

SENTIMENT CORRELATION DISCOVERY FROM SOCIAL MEDIA TO SHARE MARKET

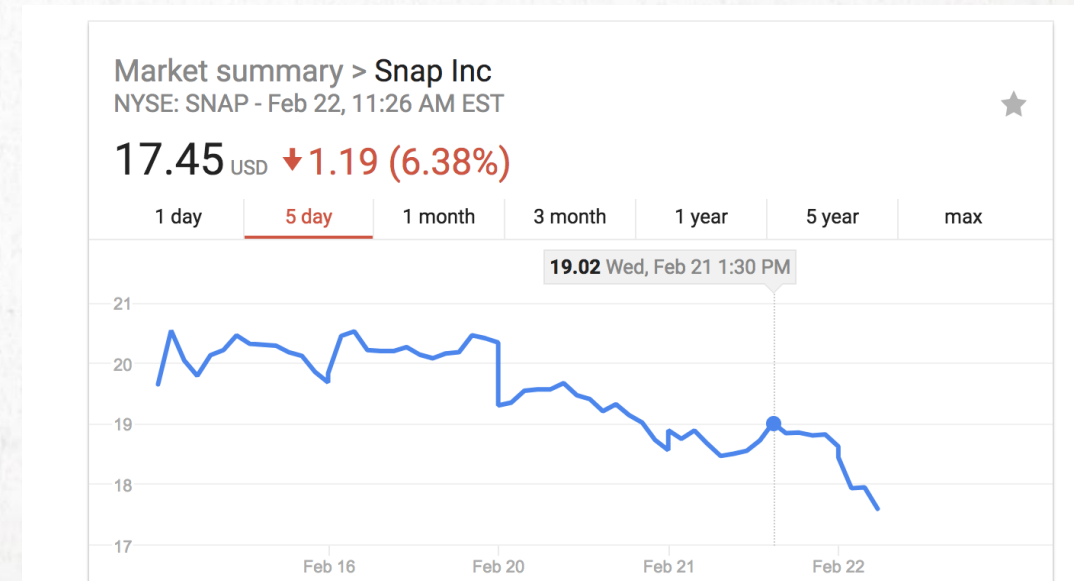
SIMON XIE, MAN LI, JIANXIN LI

POWER OF SOCIAL MEDIA

- Rapidly increase from 7% since 2005 to 65% in 2015
 - 68% of adults in U.S uses Facebook in 2018
 - Some example shows social media affects share price significantly
 - Attracts researchers and analytics
 - Becomes an important tool for B2C communication, maintain relationship between business and customer
-

SOCIAL MEDIA INFLUENCES TO SHARE PRICE

- Researches shows small business has significant effects by social media
- Special Case: Kylie Jenner and Snap Inc.
 - Share value drops 6.4% after negative sentiment comment exposed to Twitter on February 2018



MOTIVATION

- Why that could happen?
 - Does this replicable?
 - What is the correlation between?
-

