**708/E3**

# Identifying hate content with Facebook posts in Sinhala language using emoticons and reactions-based text analysis

W.A.S.N. Perera[1*], I. Perera[1] and S. Ahangama[2]

*[1]Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa, Katubedda, Sri Lanka*
*[2]Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Katubedda, Sri Lanka*

The escalating popularity of online social networks has become an integral part of the communication medium. Facebook is the most widely used social media platform, which enables people to express their opinion widely online in Sri Lanka. Such platforms bring novel opportunities but also poses several malicious phenomena such as the propagation of hate speech online with its anonymity and mobility. Hate speech on Facebook is significantly ramping up and it is paramount to prevent the dissemination of hate online. However, automatic hate speech detection is challenging in low-resourced settings and with morphologically rich languages like Sinhala. Meanwhile, much expressive facilities are being introduced frequently in Facebook. There is a new generation of emoticons, called emojis, which have dominated the social media platforms representing the thoughts, feelings, moods and emotions of a user to facilitate more expressive text content in computer-mediated communication. The emoticons dominated the sentiment associated with the text segment. Moreover, Facebook users can react to a post with different emotional reactions: love, haha, wow, sad, care, and angry. However, the sentiment inherited from reactions towards a Facebook post has not been supported by empirical evidence. In this paper, we present an experiment of effective identification of hate content in Facebook posts in Sinhala, by incorporating the impact of sentiment expressed by emoticons and reactions to FB posts for the linguistic text classification. The corpus was constructed from 2049 Facebook posts in the Sinhala language, annotated for hate/non-hate/undecided by human annotators using a crowdsourcing platform. The results evidenced that all the classifiers (Random Forest, Support Vector Machine, and Naïve Bayes) performed well when considering emojis embedded in the linguistic text and the reactions by users to the post. Finally, the study experimentally verified that the text analysis with machine learning algorithms considering emoticons embedded in FB posts and reactions to posts significantly improves hate speech classification accuracy by 97%.

**Keywords:** Hate speech, social media platforms, emoticons, reactions

**E-mail:** sureshap@cse.mrt.ac.lk