[Kwantx]

# Initial Research on NLP

Citation: *arXiv*

*By: Jimmy Lin,1 Rodrigo Nogueira,and Andrew Yates (2021)*

(650)690-2002 3101 Park Blvd, Palo Alto, CA adrian@kwantx.com

## *Content*

- Motivation
- Background Information
- General framework
- BERT: Important elements
- Where to go from here

## *Background Information/Motivation*

- Premise:
  - Based on a study that measures how quickly Twitter news is disseminated in comparison to major news outlets, we have reason to believe that Twitter could be a better source to use when making investment decisions
- Goal:
  - Because we can react to breaking news quick through Twitter than through major news outlets, we're trying to see if we can somehow automate the process of identifying, filtering for, and providing a basic sentiment analysis of the breaking news
- Methods:
  - Using NLP, we could teach our machines to get familiar with financial Twitter language. The specifics of our implementation will be discussed in the following slides

## *Implementation*

- Part 1: How to identify breaking news data and determine relevance
  - “Novel and topical business news and their impact on stock market activity”
  - Annotate training data so we can teach our machine how to identify
  - Also looking at how time should affect the importance of the news/sentiment score
- Part 2: How to perform sentiment analysis on various tweets relating to the market and use them
  - “FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining”
  - Must annotate to teach machine what is positive, negative, and relevant news

## *Part 1*

- Uses two measures to see if an article is worth looking at
  - Topicality and novelty
    - Novelty: a particular news v.s. what has been published in the past

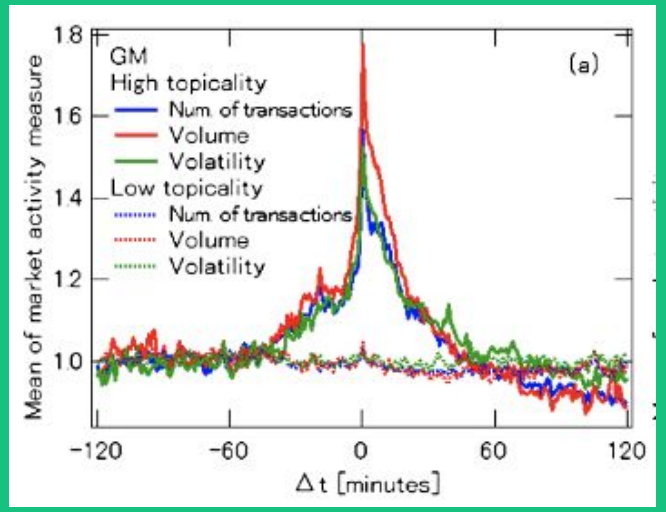
$$Nov(a_t) = \sum_{0<\Delta t \leq \tau} SIM(a_t, a_{t-\Delta t}).$$

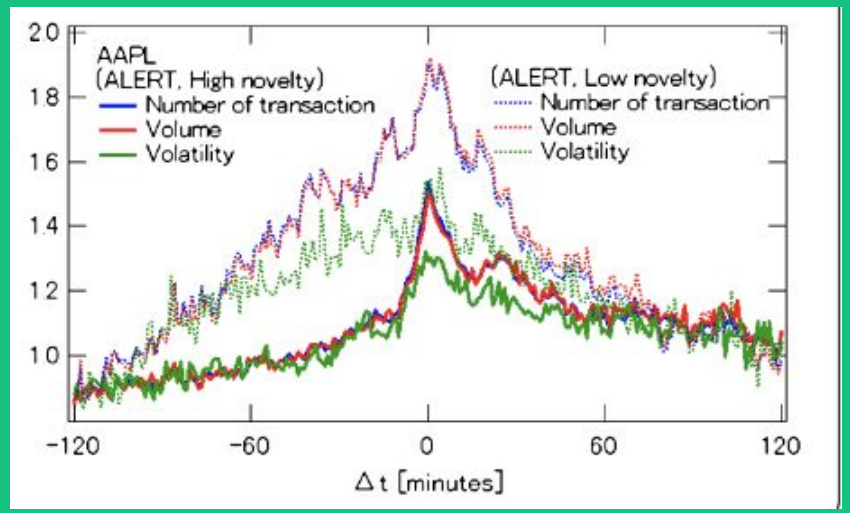
    - Topicality: a particular news v.s. what is published by other agencies

$$Top(a_{t,k}) = \sum_{j \neq k, j \in K} SIM(a_{t,k}, a_{t,j}).$$

- Mainly relies on cosine similarity, which is a very reputable method of comparing similarity between articles
  - https://github.com/huggingface/transformers/issues/876

## *Part 1*


GM
(a)
High topicality
Num. of transactions
Volume
Volatility
Low topicality
Num. of transactions
Volume
Volatility
Mean of market activity measure
18
16
14
12
10
−120
−60
0
60
120
Δt [minutes]


AAPL
(ALERT, High novelty)
Number of transaction
Volume
Volatility
(ALERT, Low novelty)
Number of transaction
Volume
Volatility
20
18
16
14
12
10
−120
−60
0
60
120
Δt [minutes]

## *Part 2*

- Implement BERT
- Fine-tune BERT to FinBERT:
  - Capitalization Prediction pre-training task
    - Predicting whether the word is capitalized or not
    - Beneficial for financial named-entity recognition
  - Token-Passage Prediction pre-training task
    - Predict whether the token appears in segments of the original passage
    - This token would be the “key point” of the article
- Pre-trained text:
  - English Wikipedia, BookCorpus, Financial Web, Yahoo Finance, Reddit QA

[Kwantx]

www.kwantx.com

(650)690-2002 3101 Park Blvd, Palo Alto, CA adrian@kwantx.com