Twitter Breaking News

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
- Directional Prediction of Stock Prices using Breaking News on Twitter
- Breaking News Detection and Tracking in Twitter
- Where to go from here

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

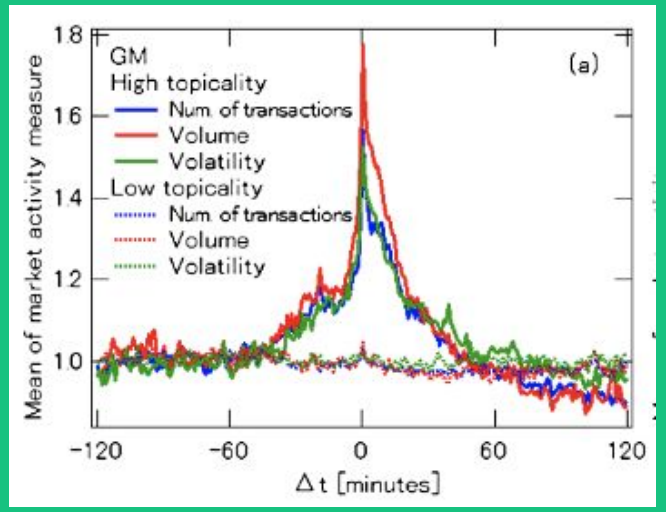
$$Nov(a_t) = \sum_{0<\Delta t \leq \tau} SIM(a_t, a_{t-\Delta t}).$$

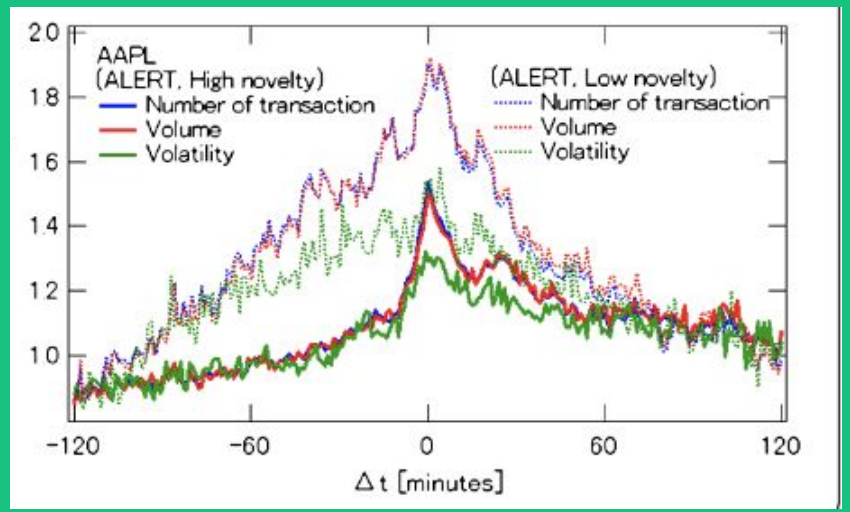
    - Topicality: a particular news v.s. what is published by other agencies

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

- 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





Novel and topical business news: Verdict

- The method of calculating novelty and topicality is a great place to start
  - The article has proved that it has worked for news articles. The first step to take from here is to see if their method works for Twitter news data
- Perhaps we can approach novelty and topicality as a type of optimization problem, where a novelty level below a certain threshold (most novel = 0) and a topicality level above a certain threshold will be identified as breaking Twitter news
  - Must remove some words like “and” so that it doesn’t count those for similarity

Directional Prediction of Stock Prices using Breaking News on Twitter

- Data:
  - Simply used Twitter Streaming API from March 2014 to September 2014
- Method:
  - $$Breakout = \begin{cases} True & \text{if } N(d,h) \geq \mu[20](d,h) + 2\sigma(d,h)) \\ False & \text{otherwise} \end{cases}$$
  - Used to calculate whether or not a tweet is a breakout tweet. "N represents tweet volume on specific date d, and hour h. μ [20] is 20-hour simple moving average applied on tweets' volume, μ[20](d, h)+2σ(d, h) represents the upper band for simple moving average. "
- Verdict:
  - Method of identifying breaking news seems a little too brute force/mathematically naive
  - Twitter data set simply uses API, no filtering process

Breaking News Detection and Tracking in Twitter

- Method:
  - $$sim(m_1, m_2) = \sum_{t \in m_1} [tf(t, m_2) \times idf(t) \times boost(t)] \quad (1)$$
  - $$tf(t, m) = \frac{count(t \text{ in } m)}{size(m)} \quad (2)$$
  - $$idf(t) = 1 + \log\left[\frac{N}{count(m \text{ has } t)}\right] \quad (3)$$
  - Then group them based on similarity by comparing the first message in the group to the current message
  - If score is above a certain threshold, the message is added to the group
- Verdict:
  - I believe that there are more efficient, accurate ways to categorize Twitter data if that is the focus of this article.

Reflection

- First article is very nutritious, perhaps look into additional literature that is related/cited to figure out more about how it is implemented
- Overall, the second and third article fall behind in comparison to the first one
- What we can focus on from here on out is the following:
  - Check if it's also an efficient way of checking for similarity in Twitter data
  - Dive deeper into optimizing cosine similarity and better implementing it (github)

[Kwantx]

www.kwantx.com

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