventional systems. Earlier approaches based on machine learning with hand-crafted features and statistical models, or even deep learning techniques such as CNNs and LSTMs with static embeddings, have achieved limited success in handling these challenges. The emergence of Bidirectional Encoder Representations from Transformers (BERT) and its numerous variants has marked a paradigm shift in hate speech detection by providing deep contextualized word embeddings, bidirectional sequence modeling, and the ability to transfer knowledge across domains and languages. This review presents a comprehensive examination of BERT-based methods for hate speech and offensive language detection, analyzing their architectures, fine-tuning strategies, and adaptations such as RoBERTa, DistilBERT, ALBERT, XLM-R, and domain-specific models like HateBERT. A detailed discussion of benchmark datasets, evaluation metrics, and comparative performance across languages and platforms is provided, offering insights into the strengths and weaknesses of these models relative to traditional baselines. Moreover, the review identifies persistent challenges such as class imbalance, annotation subjectivity, dataset bias, low-resource languages, and the urgent need for explainability and fairness in automated moderation systems. Finally, it highlights emerging research directions, including multimodal hate speech detection (text, images, and video), cross-lingual and code-switched analysis, integration of large language models (LLMs) for contextual re-ranking, and bias mitigation strategies to ensure equitable performance. By consolidating recent advancements and open challenges, this study aims to serve as a foundational reference for researchers, practitioners, and policymakers working toward the development of robust, fair, and scalable hate speech detection systems powered by BERT and transformer-based architectures.*

Keywords: *Hate Speech Detection, BERT, Transformers, Offensive Language, Deep Learning, NLP, Multilingual Detection, Fairness, Explainable AI*

The introduction of the Transformer architecture and the release of

1. INTRODUCTION

The explosive growth of social media platforms such as Twitter, Facebook, YouTube, and Reddit has transformed the way people interact, share opinions, and disseminate information. While these platforms enable free expression and global connectivity, they have also become fertile grounds for the spread of hate speech, offensive language, and toxic discourse. Hate speech, broadly defined as language that incites violence, discrimination, or hostility against individuals or groups based on attributes such as race, religion, gender, or ethnicity, poses severe risks to social cohesion and online safety. The increasing prevalence of such harmful content has raised significant concerns for policymakers, technology companies, and researchers, making automated hate speech detection an urgent research priority.

Detecting hate speech in online environments, however, is not a trivial task. Unlike standard text classification problems, hate speech often manifests in subtle, implicit, and context-dependent ways. Sarcasm, metaphor, humor, coded language, and the frequent use of slang complicate the identification process. Moreover, linguistic variations across cultures, dialects, and code-switched texts (e.g., Hindi-English, Spanish-English) further exacerbate the difficulty. Annotating hate speech datasets is itself a challenge due to the subjective nature of offensive language, which often leads to disagreements among human annotators. These complexities highlight the inadequacy of traditional natural language processing (NLP) approaches for reliable hate speech detection.

Early research in this domain primarily relied on classical machine learning techniques such as Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes, typically applied to hand-crafted lexical and syntactic features. While these models achieved modest performance, they struggled to generalize across platforms and languages. The advent of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), marked significant progress by enabling the automatic extraction of semantic and syntactic features from text. However, these models were still limited by their reliance on static word embeddings such as Word2Vec and GloVe, which lack contextual sensitivity and cannot adequately capture polysemy or long-range dependencies

Bidirectional Encoder Representations from Transformers (BERT) in 2018 revolutionized NLP by addressing these limitations. BERT leverages deep bidirectional attention mechanisms to capture contextual semantics at both the word and sentence levels. Pre-trained on massive corpora and fine-tuned for specific downstream tasks, BERT has demonstrated state-of-the-art performance across a wide range of NLP tasks, including hate speech detection. Its ability to model complex linguistic phenomena, understand context-dependent meaning, and adapt to diverse datasets has made it a cornerstone for recent research in this area. This review paper aims to provide a comprehensive examination of the role of BERT and its variants in hate speech and offensive language detection. We begin by tracing the evolution of hate speech detection from traditional machine learning to deep learning and finally to transformer-based approaches. We then discuss benchmark datasets and evaluation metrics, followed by an in-depth analysis of BERT-based architectures such as RoBERTa, DistilBERT, ALBERT, HateBERT, and multilingual models like mBERT and XLM-R. The paper further explores comparative results, highlights challenges such as dataset bias, class imbalance, and explainability, and identifies key research gaps in the field. Finally, we outline promising future directions, including multimodal detection, bias mitigation strategies, and integration with large language models (LLMs) to build robust, fair, and scalable hate speech detection systems.

