氏　　名	TANJIM MAHMUD
授 与 学 位	博士（工学）
学 位 記 番 号	博甲第 222 号
学位授与年月日	令和 7 年 3 月 21 日
学位授与の要件	学位規則第 4 条第 1 項
学 位 論 文 題 目	Multilingual Cyberbullying Detection in Low-Resource Languages and Dialects （少資源言語及び方言における多言語有害情報検出）
論 文 審 査 委 員	主査　准教授　ﾌﾟﾀｼﾝｽｷ ﾐﾊｳ ｴﾄﾞﾑﾝﾄﾞ 教　授　桝井　文人 教　授　升井　洋志 教　授　前田　康成 教　授　奥村　貴史

学位論文内容の要旨

Cyberbullying detection in low-resource languages and dialects is a critical challenge, particularly as the prevalence of online abuse continues to rise globally. This thesis addresses this issue with a focus on Bangla and its dialects, including Chittagonian, by introducing novel methodologies for automatic cyberbullying detection through linguistic similarity, cross-lingual transfer learning, and machine learning approaches. The research aims to bridge the gap in existing literature, which predominantly focuses on high-resource languages such as English, and presents solutions for effective abusive language identification in low-resource contexts.

Firstly, I present a newly annotated Bangla cyberbullying dataset, consisting of abusive and non-abusive remarks, and demonstrate the effectiveness of machine learning classifiers, with logistic regression achieving 97% accuracy in detecting abusive language.

I also expand the scope of the research by conducting a systematic survey of over seventy studies on cyberbullying detection in low-resource languages from 2017 to 2023, identifying key gaps such as unreliable definitions of cyberbullying and biases in dataset annotation. As a response to these challenges, I release a new dataset in the Chittagonian dialect of Bangla, contributing valuable insights to this underrepresented area.

Secondly, I explore the methods for detecting offensive language in Chittagonian, specifically focusing on keyword matching and machine learning techniques. While keyword matching serves as a baseline, I demonstrate that machine learning and deep learning approaches, particularly for recognizing linguistic patterns and contextual variations, provide superior performance. Furthermore, I introduce an automatic method for extracting vulgar words from linguistic data, which adapts to evolving linguistic trends, achieving near-human performance.

Thirdly, I compare traditional machine learning models with deep learning architectures for cyberbullying detection in both Bangla and Chittagonian. The results show that deep learning models, particularly CNNs and hybrid networks like BiLSTM+GRU, outperform traditional methods such as SVMs and ensemble techniques. I also demonstrate the effectiveness of transformer-based models, such as mBERT and Bangla BERT, achieving up to 84% accuracy in detecting cyberbullying.

Fourthly, I tackle the challenge of detecting sarcasm, an often-overlooked aspect of cyberbullying detection. By integrating sarcasm detection into my framework, I utilize machine learning models alongside explainable AI techniques like LIME (Local Interpretable Model-agnostic Explanations), improving the model's ability to correctly classify sarcastic remarks as abusive and significantly reducing false positives.

Fifthly, I delve into zero-shot cross-lingual transfer learning, addressing the challenge of training models on low-resource languages without annotated data. I investigate the selection of an optimal transfer language by utilizing multilingual transformer models, specifically mBERT and XLM-RoBERTa, trained on publicly accessible datasets from 21 languages. Through linguistic similarity scores such as eLinguistics, lang2vec, WALS, and EzGlot, I demonstrate the correlation between language similarity and classifier performance, providing a method for selecting the optimal transfer language for cyberbullying detection in low-resource scenarios.

Lastly, the contributions of this thesis provide a comprehensive framework for cyberbullying detection in low-resource languages, utilizing innovative machine learning approaches, linguistic similarity analysis, and cross-lingual transfer learning to overcome the limitations of scarce labeled data. This research not only enhances the detection of abusive language in Bangla and Chittagonian but also offers a methodology that can be extended to other low-resource languages, providing valuable insights for future advancements in multilingual cyberbullying detection.

審査結果の要旨

この論文では、少資源言語及び方言における有害情報検出について調査する。まず、多言語での有害情報（誹謗中傷）の特徴を明確にし、有害情報検出において少資源言語に関する従来研究を徹底的にレビュー調査を行なった。次に、言語間転移学習を活用した多言語有害情報検出手法を提案している。複数の少資源言語と方言の有害情報検出用のデータセットを収集し、その内容には人手によるアノテーションを用い多様性を保ったデータセットを作成した。さらに、徹底的な実験過程を通じて、性能の高い多言語言語モデルで有害情報検出を特定している。そして、少資源言語に最適な設定を実験的に特定し、有害情報検出のためのモデルの微調整（ファインチューニング）を行った。最後に、言語間類似度定量化手法を提案し，類似度の高い言語を少資源言語の有害情報検出というタスクにおいて訓練データとして使用することで，少資源の言語でも十分な応用性で有害情報を検出できることを確認した。本研究において明らかにしたことを、研究対象言語以外の言語にも拡張することを期待でき、多言語での有害情報検出の理解を深めることにさらに貢献できると考えられる。

結論として、著者は、今後の多言語有害情報とその検出手法開発への根本的な貢献をした他、人工知能、自然言語処理分野、そして特に少資源言語処理の分野において課題となっている少資源の訓練データの場合でも多言語からのデータ拡張により高精度な検出を実現するという課題について示唆を与える新知見を得たと判断される。なお、自然言語処理、人工知能など複数の分野に跨る深淵な課題に対して貢献するところ大なるものがある。よって著者は、博士(工学)の学位を授与される資格があるものと認める。