2020/12/20

Final Report

장호우 (2018000337)
컴퓨터소프트웨어학부

Project Goal

To predict the specific stocks rise or down based on historical finance news.

Financial information data sources

There are three kinds of information data sources, twitter tweets, new-api and wallmine. It should be noticed that Twitter-API or News-API both only support collecting the last 30 days' data on free requests. The wallmine information sources is recommended, in general, it supports the last 5 years historical financial news data at maximum.

To collect the specific company historical news data, only needs to set the stock code when running the script. The details of the script is here. It needs to be noticed that if using the twitter-api or news-api, you need to apply for your own requests key or token. The wallmine source does not need a request key.

Transform financial information data to features

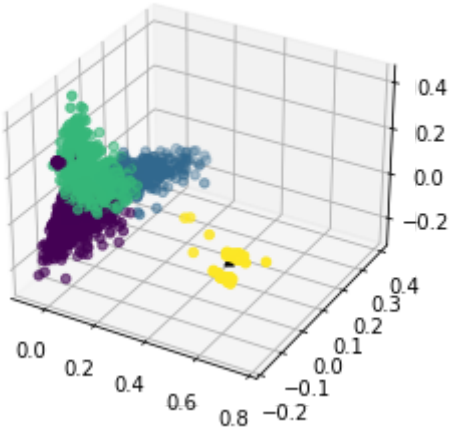
There are three methods used to transform financial information data to features. Most common three methods are the TF-IDF method, Doc2vec method and VADER sentiment method. The last method is difficult from the other two methods, it depends on the sentiment analysis to reduce the features.

TF-IDF method

TF-IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term that occurs in the text has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF-IDF weight of that term. Simply, the higher the TF-IDF score (weight), the rarer the term and vice versa.

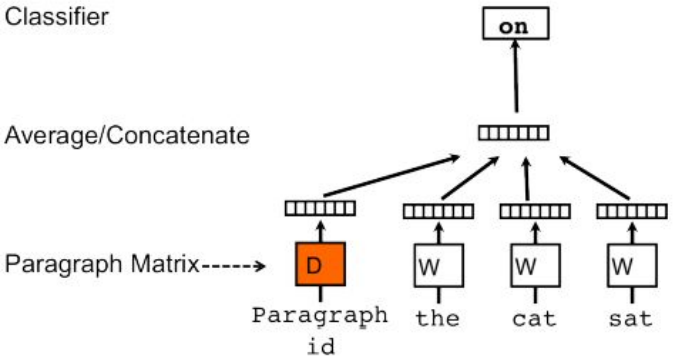
The TF-IDF algorithm is used to weigh a keyword in any content and assign importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the corpus.

The visual representation of features in TF-IDF method processing as follows.

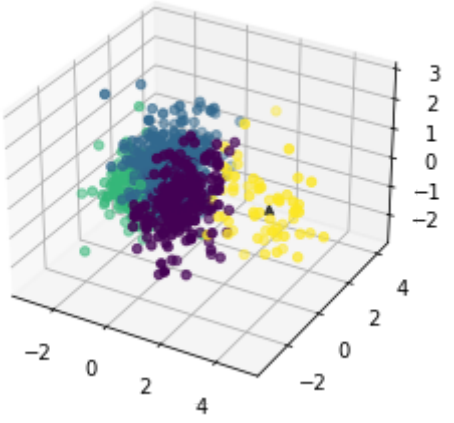


Doc2vec method

Doc2vec is to create a numeric representation of a document, regardless of its length. But unlike words, documents do not come in logical structures such as words, Therefore, Mikilov and Le have used was simple, yet clever: they have used the word2vec model, and added another vector (Paragraph ID below), like so:



The visual representation of features in Doc2vec method processing as follows.



VADER sentiment analysis

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. There are four return values as following, positive, negative, neutral and compound. In addition, the compound value is the most important of them.

1. How to Streamline the 6,714 Photos Cluttering Your Phone
2. Fire and Fury: Kimmel, Colbert Skewer Trump Over Wolff Book Fortune
3. This entrepreneur is ringing up sales restoring vintage telephones - MarketWatch

news1	how	to	streamline	the	6,714	photos	cluttering	your	phone		
news2	fire	and	fury	kimmel	colbert	skewer	trump	over	wolff	book	fortune
news3	this	entrepreneur	is	ringing	up	sales	restoring	vintage	telephones	marketwatch	

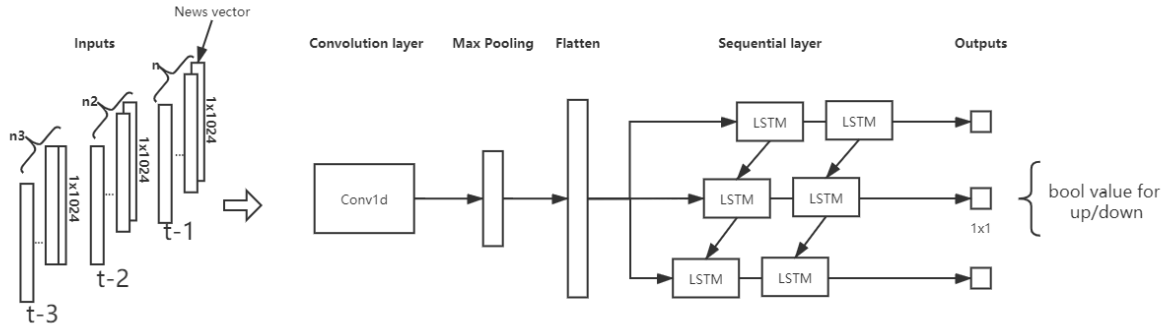
	compound	negative	neutral	positive
news1	0.0	0.0	1.0	0.0
news2	0.7269	0.404	0.596	0.0
news3	0.296	0.0	0.804	0.196

The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive).
The normalization of compound:

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

* where x = sum of valence scores of constituent words, and α = Normalization constant

Model



The model structure as above, a simple CNN-LSTM model.

Metrics

The project used the Daily Return value as target value in the beginning but the models could not converge.

$$R_i = \frac{ClosingStockPrice_t - ClosingStockPrice_{t-1}}{ClosingStockPrice_{t-1}}$$

Therefore, the metrics of the whole project was changed to **UP/DOWN** value to determine the accuracy of the models.

Results

Using the last 3 day's historical finance news and picking up 16 news each day, 300/4 features each news as input. The numbers of features via three kinds of different methods.

The accuracy of prediction as following table.

	TF-IDF	**Doc2vec**	**VADER**
days	3	3	3
news/day	16	16	16
features/news	300	300	4
accuracy	**0.52**	**0.54**	**0.58**

Conclusion

The VADER sentiment analysis gave us an edge over TF-IDF and Doc2vec methods. The whole project using the historical financial news data to predict, in the further, using the do sentiment analysis on tweets would be get better accuracy.

Appendix

Daily return with VADER sentiment analysis four return value.