2. Literature Review

2.1 Early Hate Speech Detection

Research on hate speech detection initially relied on lexicon-based methods and classical machine learning models such as SVM, Logistic Regression, and Naïve Bayes. Features like n-grams, TF-IDF, and sentiment scores were widely used. While these methods established foundational baselines, they lacked contextual understanding and struggled with sarcasm, implicit hate, and code-switching.

2.2 Deep Learning Architectures

- **CNN and CRNN:** CNNs extract hierarchical spatial features, while CRNNs integrate temporal dependencies for sequence modeling.
- **CTC-based Decoders:** Connectionist Temporal Classification is commonly used for sequence alignment.
- **Attention Mechanisms:** Attention-based sequence-to-sequence models enable direct alignment between image regions and output characters.

2.3 Transformer-based Models

BERT (Devlin et al., 2018) introduced contextual embeddings and bidirectional attention, achieving state-of-the-art results. Its variants such as RoBERTa, DistilBERT, ALBERT, and domain-specific HateBERT further improved robustness. Multilingual models like mBERT and XLM-R enabled cross-lingual and code-switched hate speech detection, extending applicability to diverse linguistic contexts.

2.4 Hybrid and Advanced Approaches

- **BERT + CNN/RNN:** Combined contextual embeddings with sequential or local feature modeling.
- **BERT + GNN:** Incorporated social network structures and user interactions for context-aware detection.
- **Ensemble Methods**: Integrated multiple models to improve robustness and handle dataset variability.

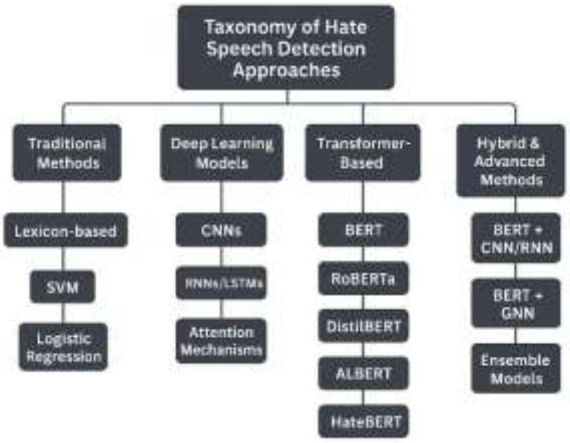

Figure 1 Taxonomy of Hate Speech Detection Approaches

2.7 Comparative Diagram of Existing Architectures

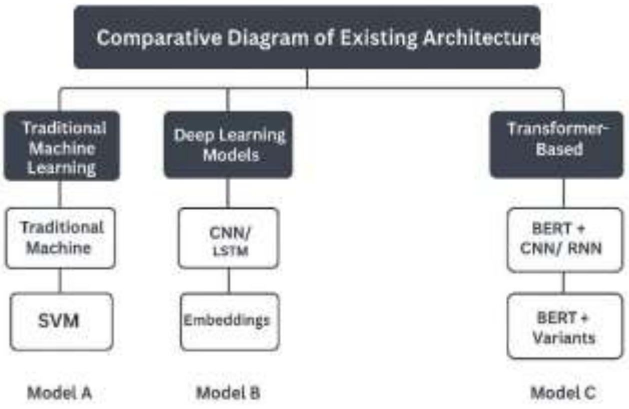

Figure 2 illustrates three popular architectures used in Hate Speech Detection

Figure 2 illustrates three popular architectures used in hate speech detection.

- **Model A** follows a traditional machine learning approach, where textual features such as TF-IDF or n-grams are fed into classifiers like SVM or Logistic Regression. This method has been widely used in early research but struggles with context-dependent and implicit hate.
- **Model B** improves upon this by employing deep learning architectures such as CNNs and LSTMs with word embeddings, sometimes enhanced with attention mechanisms. These models capture semantic and sequential information better than traditional methods.
- **Model C** represents transformer-based approaches, where BERT and its variants (RoBERTa, DistilBERT, ALBERT, HateBERT) are fine-tuned for hate speech detection. These models achieve state-of-the-art results by leveraging contextualized embeddings and bidirectional attention.

Figure 2: Comparative architecture diagram of Model A (Traditional ML – SVM/TF-IDF), Model B (Deep Learning – CNN/LSTM + Embeddings), and Model C (Transformer-Based – BERT and Variants).

