Abstract No: SO-23 **Software Intensive Systems**

Predicting box office success of movies using sentiment analysis and opinion mining

Hasara Basnayake* **and Shantha Jayalal**

Department of Industrial Management, Faculty of Science,
University of Kelaniya, Sri Lanka
hasarabas@gmail.com

Movies and social media come together as a result of people sharing their opinions on social media and movie makers using the same platforms for movie promotions. From movie makers to movie goers, many parties are interested in the success or failure of a movie. Forecasting the success of a movie before its release has been a difficult task for many industry analysts. Since film industry's unpredictable nature, many analysts have come up with different algorithms and mechanisms to predict the success of a movie.

One of the mechanisms to predict the box office success is hype analysis. Hype is one of the factors that drive people to the theatres to watch a new movie. Box office opening of a new movie depends on this hype and it will boost up the total box office collection. Hype can be estimated through social media platforms like Twitter. Twitter can be used as a corpus for sentiment analysis and opinion mining.

A movie's success cannot be predicted in a high accurate level solely based on social factors. Classical factors like movie's brand name, cast, director, etc. are also important aspects in movie's performance at box office and should be considered as well. However, a highly accurate method for movie box office prediction integrating both social and classical factors is yet to be introduced for this research area.

In this study, tweets related to the particular movie before releasing are collected using an archiver tool and are used as input data. Then the collected data is pre-processed in order to get a clean dataset. As a part of sentiment analysis and opinion mining, feature selection is performed using N-gram method in order to filter out irrelevant data records and unlike Bag of words method, this does not require an extensive dictionary of words since it uses combinations of words and letters. Afterwards the data related to classical factors are integrated with the proposed formula in order to predict the opening box office collection of the movie. The proposed formula is an extension of a formula used in a previous research and the new extension represent the inclusion of classical factors. Finally, the results are compared with actual box office data and the previous formula results in order to compare and determine the level of accuracy.

Based on initial results, the proposed formula showed of an accuracy level more than 85 percent when the results were compared with actual box office data. Even though it produced a higher accuracy level, the results produced were less than the actual box office values. Thus further testing is needed to determine the actual accuracy level.

Keywords: - Movies, Box office, Sentiment analysis, Opinion mining, Social media