REPORT ON NLP BRACNH

A progress update.

Contents

[I. THE END RESULT 1](#_Toc131081704)

[II. THE CURRENT PROGRESS 1](#_Toc131081705)

[III. THE DETAILED WORK 2](#_Toc131081706)

[A language model 2](#_Toc131081707)

[a) Why a language model is needed? 2](#_Toc131081708)

[b) How did you use the news data? 2](#_Toc131081709)

[c) What language model did you use and why? 2](#_Toc131081710)

[A Price model 4](#_Toc131081711)

[A Combined model 4](#_Toc131081712)

[Website for users 4](#_Toc131081713)

[A system for data, models 4](#_Toc131081714)

[a) How the data is organized? 4](#_Toc131081715)

[b) How the models are organized? 4](#_Toc131081716)

[c) What you’re building? 4](#_Toc131081717)

# THE END RESULT

The final result we’re aiming is a website allows users to search a company code and gives a buy advise based on history news and prices. For that reason, we need these things to complete the project:

1. A language model that can do sentimental analysis. It takes a news title and gives a label whose value is in {Negative, Positive, Neutral}.
2. A price model that can predict the stocks movement.
3. A combined model that takes the inputs from the 2 above models and gives out a buy advise.
4. A website for user to make a buy advise.
5. A system for automatic data updates and data labeling.

Finishing 5 key points above will mark our project as completed. In the next section, you can find a table listing the completion percentages for each key points.

# THE CURRENT PROGRESS

So far, these are the things I can list out. Feel free to correct me if there’s any mistake. Also, the next section I’ll discuss each key point in details so feel free to skip unclear parts.

|  |  |  |
| --- | --- | --- |
| Key Point | Completion | Note |
| A language model | 90% | A simple BERT model |
| A price model | 10% | A selection of models and plans |
| A combined model |  |  |
| A website for users |  |  |
| A system for data, models | 40% | A system of data retrieval |

# THE DETAILED WORK

## A language model

The first and foremost, a language model that can label give news with one of these labels {Negative, Positive, Neutral}. This is has to be the most time-consuming task in the whole, however, I believe we need to review things before jumping to any conclusion. Please note that we’re doing sentence-level sentimental analysis.

### Why a language model is needed?

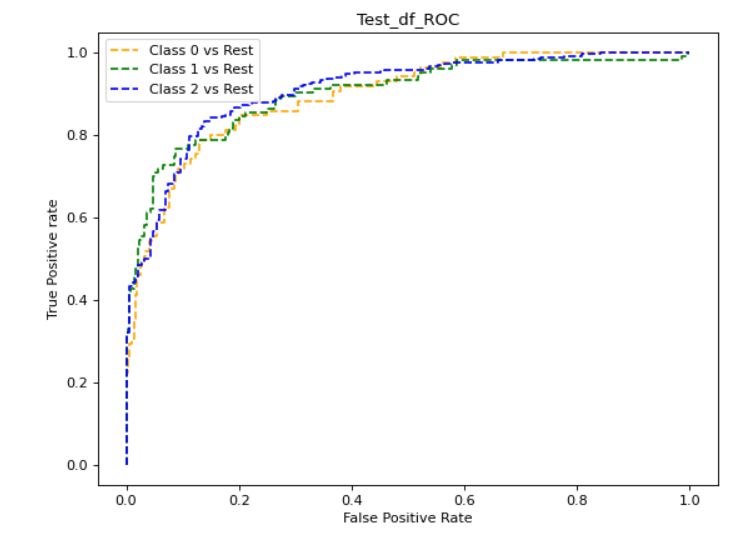
The most important question. With this question I believe it can be clearly answered. As you may have already known, one way to predict movement of a stock is by looking its prices. But the price data is rare and relying on one source of data is not enough to make an educated guess. Therefore, we decided to include news data as it is more diverse and plentiful. The news data can be useful as it provides an overall view of an interest company. You can tell how many achievements a company had, downfalls, future expectations, ...

### How did you use the news data?

The news data is mainly used for sentiment labeling. One thing we did consider is calculating how important a news. But we decided to postpone it as it doesn’t play an important role to the whole project. However, with the knowledge from my Big Data course, I believe we can use TF-IDF to achieve this. The idea is by looking at the important words in each news and how frequent they’re mentioned.

### What language model did you use and why?

Originally, I decided that BERT is a great choice and we also trained some models. The most current one is the bert-muti-lingual.



The image above is the ROC curve based on its predictions on the test dataset. The image shows that the model made good guesses for each labels indicated by how close each curve is to the top left 1.0. However, it’s also clear that it can be improved with more training time. We also briefly look at models such as GPT, PhoBert, ...

For now, we’re still improving the language model. But the main focus is on BERT family models, and I believe GPT despite its extraordinary performance is not a good choice.

|  |  |  |
| --- | --- | --- |
| Feature | BERTs | GPTs |
| Availability | hundreds on hugging face hub. | a few on hugging face hub. Probably not tested |
| Vietnamese | PhoBert, VinAI core for VN preprocessing... | So far none |
| Size | ~= 340 million parameters | ~= 1.7 billion parameters |
| Sentimental analysis | BERT takes into account both left and right context. Useful for VN. | GPT-3 only considers the left context when making predictions. |

There are more in the comparison between these models, but the above things convince me that GPT model is not a good choice.

## A Price model

Nothing much to update aside from a selection of choices including Prophet, CNN...

## A Combined model

Nothing much to report

## Website for users

I’m thinking of re-using the prototype we have shown in January.

## A system for data, models

A complete system for data provisioning, automatically data crawling and predictions.

### How the data is organized?

So far, I have gathered a group of software for data versioning and data provisioning. All the data is stored in a DVC storage provided by Dagshub. This helps me version control the dataset and stream data from online with its Direct Data access feature.

### How the models are organized?

The models can be tracked and used from the MLFlow repository. The platform provides me with lots of helper functionalities.

### What you’re building?

I’m building 2 platforms that provide data and make predictions respectively. And I almost done with the first flat form. These flat forms can be though as duct-tape connecting all functionalities at one place.

<https://symbl.ai/blog/gpt-3-versus-bert-a-high-level-comparison/>

<https://blog.invgate.com/gpt-3-vs-bert#:~:text=Differences%20between%20GPT%2D3%20and%20BERT&text=While%20GPT%2D3%20only%20considers,sentence%20or%20phrase%20is%20essential>.