**CE807-7-SP: Text Analytics**

**Assignment 2**

**Task 1**: Report

**Professor**: Dr. Habil Ansgar Scherp

**Student Name**:

Atiwat Onsuwan 1802514

Fazal Jamadar 1802447

**Abstract**

In the world of the digital era, with the rapid growth of the internet, a huge amount of reviews are sent by the customers in describing the polarity between the movies. The reviews are shared by the customers in the form of reviews and comments such as positive, negative or neutral. This kind of websites makes people to have essential decision making. The review of movies is done by the sentiment analysis. In this assignment, we make use of Neural Network LSTM (Long Short term memory) and then make a comparison with another algorithm such as Random forest. It is seen that LSTM has better accuracy, Precision, and Recall.

***Keyword****s-* Sentiment analysis, Rotten Tomatoes, LSTM

# Introduction

Sentiment analysis in machine learning is defined as people’s sentiment or feelings on certain topics and objects. Nowadays, people have progressively expressed their opinions on the web or on social media discussing their views about the latest movies which have hit the theatres. Sentiment analysis uses a computational approach that identifies the opinion and categorizes as positive, negative or neutral. Collecting the data from the people is important to organizations and companies to improve their product by analyzing what satisfies or dissatisfies customers. Using people’s text and the sentiment is an application of NLP (Natural language Processing) [1].

There have been two major approaches which have been used in sentiment analysis from the previous research: lexical based approach and supervised machine learning approach [2]. The Lexical approach such as SentiWordnet has sentiment polarity on them. Furthermore, statistical analysis has been used in supervised machine learning to predict sentiment polarity. In the given dataset Rotten Tomatoes there are five classes of sentiment polarity, Negative, somewhat negative, Neutral, somewhat positive and positive. The goal is to test the data to have an accurate model with the highest possible score in Kaggle competition. Thus, the problem occurs in supervised machine learning because the domain is self-dependent. To get high accuracy, machine learning needs to be retrained if different data set in the domain. Sentiment analysis is categorized in 3 different levels

* Document-level – it obtains the sentiment of complete document or paragraph
* Sentence level – obtains the sentiment of a single sentence
* Sub-sentence level – obtain the sentiment of sub-expressions within a sentence

# Literature review

This section surveys previous work which was carried out. The different approaches that are applied to sentiment analysis are given below.

[3] classified binary sentiment on movie reviews using Naïve Bayes, Maximum Entropy (ME) and SVM. The dataset was collected from IMDB and used various features in which SVM gave the better accuracy about 82.9% with unigram feature [3]. The results obtained from machine learning techniques were quite good compared to the human-generated baseline. Thus SVM provided the better result where as Naïve Bayes gave the poor performance. Although they tried different features, Unigram turns out to be the most effective feature. As we know, any machine learning algorithm surpasses 50% of baseline. They tend to beat the baseline by using unigram feature of 64%.

[4] Classified sentiment using SVM for the multi-domain dataset. 305 positive reviews and 307 negative reviews were collected from a digital camera. Information gain was used as feature selection with an accuracy of 84.15% [4]. The research states that these approaches were combined into one framework to develop a lexicon enhanced method to produce a set of sentiment words. These sentiment features were then combined with a content-free feature which was used in a machine learning approach.

[5] Classified sentiment from tweets using SVM. They investigated with 11875 labeled tweets annotation. The feature which was used was unigram and dependency tree to get an accuracy of 75.39%. Social network sites like Twitter, where people post their real-time reaction and opinions face different challenges. In this research, they used tweets in positive emoticons like “☺” as positive and negative emoticons “☹” as negative [5]. Using Naïve Bayes a model was built for two classification tasks, a binary classifier into positive and negative and three-way task classified into positive, negative and neutral. It experimented in three types of model unigram model, a feature based model and a tree kernel model [5]. The main advantage of this data compared to other data sets used before is that tweets are in streaming fashion and categorize them into a true sample of actual tweets. Moreover, for training and testing, the data set were used was collected by search queries and is biased. Also, different methods of data representation were explored.

[6] used SentiWordnet to classify play store review dataset. The result of the experiment was about 67%. SentiWordNet does not support any other language other than English. In this research, they make use of Bahasa Indonesia language and focus on user opinion. Firstly, the data was translated from Bahasa Indonesia into English and analyzed using SentiWordNet. User opinion data was taken from the Google play store and the Apple store. Secondly, the result is expected to drive new studies to lexicon based semantic analysis.

In the past recent years, we have seen that the neural network plays an important role in the application in natural language processing. [7] Considered in understanding the documents with partial reading, and proposed an alteration that forces them to read input text with skipping. It was found that this approach was faster as it skips unrelated information. [7] Used basic LSTM network as a base model using classification data set such as Rotten Tomatoes. The accuracy they got using the data set Rotten Tomatoes is 79%. From the research, it is observed that LSTM jump is faster than LSTM because the higher the sequence is more the speed will be.

# Conclusion

Nowadays due to the growth of social media, the review of movies has become popular. Sentiment Analysis research which is done in the last decades due to the rise in the growth of big data.

In this paper, we researched various approaches based on Document-Level, Sentence-Level and sub-sentence analysis automated multi-task learning models can consistently outperform the LSTM in sentiment analysis, topic prediction. We also studied about different State-of-Arts in the same area as (Naïve Bayes, and SVM). The perception of automated tasks can be the perception of non-NLP sequence tasks such as image categorization with next row prediction as to the automated task because automated MTL can be integrated into an existing network by adding a new branch to a pre-existing graph.

# Reference

[1] B. T. Nj and R. Y. Nk, “Sentiment Analysis on Movie Reviews Using Support Vector Machines with Linear Kernel,” pp. 1–18, 2016.

[2] A. Firmanto, “Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet,” pp. 202–206, 2018.

[3] B. Pang, L. Lee, H. Rd, and S. Jose, “Thumbs up ? Sentiment Classification using Machine Learning Techniques,” no. July, pp. 79–86, 2002.

[4] Y. Dang, Y. Zhang, and H. Chen, “A Lexicon-Enhanced Method for Sentiment Classification : An Experiment on Online Product Reviews,” *IEEE Intell. Syst.*, vol. 25, pp. 46–53.

[5] R. Passonneau, “Sentiment Analysis of Twitter Data,” no. June, pp. 30–38, 2011.

[6] E. W. Pamungkas, “An Experimental Study of Lexicon-based Sentiment Analysis on Bahasa Indonesia,” *2016 6th Int. Annu. Eng. Semin.*, pp. 28–31, 2016.

[7] H. Lee, “Learning to Skim Text,” 1997.