Foody.vn Sentiment Analysis

Created by Team 1, ML class INT3405E 40.

Vu Duc Hieu
Team Leader
Student, UET-VNU
Hanoi, Vietnam

Nguyen Sinh Hien
Team Member
Student, UET-VNU
Hanoi, Vietnam

Tran Hai Nam
Team Member
Student, UET-VNU
Hanoi, Vietnam

***Abstract*—This document is a report for a Kaggle entry created by the group consisted of 3 members mentioned above. It includes data processing methods, model creations, Kaggle results and future developments.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document shows the process behind the creation and submission of the given model on the Kaggle sentiment analysis contest. The contest includes data from foody.vn, a Vietnamese catering networks for majority of the population. The given data set is compromised of reviews, rating score and additional images. In order to process the sentiment analysis tasks on the dataset, a RNN model will be applied with a language prepossessing module written under NLTK library with an additional Vietnamese stopwords file.

II. Data Analysis

A. Data Analysis

The dataset is delivered under 5 categories: UserID, RevID, Comment, Rating and attached images, with all of given entries are written in Vietnamese. Our training dataset has 5103 entries following the given rules; with Rating entries ranged between 0 and 1, some comments have critical symbols ('¡3', laugh in tears,...) together with worthless symbols ('vn') and duplicate comments are found.

III. Data Processing and Modelling

In this section, we concentrate on the procession and modulation on the given training dataset. Each procedure is coined into each subsection.

A. Data Processing

In order to preprocess the data, our group use NLTK library with a separate Vietnamese stopwords file. The reason behind our choice is the depreciation of Vietnamese natural language processing libraries, for example VNToolkit was written majorly in Java and the only Python API that we found was only for reading the data rather than processing the data; and the fact that NLTK libraries has no built-in dataset for Vietnamese stopwords besides from being the most powerful tool for natural language processing tasks. Our data processing module includes a text splitting method, a stopword removal method, a punctuation removal method, a tokenization method and finally, a denoise method consisted of the above methods.

B. Data Modelling

In this paper, we propose a RNN model under a TensorFlow Sequential method with a LSTM layer included in the model. The model is then trained with Adam optimiser and binary cross entropy loss aiming to get the highest accuracy rating as possible. In details, our model is created from the following layers:

- This model is built under a Sequential model from TensorFlow. With the pre-built model, it provide stacks of layer with one input tensor and one output tensor, which is suitable in this sentiment analysis with exactly one input tensor and one output tensor can be created from each raw entry from the dataset.
- The first layer in the model is Embedding layer. This layer is created in order to put post-processed data into the model, by converting them into sequences of vectors after receiving data from the previously built encoder.
- The second layer, wrapped in a Bidirectional layer, is a LSTM layer. This core layer, pre-built by TensorFlow, is made of three gates (an input gate, an output gate and an forget gate; all of them are powered by Sigmoid function), two cell functions (a tanh-run cell update function and a linear cell state function) and one hidden state indicator created from all of the above entities. It is created with 64 units in order to reduce vanishing gradient. Also, this Bidirectional wrapper is included in the layer in order to reduce time consumption for the sake of less tracking by letting it run till the processed data end.
- The third layer in the model is a Flatten model, built to make data flattened, thus making the model run smoother.
- The fourth layer is a BatchNormalization layer, followed by a Dropout layer set at 0.30 and Dense layers with Relu activation. This layers' setup, with double repetition, applies a transformation in order to maintain the mean output close to 0 and the output standard deviation close to 1 and then apply Relu activation with a Dropout rate of 0.30 in order to to train the distinct elements of the data faster without creating an overfitting scenario to the model.

- Finally, the model is concluded by a Dense layer with Sigmoid activation, which allows the whole data in entries to be trained altogether.

Our training process includes 10 epoches with Adam optimiser and Binary Loss Entropy enabled. By applying all of those into the training process, the initial accuracy is increased and maintained at 0.9, thus contributing to the final score. In this model, we use TensorFlow for model development, because the library has pre-built layers (including LSTM), which reduces the bugging chance and debugging resource altogether.

IV. Experimentation

A. Kaggle Score

Our final Kaggle score on private test data is around 0.68. Comparing to the leaderboard, this is not a good score due to some reasons. Firstly, our language processing module does not process symbols. As can be seen from section II, our given data contain symbols in some entries, with some critical symbols and non-critical ones. Unfortunately, our additional stopword file has no useless symbols included and our model does not process useful ones. Secondly, our model does not process additional images. While some of them are attached for seeding purposes (such as gaining promotional token from Foody), a majority of them is relevant to the reviews themselves, and we have no possible modules for analysing the images and give sentiment scores for each of them.

V. Conclusion

A. Result Evaluation

In conclusion, our given solution is capable of doing sentiment analysis tasks on foody.vn data, but with some restrictions on its efficiency.

B. Future Development

In future, our model will be updated on three aspects. Firstly, a new module for processing symbols will replace the original stopword file and contribute into the language processing module in two distinct directions depending on its value on the whole data. Secondly, a new sentiment model will be created for processing and modelling the attached images in order to create a criteria for the overall score besides from the original models. Thirdly, our original model will be reviewed and updated in the correspondence with the projected image processing model, and a new sentiment analysis based on two models will be developed in order to combine them into the final prediction and increase its efficiency and accuracy.

C. Peer Evaluation

In the project, the group leader is responsible for 40 percents of the workload. Each team member is responsible for 30 percents of the workload.

VI. Acknowledgements

We are thankful to TensorFlow for providing us their RNN sentiment analysis tutorial and their documents. Without those resources, the team project cannot be completed. We are also glad to be part of the Machine Learning course led by Mr. Le Duc Trong and his affiliates, which provides us valuable theoretical knowledge and vision for completing the project.