World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:02, 2022

Native Language Identification with Cross-Corpus Evaluation Using Social Media Data: ’Reddit’

Authors : Yasmeen Bassas, Sandra Kuebler, Allen Riddell

Abstract : Native language identification is one of the growing subfields in natural language processing (NLP). The task of native language identification (NLI) is mainly concerned with predicting the native language of an author’s writing in a second language. In this paper, we investigate the performance of two types of features; content-based features vs. content independent features, when they are evaluated on a different corpus (using social media data “Reddit”). In this NLI task, the predefined models are trained on one corpus (TOEFL), and then the trained models are evaluated on different data using an external corpus (Reddit). Three classifiers are used in this task; the baseline, linear SVM, and logistic regression. Results show that content-based features are more accurate and robust than content independent ones when tested within the corpus and across corpus.

Keywords : NLI, NLP, content-based features, content independent features, social media corpus, ML
Conference Title : ICCLST 2022 : International Conference on Computational Linguistics and Statistical Translation
Conference Location : Jeddah, Saudi Arabia
Conference Dates : February 17-18, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:02, 2022 publications.waset.org/abstracts/142396.pdf