

Chapter 14

Telugu News Data Classification Using Machine Learning Approach

Bala Krishna Priya G.

Sri Padmavathi Viswa Vidyalayam, India

Jabeen Sultana

Majmaah University, Saudi Arabia

Usha Rani M.

Sri Padmavathi Viswa Vidyalayam, India

ABSTRACT

Mining Telugu news data and categorizing based on public sentiments is quite important since a lot of fake news emerged with rise of social media. Identifying whether news text is positive, negative, or neutral and later classifying the data in which areas they fall like business, editorial, entertainment, nation, and sports is included throughout this research work. This research work proposes an efficient model by adopting machine learning classifiers to perform classification on Telugu news data. The results obtained by various machine-learning models are compared, and an efficient model is found, and it is observed that the proposed model outperformed with reference to accuracy, precision, recall, and F1-score.

INTRODUCTION

Natural Language Processing-NLP is a sub-extent of Artificial Intelligence that describes communications between computers and languages of people. Recently numerous individuals accept online multimedia platforms such as blogs, online shopping review websites, feedback forums, social networking sites – Facebook, Twitter, WhatsApp, Instagram, LinkedIn and so on to mention their opinions and perspectives on a particular thing. The *Sentimental Analysis* is a significant portion of NLP and is the study of analyzing opinions, sentiments, emotions, appraisals, evaluations and attitudes of human being on specific

DOI: 10.4018/978-1-7998-7685-4.ch014

objects such as topics, products, events, firms, people, point outs, services and properties (Liu & Bing, 2012). It helps us in understanding the sentiments, in most cases the opinions. Document, Sentence, and Aspect/Feature level are three distinct levels of opinion mining can be applicable to text. These levels of analysis respectively evaluate the document-wise polarity, sentence-wise polarity in specified document and word-wise polarity of aspects in specified sentence or entire document.

Greater part of research in the field of opinion classification has been worked out in English language than the contribution of work for Indian regional languages. Indian dialects are mostly morphologically capable and agglutinative that creates job of producing specific tool for proficient language tricky and grave. Authors are concentrating on one of the territorial spoken language Telugu transcendentally in Andhra Pradesh and Telangana states and exist approximately 93 million native speakers of Telugu all over the world (“List of languages by total number of speakers”, 2020). At present majority of the sites, web journals, twitters and so forth, about news are wealthy in Telugu content. Hence there is a necessity to analyse the sentiments of news in Telugu language.

Data Mining techniques have been employed to natural language processing with some success (J.Sultana et al., 2019). Knowledge Discovery in Real time applications, for example, clinical analysis (J.Sultana et al., 2018, 2019) in business of marketing utilizing Association Rule mining (J.Sultana & G.Nagalaxmi, 2015) and system of education (Jabeen et al., 2019) require lean toward information disclosure ways to deal with comprehend the prediction algorithms. The initiation of learning machine and deep learning in the area of NLP was made arduous and troublesome assignment of preparing opinions simple and conveniently.

In this work, News in Telugu text translated into English by using *Google Translator* library available in Python. Finding sentiment score and labeling as positive or negative by using different tagging techniques. After that, attempted to classify the polarity value of Telugu news statements utilizing several Machine Learning classifiers namely Naive Bayes, Random Forest, Passive Aggressive Classifier, Perceptron and SVM (Support Vector Machines). The authors built two models for classifications: one is a binary-class and another is multi-class. In binary classification, the system classifies the sentiment as positive and negative polarities whereas in categorise(multi-class) task, the system furtherly classifies the sentiment into business, editorial, entertainment, nation and sports. Performed the results on test data through performance parameters.

Next, this paper is organised as sections as follows: Section II explains literature and related research work about NLP problems on Indian dialects. Section III explains the dataset description, translation and pre-processing of data. In Section IV, discuss the methodology by propose a frame work which includes feature selection, different classifiers and tools used, training & testing the data by using machine learning models and performance metrics. Section V discuss the results. Final Section VI conclude with future work.

LITERATURE WORK

Research analysts have exposed the attention of sentiment analysis with regards to Indian dialects, for example, Hindi, Malayalam, Telugu, Odia, Marathi, and so on. In Malayalam, the corpus is gathered from the Malayalam sites to do the sentiment classification. But, the most important problematic with corpus is user’s feedback containing spelling mistakes which will enormously influence the accuracy of the analysis. A rule-based approach is proposed by Deepu S. Nair and Co. (Deepu et al., n.d.) on resulting

12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/telugu-news-data-classification-using-machine-learning-approach/287238

Related Content

An Infrastructure for Wireless Sensor-Cloud Architecture via Virtualization

S. P. Anandarajand S. Poornima (2015). *Next Generation Wireless Network Security and Privacy* (pp. 143-170).

www.irma-international.org/chapter/an-infrastructure-for-wireless-sensor-cloud-architecture-via-virtualization/139429

Authentication in Next Generation Network

G. C. Mannaand Vishnu Suryavanshi (2015). *Next Generation Wireless Network Security and Privacy* (pp. 238-264).

www.irma-international.org/chapter/authentication-in-next-generation-network/139432

Reconfigurable Antenna Systems for the Next Generation Devices Based on 4G/5G Standard

Massimo Donelli (2021). *Research Anthology on Developing and Optimizing 5G Networks and the Impact on Society* (pp. 44-65).

www.irma-international.org/chapter/reconfigurable-antenna-systems-for-the-next-generation-devices-based-on-4g5g-standard/270186

Cooperative Space Time Coding for Semi Distributed Detection in Wireless Sensor Networks

Mohammad A. Al-Jarrah, Nedal K. Al-Ababneh, Mohammad M. Al-Ibrahimand Rami A. Al-Jarrah (2012). *International Journal of Wireless Networks and Broadband Technologies* (pp. 1-15).

www.irma-international.org/article/cooperative-space-time-coding-for-semi-distributed-detection-in-wireless-sensor-networks/85002

AI-Based Wireless Communication

Sanjana T., Lalitha S., Surendra H. H.and Madhusudhan .. K. N. (2022). *Challenges and Risks Involved in Deploying 6G and NextGen Networks* (pp. 42-60).

www.irma-international.org/chapter/ai-based-wireless-communication/306814