**Proposed Problem**

The objective of this project is to explore the application of deep learning (DL) on sequences through sentiment analysis, an automated process which categorizes the attitude or emotional tone expressed by a digital text. [1] More specifically, we intend to employ recurrent neural networks (RNNs) to analyze IMDb movie reviews and classify them as either positive or negative based on the sentiment expressed in the text.

**Motivation**

Sentiment analysis is a tool used to automatically assess the tone of digital text. It parses through vast amounts of text data, providing insights into individuals' attitudes toward products, services, or brands. [1] Sentiment analysis presents an unbiased perspective of consumer sentiments to businesses. This is useful for analyzing movie reviews, allowing studios to improve marketing strategies and movie content based on audience reactions. As a result, sentiment analysis has become a key tool for improving product development in all industries.

**Relevant Dataset**

The chosen dataset comes from Stanford's Large Movie Review Dataset, which contains 50K movie reviews from IMDB. [2] The dataset is split into 25K reviews for training and 25K reviews for testing. The outcomes are binary, so a review is either "positive" or "negative".

**Prior Work**

**Sentiment Analysis of US Airlines Tweets using LSTM/RNN**

The paper explores the application of DL techniques including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, for sentiment analysis on tweets about US airlines. The study aims to classify these tweets into positive, negative, and neutral sentiments based on the content related to flight services. The authors utilized word embedding models such as Word2Vec and GloVe to process tweet data and applied LSTM units within RNNs to handle long-term dependencies in text data effectively. The dataset of the paper includes 14640 tweets collected from six different US airlines. The dataset is split into 80% for training and 20% for testing. Also, the authors proposed that Bidirectional LSTM models can enhance performance.

**Sentiment Analysis Using Word2Vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews**

The article investigates the application of sentiment analysis on Indonesian hotel reviews utilizing the Word2Vec and LSTM models. The research focuses on processing the large volume of hotel reviews available online, which presents a challenge for service managers to evaluate the feedback on their services. The study employs Word2Vec for word embedding and LSTM, to classify sentiments as either positive or negative based on various factors such as services, prices, location, food, and facilities. The methodology part consists of two main phases: dataset preparation using the Word2Vec model, and the creation of a sentiment division model based on LSTM. The dataset comprised 2,500 review texts, divided evenly between positive and negative sentiments, collected from the Traveloka website using web crawling techniques. Pre-processing steps included case folding, tokenization, stemming, stopword removal, and padding to prepare the data for training. The Word2Vec model was trained using a

combination of architectures, evaluation techniques, and vector dimensions, while the LSTM model was evaluated with varying dropout values, pooling methods, and learning rates. A 10-fold cross-validation technique was used for dataset division to ensure balanced classes and model evaluation. The results indicated that the best performance was achieved with specific parameter combinations for Word2Vec and LSTM, resulting in an accuracy of 85.96%.

**Proposed Architecture**

Since our goal is to determine the tone of a text, we must consider the context by examining a sequence of words. A recurrent neural network (RNN) would be the most fitting architecture, as it allows for the processing of sequences of text with variable lengths [5]. However, since a conventional RNN suffers from vanishing and exploding gradient problems, we will use the most widely accepted solution, long short-term memory (LSTM), in our sentiment analysis.

The basic structure of an LSTM model consists of a series of LSTM units, each designed to process sequential data while retaining information over time. Each unit contains four interconnected layers that generate the cell's output and update the cell state. These two components will then be passed onto the next LSTM unit. There are three gates that control the cell state: the forget gate, the input gate and the output gate. The forget gate controls how much information should be retained from the previous cell state. The input gate controls the amount of information to be accepted into the current cell state. Lastly, the output gate determines what information will be passed into the LSTM in the next instant of time [6]. Thus, LSTM is suitable for processing sequential text information.

**Proposed Architecture**

1. Input layer
2. Word Embedding Layer (word2vec)
3. Long short-term memory (LSTM)
4. Dropout Layer
5. Dense Layer
6. Activation Layer (Softmax)

**Proposed Procedure**

Firstly, we will clean the data by tokenizing the text to split passages into individual tokens and removing characters that do not contribute much to the semantic meaning of the text, such as punctuations, digits, and common stopwords. After data cleaning, we will transform the words into vector representations using the pre-trained word embedding Word2Vec. Then, we will define and train our LSTM model using the training set. Finally, we will tune the hyperparameters using a validation set, and evaluate the model using the test set.

**Computing Resources**

Since we intend to train our LSTM models using 25K IMDB reviews, we expect that a GPU is needed for our project. We intend to make use of the Department’s Teaching Lab Computers, which provide sufficient computing power for the purpose of our project.

## References


1. What is sentiment analysis? - sentiment analysis explained - AWS. (n.d.-b). https://aws.amazon.com/what-is/sentiment-analysis/

2. Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher, “Learning Word Vectors for Sentiment Analysis,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June, 2011, Portland, Oregon, USA, Association for Computational Linguistics, 142--150, http://www.aclweb.org/anthology/P11-1015

3. Putra Fissabil Muhammad, Retno Kusumaningrum, Adi Wibowo,“Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews,” Procedia Computer Science, Volume 179, 2021, Pages 728-735, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.061. (https://www.sciencedirect.com/science/article/pii/S1877050921000752)

4. R. Monika, S. Deivalakshmi and B. Janet, "Sentiment Analysis of US Airlines Tweets Using LSTM/RNN," 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 92-95, doi: 10.1109/IACC48062.2019.8971592. keywords: {Twitter;Sentiment;Recurrent neural networks;LSTM;Word embeddings},

5. Murthy, Dr & Allu, Shanmukha & Andhavarapu, Bhargavi & Bagadi, Mounika. (2020). Text based Sentiment Analysis using LSTM. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS050290.

6. Understanding LSTM networks. Understanding LSTM Networks -- colah’s blog. (n.d.). https://colah.github.io/posts/2015-08-Understanding-LSTMs/