Breaking News Technique: Idea

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

- Novel and topical business news
- How to use this method for our purpose

Novel and topical business news

- 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

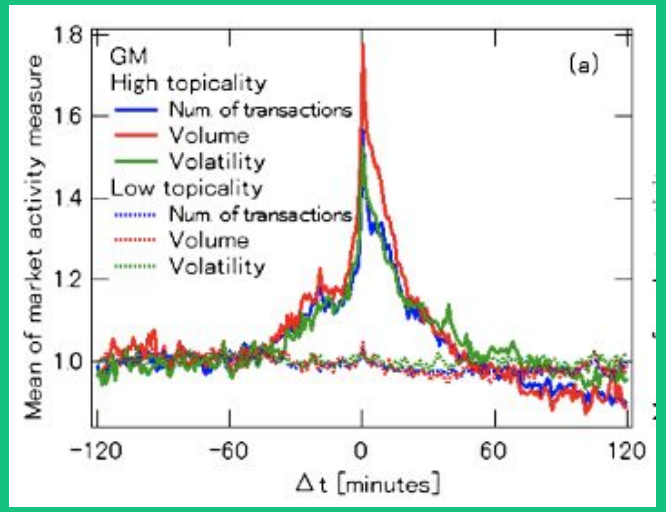
$$Nov(a_t) = \sum_{0<\Delta t \leq \tau} SIM(a_t, a_{t-\Delta t}).$$

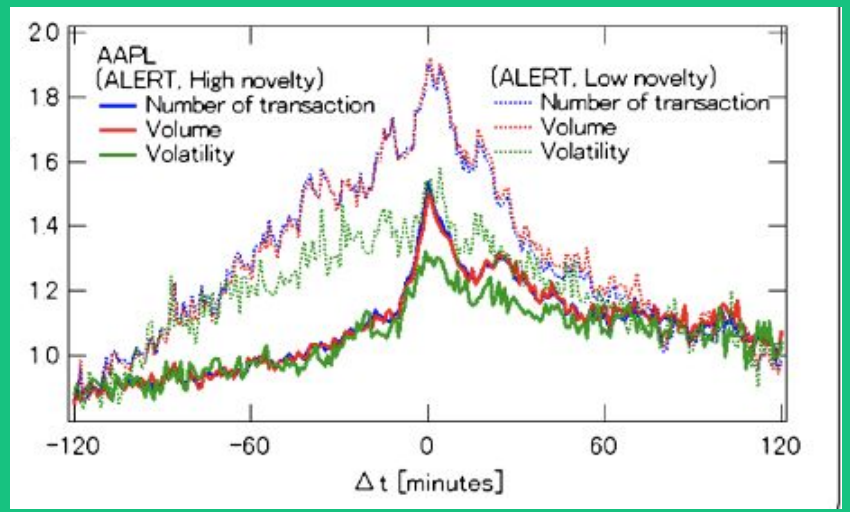
    - Topicality: a particular news v.s. what is published by other agencies

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

- Mainly relies on cosine similarity, which is a very reputable method of comparing similarity between articles
  - https://github.com/huggingface/transformers/issues/876
  - The first reply to this link might be something that we want to use.

Proof of the method working





How to use this idea for our purpose

- We need a method to calculate similarity, or SIM() in the equation mentioned in the previous slide
  - The github repo takes care of this issue
  - This transformer is faster than a BERT model because “every sentence / article is mapped to a fixed-sized vector”
  - On the other hand, “if you have 10,000 sentences/articles in your corpus, you need to classify 10k pairs with BERT”
- For this reason, this transformer is a much better way of measuring similarity than BERT out-of-the-box. The example is linked here.
- The github code above is only really used for mapping the sentences to vectors. We can modify it to check for cosine similarity, for which the formula is where w represents the vector of a specific article

$$SIM(a_i, a_j) = \frac{\vec{w}_{a_i} \vec{w}_{a_j}}{|\vec{w}_{a_i}||\vec{w}_{a_j}|}.$$

[Kwantx]

www.kwantx.com

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