Sentiment analysis of Financial Tweets Machine Learning for Natural Language Processing 2020-2021

Nathan Bry
ENSAE-MS Data Science
nathan.bry@ensae.fr

Bertrand Vuillemot
ENSAE-MS Data Science
bertrand.vuillemot@ensae.fr

Abstract

We used Bert architecture on a dataset of financial tweets concerning US financial markets. We labelized the dataset *by hand* in order to obtain a qualitative and sufficiently large dataset. We will then train Bert with a view to predicting if a tweet is *positive*, *neutral* or *negative* about the concerned institution.

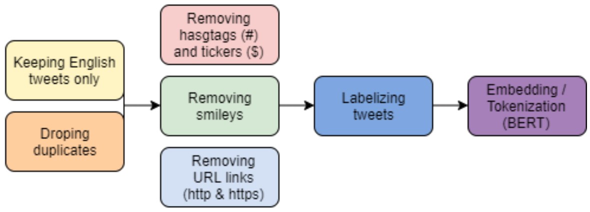


Figure 1: Preprocessing of tweets

1 Problem Framing

Retail trading frenzy was highlighted a couple months ago by "meme stocks" (*GameStop* or *AMC* for example). Hedge funds are now scrapping Social media like *Reddit* and *Twitter* in order to analyze wich stock is going to be targeted next and what will be the direction: Long or Short.

In this context, we believe that using machine learning with a view to analyzing quickly and precisely posts could be a major strength. But it requires two things: the raw materials (mostly tweets or posts) and a sentiment to give an idea about the direction of the stock's price.

We found a dataset of financial tweets from *Kaggle* in which the tweets have been scrapped from Twitter[1]. Those 28k+ posts concern publicy traded companies and cryptocurrencies (also a very hot subject). However, one of the major challenges posed by the dataset was the absence of labels, compulsory for using the state of the art algorithm in NLP: Bert (Devlin et al.).

2 Experiments Protocol

As expected, the tweets weren't perfectly cleaned and required some preprocessing. First we had to keep only the tweets written in English, and then drop the duplicates. Then, our preprocessing pipeline involved removing everything that could disturb the models, such as URL links (http and https), all the hashtags (#) and ticker symbols ($) that were followed by company names for example (to avoid overfitting on the company ticker), and the smileys thanks to regular expressions (re library). We then used regular expressions one more time to labelize the tweets and give them a sentiment (*Positive*, *Neutral* or *Negative*). For example "*outperforms*" is a strong signal of positive tweet while "*bearish*" is a strong signal of negative tweet. However, the sense of tweets can be complex ("*The market outperforms Apple AAPL*") which is actually a negative tweet. We then checked every of the 6k+ tweets selected[2] to train the algorithm (allowing us to eliminate absurd or useless tweets at the same time).
Finally, the embedding and tokenization took place inside the Bert model.

3 Results

We then evaluate the scores and create a confusion matrix for the predictions, and we observe really good predictions (over 90% f1_score) for positive and negative tweets (Figure 2). Neutral predictions are less precise maybe due to quality of tweets or a more questionable classification.
Among common errors concerning neutral labels, we observe that "*Tuesday's Top Analyst Upgrades and Downgrades...*" is classified as positive (label 2) while it's negative (label 0).

[1] *https://www.kaggle.com/davidwallach/financial-tweets*

[2] available at https://github.com/nathanbry2/NLP-ENSAE

Classification Report:

	precision	recall	f1-score	support
2	0.9028	0.9378	0.9200	901
1	0.7935	0.8359	0.8141	262
0	0.9791	0.8792	0.9265	480
accuracy			0.9044	1643
macro avg	0.8918	0.8843	0.8869	1643
weighted avg	0.9077	0.9044	0.9050	1643

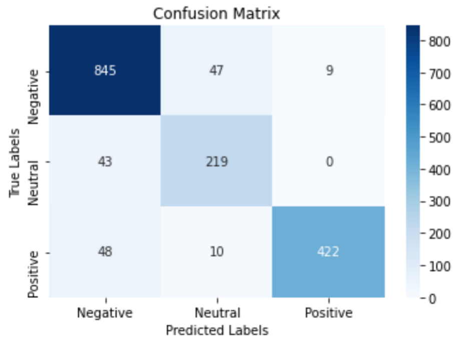


Figure 2: Scores of Bert-3 labels

That may be induced by the fact that 'losers' and 'winners' are mentioned within the tweet (while we only analyse one company per tweet). But the majority of misclassification of neutral comes from the tweets of type "*open interest for maturity 07/20/2018. High pu...*" that the algorithm predicts as neutral while we choose to classify it as negative (since it concerns put options and then a strategy benefiting from a slumping price of the stock). Even so, we believe that a lot of those neutral tweets are not very meaningful.

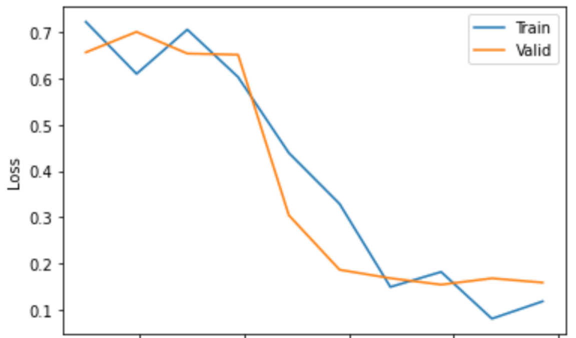
We then decided to focus on a more "economic standpoint" and giving up on neutral tweets: equivalent to a buy (positive) or sell (negative) approach. Thus, we only try to predict negative or positive tweets. The scores of the Bert-2 labels (Figure 4) show an non-negligible increase towards 95% (weighted or not). The train and valid score curves look 'healthy'(Figure 3) i.e. showing no under or over-fitting phenomenon.

We observe that the majority of the errors are due to misclassification of $predicted = negative$ while $true\_label = positive$. Among pitfalls, we still face the problem when a tweet speaks about winners and losers within the same post.

4 Discussion/Conclusion

We trained a pretty efficient algorithm in predicting either a tweet is positive or negative. An enhancement of our model would be the ability to take into account that different companies are concerned within a tweet and then return a sentiment for each company mentioned.

Figure 3: Train-Val of Bert-2 labels



Classification Report:

	precision	recall	f1-score	support
1	0.9577	0.9789	0.9682	901
0	0.9587	0.9187	0.9383	480
accuracy			0.9580	1381
macro avg	0.9582	0.9488	0.9532	1381
weighted avg	0.9580	0.9580	0.9578	1381

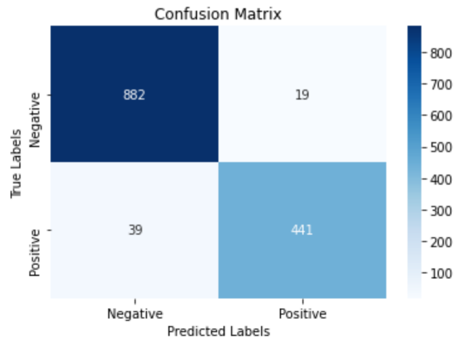


Figure 4: Scores of Bert-2 labels

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.