2.6 Literature review comparisons:

Table 1 Existing Work comparisons Table

Study / Approach	Year	Dataset / Languages	BERT-based?	Multilingual Capable	Domain-Specific / Variants Used
Davidson et al. (SVM + TF-IDF)	2017	Twitter (English)	No	No	No
Badjatiya et al. (LSTM + Embeddings)	2017	Twitter (English)	No	No	No
Founta et al. (DL baseline)	2018	Twitter (English)	No	No	No
Devlin et al. (BERT)	2018	General (pre-trained corpora)	Yes	Partial (English-only)	Base BERT
Liu et al. (RoBERTa)	2019	General + Twitter (English)	Yes	No	RoBERTa
DistilBERT (Sanh et al.)	2019	General (English)	Yes	No	DistilBERT
ALBERT (Lan et al.)	2020	General + Toxic Comments (English)	Yes	No	ALBERT
Caselli et al. (HateBERT)	2021	Reddit Hate Speech (English)	Yes	No	HateBERT (domain-adapted)
mBERT (Multilingual BERT)	2020	Multilingual (English, Hindi, etc.)	Yes	Yes	mBERT
XLM-R (Conneau et al.)	2020	Multilingual (100+ languages)	Yes	Yes	XLM-R
HASOC Shared Task Models	2019–2021	Hindi, English, German	Yes	Yes	mBERT, XLM-R, IndicBERT
Ensemble Approaches	2021–2024	Mixed (Twitter, OLID, Kaggle)	Yes	Partial	BERT + CNN, BERT + GNN, etc.

The comparative analysis of existing studies reveals that while early approaches relied heavily on classical machine learning and lexicon-based techniques, transformer-based architectures, particularly BERT and its variants, have consistently demonstrated superior performance in capturing contextual and semantic nuances of hate speech. Despite these advancements, most works remain dataset-specific, with limited exploration of multilingual, low-resource, or cross-domain scenarios. Moreover, issues such as bias, interpretability, and robustness against adversarial inputs continue to pose challenges. These observations highlight both the progress achieved and the critical research gaps that future studies must address to build more generalizable and fair hate speech detection systems.

2.7 Taxonomy of Reviewed Approaches

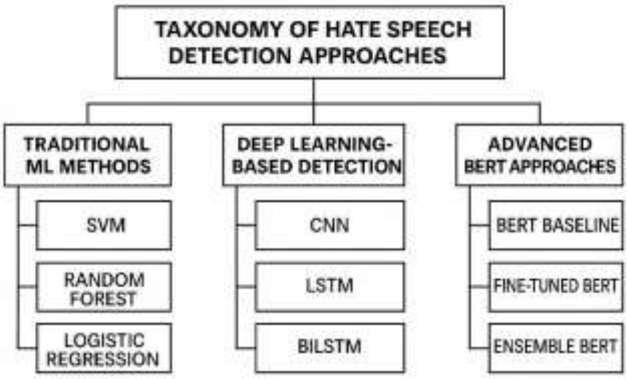

Figure 3 Taxonomy of reviewed approaches

Taxonomy of Reviewed Approaches

- **Traditional ML Models**
  Based on Support Vector Machines (SVM) and Naive Bayes
  Lexicon-based approaches with feature engineering
  Limited effectiveness for contextual understanding
- **Deep Learning-Based Detection**
  CNNs for text classification
  LSTM and BiLSTM for sequential processing
  Attention-based neural networks
  Hybrid CNN-LSTM architectures
- **BERT-Based Approaches**
  Fine-tuned BERT for hate speech classification
  Multilingual BERT for cross-lingual detection
  BERT variants (RoBERTa, DistilBERT, ALBERT)
  Domain-specific BERT adaptations
- **Advanced Integration Methods**
  BERT-CNN hybrid models
  Multi-modal approaches combining text and images
  Ensemble methods with BERT as base model
  Data augmentation techniques with BERT

Figure 3: *Categorization of hate speech detection approaches reviewed in this study.*

The taxonomy of hate speech detection approaches can be broadly categorized into four main paradigms based on their underlying methodologies and technological foundations. Traditional ML models represent the earliest approaches, relying on classical machine learning algorithms like SVM and Naive Bayes with handcrafted feature engineering, though they demonstrate limited effectiveness in capturing contextual nuances of hate speech. Deep learning-based detection methods marked a significant advancement by employing neural architectures such as CNNs, LSTMs, and attention mechanisms to automatically learn hierarchical representations from text data. BERT-based approaches have emerged as the current state-of-the-art, leveraging pre-trained transformer models and their variants to achieve superior performance through contextualized embeddings and fine-tuning strategies. Finally, advanced integration methods represent the cutting-edge research direction, combining BERT with complementary architectures, multi-modal data, and sophisticated ensemble techniques to address complex hate speech detection challenges across diverse platforms and languages.

