Insight Medical Publishing

International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering

2020

Sp. Iss. 117

Content Identification in Hindi and Bangla.

Subhabrata Banerjee

HCL Technologies, NOIDA, India

Abstract

This poster tries to focus on content identification on the two most popular languages of the Indian sub-continent, Hindi and Bangla. The emergence of substantial online content in Indian languages has given us the forensic linguistics challenge to detect the content of languages. In our effort, we could develop an online content detection system, which may identify the contents of both Hindi and Bangla. The system exploited another issue, the use of gold standard data, to overcome the crisis of data in Indian languages. The system used gold-standard data for the development of this system. It used trigram, and Named Entity(NE) data to identify contents, with the use of standard classifier. The result of both the system created from two monolingual corpora is above 90.0 in the F1-score measure. The work also gives a comparative study of the nature and distribution of trigram and NE data of the languages.

Biography:

Mr.Subhabrata Banerjee, works as a Computational Linguist in HCL Technologies, Noida. He has worked in places like Indian Institute of Science(IISc), Bangalore, Centre for Development of Advanced Computing(CDAC), Noida and Indian Institute of Technology(IIT) Kanpur in the capacity of Computational Linguist. He has around 10 years IT experience and has delivered around 10 projects.

7th World Machine Learning and Deep Learning Congress, Webinar- June 18-19, 2020

Abstract Citation:

Subhabrata Banerjee, Content Identification in Hindi and Bangla, Machine Learning 2020, 7th World Machine Learning and Deep Learning Congress, Webinar, June 18-19, 2020

https://machinelearning.conferenceseries.com/2020