ABSTRACT

India is a multilingual country with 23 official and 122 major languages. With the ever-increasing popularity of social media platforms such as Facebook and Twitter and messenger applications such as WhatsApp and Hike, we observe an increase in communication using various low-resource Indian languages. Apart from the low resource monolingual languages, code-mixing is also prevalent on multiple online platforms. In this thesis, we study the usage of these low-resource languages along with their mixing with the English Language. In the first-of-its-kind study, we look at the social interactions in multiple Indian Languages over 281 public WhatsApp groups discussing Indian General Elections 2019. We observe a high-rate of usage of informal text in the WhatsApp communications. Besides, we also study several challenges to process the high volume of code-mixed and informal text over social-media platforms. To improve the quality of machine translation for Hindi-English code-mixed data, we create a parallel corpus (PHINC) containing 13,738 sentence pairs. We also present a pipeline (PPGT) build on top of Google Translate that outperforms the various SOTA machine translation systems on PHINC. We leverage the high-quality Hindi-English code-mixed sentences in PHINC to build a system for the Hinglish sentiment classification problem. To address the challenge of resource scarcity of various downstream tasks for code-mixed data, we generate high-quality and semantically correct code-mixed sentences from the parallel corpus. Our experiments present an opportunity to generate and upscale code-mixed datasets for various tasks such as sentiment analysis, machine translation, etc.