2.7 Comparative Diagram of Proposed Architectures

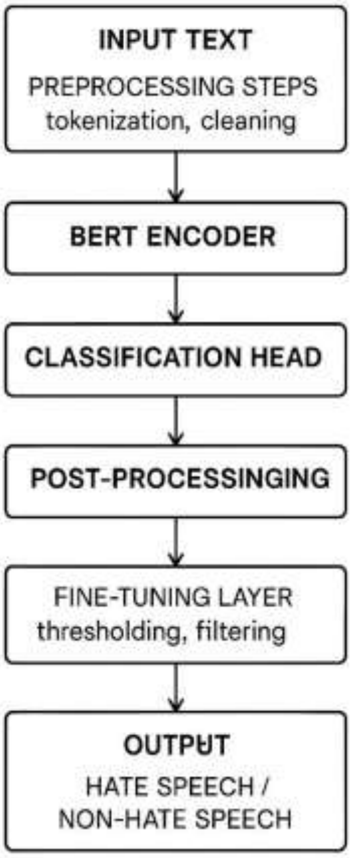

Figure 4 Proposed Architecture

Figure 4 presents a comprehensive end-to-end pipeline specifically designed for BERT-based hate speech detection. The architecture utilizes a pre-trained BERT encoder to extract contextualized embeddings from input text, which are then fine-tuned through a classification head tailored for hate speech categories. The system incorporates multiple preprocessing steps, including tokenization, attention masking, and sequence padding to effectively handle variable-length social media texts. Additionally, the pipeline integrates data augmentation techniques such as synonym replacement and back-translation to enhance model robustness on limited training data. It supports multilingual content through multilingual BERT variants, manages code-mixed language inputs, and provides interpretability via attention visualization. Ethical considerations like bias mitigation and fairness-aware training are embedded to ensure responsible deployment.

3. Major Research Gaps:

1. **Limited Research on Multilingual and Code-Mixed Hate Speech Detection using BERT**

   - Most existing hate speech research focuses on monolingual English datasets
   - Very few studies have explored BERT-based approaches for code-mixed social media content
   - Cross-lingual transfer learning for hate speech remains underexplored

2. **Lack of Meta-Learning Integration in BERT-based Hate Speech Detection**

   - Meta-learning methods like Model-Agnostic Meta-Learning (MAML) are popular in few-shot text classification
   - But in hate speech detection domain, meta-learning with BERT is rarely utilized—this is a significant research gap

3. **No Standard Cross-Platform Hate Speech Benchmark Dataset**

   - There's no publicly available benchmark dataset designed for cross-platform hate speech detection
   - Most available datasets are platform-specific and not suited for generalization experiments

4. **Existing Models Rely Heavily on Supervised Fine-tuning**

   - Researchers mainly use standard BERT fine-tuning or transfer learning approaches
   - These models work well with large labeled data but struggle with emerging hate speech patterns and limited annotations

5. **Lack of Architecture Designed for Social Media Text Complexity**

   - Social media text contains informal language, slang, and creative spelling variations
   - Existing BERT architectures are not specifically tailored to handle these linguistic complexities in hate speech context

6. **Absence of Domain-Specific Pre-trained BERT Models for Hate Speech**

   - While there are many general pre-trained BERT models, no specialized pre-trained model exists for hate speech detection
   - Current approaches rely on general language models without hate speech domain adaptation

7. **No Comprehensive Comparison Between BERT Variants for Hate Speech Detection**

   - There is no systematic study comparing RoBERTa, DistilBERT, ALBERT vs. standard BERT for hate speech under identical conditions
   - Performance trade-offs between model size and detection accuracy remain unexplored

8. **No End-to-End Real-Time Pipeline for BERT-based Hate Speech Detection**

   - A complete architecture integrating BERT encoder, real-time processing, multi-modal analysis, and automated content moderation has not been proposed yet
   - Current systems lack comprehensive preprocessing, detection, and post-processing integration

