Project Documentation

Hung Nguyen

8/13/23

Introduction

Twitter is a great open source of textual data for training specific natural language processing tasks, considering that there are millions of users expressing their own opinions, emotions or experiences regarding various matters everyday. For this, Twitter sentiment analysis is chosen to be the task of interests by many researchers or marketers to investigate how people feel towards certain topics, products, brands, etc.

Within the scope of the project, however, we will solely focus on building, assessing and comparing several machine learning models that are trained specifically for sentiment analysis. The complexity of the models varies from as simple as a Logistic Regression model to a higher level such as the LTSM. We may also discover the results from fine-tuning Transformers. From this, we hope to obtain reasonably strong and robust models and combine with another Topic Modelling model to create analysis and evaluation of several topics being discussed on Twitter, which can be obtained using hashtags.

The project will consist of the following stages:

- Data cleaning
- Modelling
  - Logistic Regression model.
  - Naive Bayes model.
  - Feed-foward neural network.
  - Simple RNN.
  - Bi-directional LTSM.
  - Fine-tuning TinyBERT.
  - Additional topic modelling using LDA.
- Analysis