Lexicon based Sentiment Analyser for the Telugu Language

Rajesh Burla, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Prof. Khurshid Ahmad

Digital media includes social media, blogs and diverse data that provides instantaneous access to information, now and can be any time in history. This data in most of the cases are unstructured, sentiment analysis has become a principal tool of making that data meaningful. In linguistics, Dravidian languages have a unique orthography, syntax, lexis, and morphology, Telugu is one among them with more than 90 million native people who write, read and speak in the language. There is a continuous exchange of data on the world-wide-web in the Telugu language that impacts products, reviews, news editorials, political concerns, and many more. Most of the analytics tools are developed with respect to the structure and functions of the English language and majorly spoken language across the world. Telugu language being second popular spoken language in India lacks in digital innovations like text analytics, sentiment analysis, etc. which can impact wide-range of markets.

This dissertation aims to design and implement a sentiment analyser for Telugu, a Dravidian language by building a SentiWordNet lexicon coupled with a benchmarked Telugu POS tagger. This system handles various morphological and syntactic rules in processing Telugu script which is very different from the existing sentiment analysis tools of English and other languages. A web application has been developed for extracting sentiment for given Telugu sentences. An experiment is conducted by collecting various e-newspaper articles with a manual annotation by three native speakers and comparing it with the results of the developed prototype. It resulted in an accuracy of 83.5%. It also examines a case study for detecting biased Telugu e-newspapers by comparing with a local governmnet official statement in both Telugu and English languages. However, results varied with different morpho-syntactic rules. Also, gives an overview of adopting this framework for other Dravidian languages like Tamil, Malayalam, and Kannada. Overall, the methods presented have limitations but show promising results for future research and expansion.