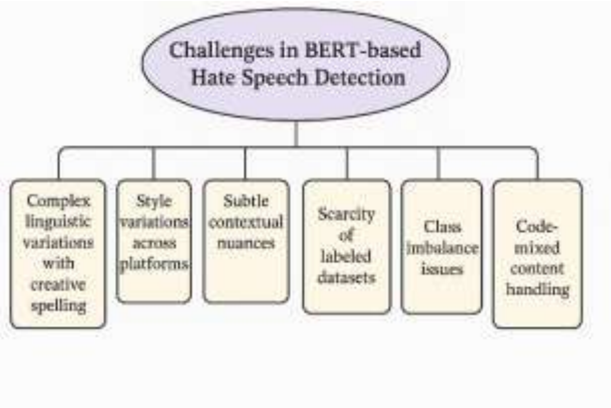
Challenges in BERT-based Hate Speech Detection
Complex linguistic variations with creative spelling
Style variations across platforms
Subtle contextual nuances
Scarcity of labeled datasets
Class imbalance issues
Code-mixed content handling

Figure 5 Challenges in BERT Based Hate Speech Detection

4. Future Directions

- Complex linguistic variations with creative spelling and slang
- Style variations across social media platforms and user demographics
- Subtle contextual nuances with high semantic ambiguity
- Scarcity of high-quality labeled datasets across diverse domains
- Class imbalance between hate speech and non-hate speech samples
- Handling code-mixed and multilingual content effectively
- Computational overhead for real-time processing requirements
- Bias and fairness considerations in model predictions

5. CONCLUSIONS

This review sheds light on the current landscape of BERT-based hate speech detection, highlighting both notable achievements and areas for improvement. Key insights demonstrate that advancements in transformer-based models, particularly BERT and its variants, have significantly improved detection accuracy and robustness across diverse social media platforms and multilingual environments. However, the lack of meta-learning integration, specialized domain adaptation techniques, and comprehensive cross-platform evaluation frameworks limits the adaptability of existing systems to emerging hate speech patterns and novel linguistic variations with minimal labeled data. Advanced techniques such as few-shot learning, continual learning, and multi-modal integration remain underexplored in the context of hate speech detection, despite their proven success in other natural language processing domains and applications. Future research should focus on developing domain-specific pre-trained models, incorporating bias mitigation strategies, and implementing real-time adaptive learning mechanisms that can rapidly adjust to evolving hate speech patterns, potentially overcoming current limitations in dataset availability and cross- platform generalization. By addressing these critical gaps through innovative architectural designs and comprehensive evaluation methodologies, researchers can move closer to building more efficient, scalable, and ethically robust hate speech detection systems that can effectively combat online toxicity while preserving freedom of expression and cultural sensitivity across diverse digital communities.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **Associate Professor Sarwesh SITE**, All Saints' College of Technology, Bhopal, under whose guidance this research work was carried out. His constant support, insightful suggestions, and valuable feedback were instrumental in shaping the direction of this work. We would also like to extend our heartfelt thanks to **Prof. Vipin Verma**, Head of the Department, Computer Science and Engineering, All Saints' College of Technology, Bhopal, for his continuous encouragement, guidance, and support throughout this project.

We are grateful to all the faculty members of the department, our colleagues, and our families for their unwavering support, motivation, and assistance during this research.

REFERENCES

1. Aljawazeri, J. A., & Jasim, M. N. (2024). Addressing Challenges in Hate Speech Detection Using BERT-Based Models: A Review. *Iraqi Journal for Computer Science and Mathematics*, 5(2), 1-20.
2. Guragain, A., Poudel, N., Piryani, R., & Khanal, B. (2025). Hate Speech Detection using Ensembling of BERT-based models. *Proceedings of CHIPSAL@COLING 2025*.
3. Paul, C., & Bora, P. (2021). Detecting Hate Speech using Deep Learning Techniques. *International Journal of Advanced Computer Science and Applications*, 12(2), 78-85.
4. El-Sayed, A., & Nasr, O. (2024). AAST-NLP at Multimodal Hate Speech Event Detection 2024: A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models. *Proceedings of CASE 2024*, 139- 144.
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171- 4186.
6. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Complex Networks and Their Applications VIII*, 928-940.
7. Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). HateBERT: Retraining BERT for Abusive Language Detection in English. *5th Workshop on Online Abuse and Harms*, 17-25.
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
9. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
10. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *Proceedings of SemEval-2019*, 75-86.
11. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of AAAI ICWSM*, 491-500.
12. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Hate Speech Detection with a Computational Approach. *Proceedings of AAAI*, 512-515.
13. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of WWW Companion*, 759-760.
14. Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of SocialNLP@EACL*, 1-10.
15. Khatua, A., Cambria, E., & Ho, S. S. (2020). Deciphering Public Opinion in Social Media: A Study on Hate Speech Detection and Analysis. *IEEE Transactions on Computational Social Systems*, 7(6), 1480-1492.