INTRODUCTION

Our group project involves building a classifier that can label financial news articles as either fact or opinion, using a given dataset that has already been labelled accordingly. We divided the project workload equally among the group members, starting with data exploration to gain insights and clues about the data. After performing in-depth pre-processing, we evaluated three models - LSTM, Naïve Bayes, and Random Forest - to classify the articles. Our analysis of the results provided suggestions for future improvement. The project allowed us to apply traditional machine learning and deep learning techniques to understand how they work in text data classification. Initially, we used TF-IDF with the traditional models, then switched to word2vec with LSTM. Our findings indicate that the LSTM model with word2vec and bigrams performed the best, while in general, all LSTM models outperformed the traditional models. We were surprised to find that the traditional models performed almost as well as LSTM, although this could be specific to the given dataset, and results may vary with other text-based datasets.

MY WORK

During the data exploration phase, I conducted initial checks such as null value check, rows/columns check, and other basic inspections. I also carried out word embedding using word2vec. In the data pre-processing stage, I executed parts of speech tagging and tokenization. Additionally, I implemented LSTM with word2vec and LSTM with word2vec and bigrams during the model implementation phase. I actively participated in all group meetings, report writing, and results analysis discussions. I provided support wherever necessary and was involved in all stages of the project.