PATH OF DISCOVERY

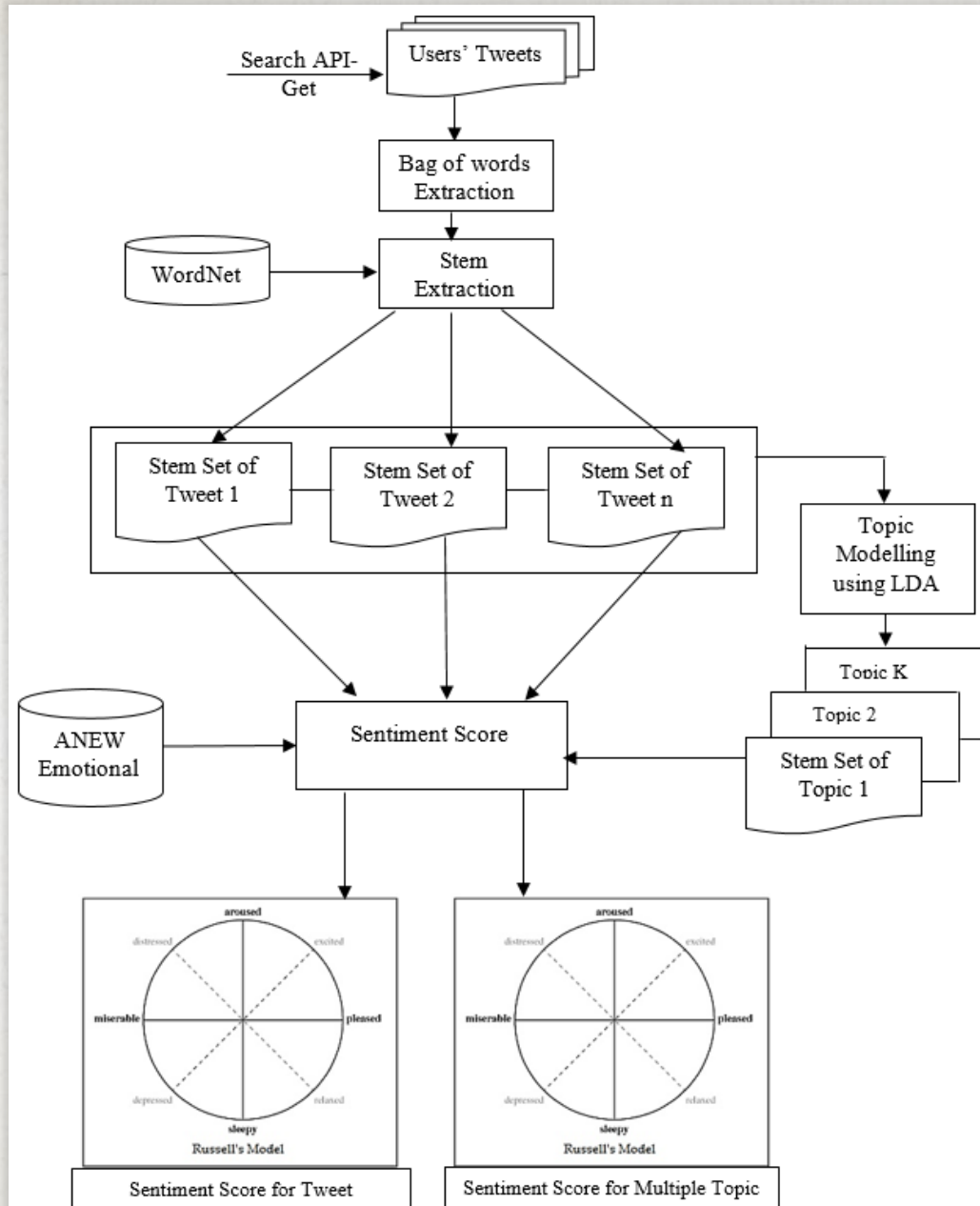
**METHOD TO CREATE NUMBERS FROM
SENTIMENT AND SHARE MARKET**

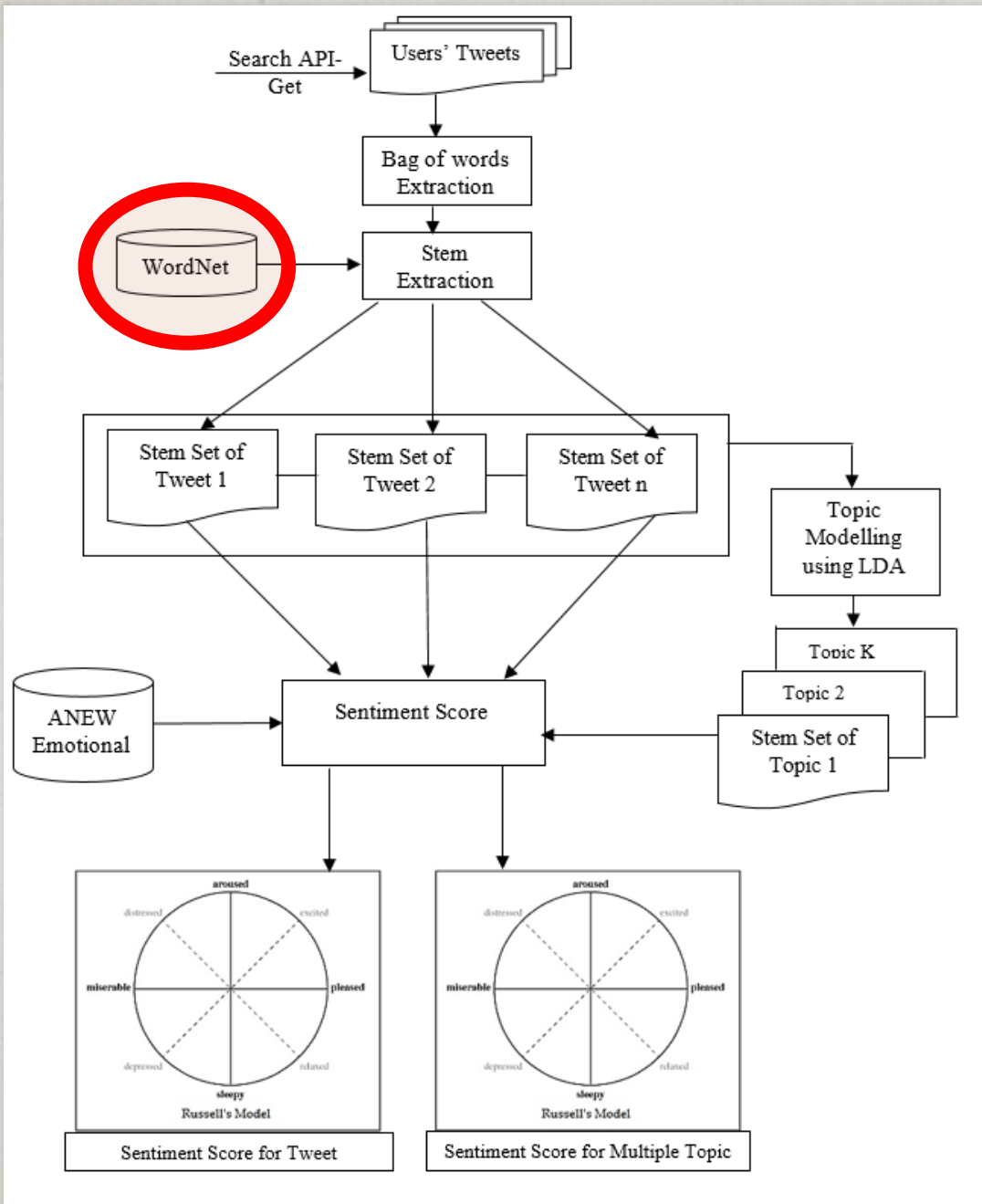
REQUIREMENTS

- Find out sentiment of specific company across public discussion
 - Extract company-related social media posts
 - Using natural language processing to discover general discussed topic from extracted posts
 - Generate sentiment score on topics
 - Compare Sentiment with share market trends
 - Collect Share price data with the same time period with posts used for sentiment extraction
 - Compare sentiment score with share price
-

INSPIRATION OF SENTIMENT EXTRACT

- A model introduced by Debashin et. al in 2016 called sent_LDA
- A model create sentiment score over set of single tweets and topic modelling
- An example uses Latent Dirichlet Allocation, generates sentiment score, and stem extraction



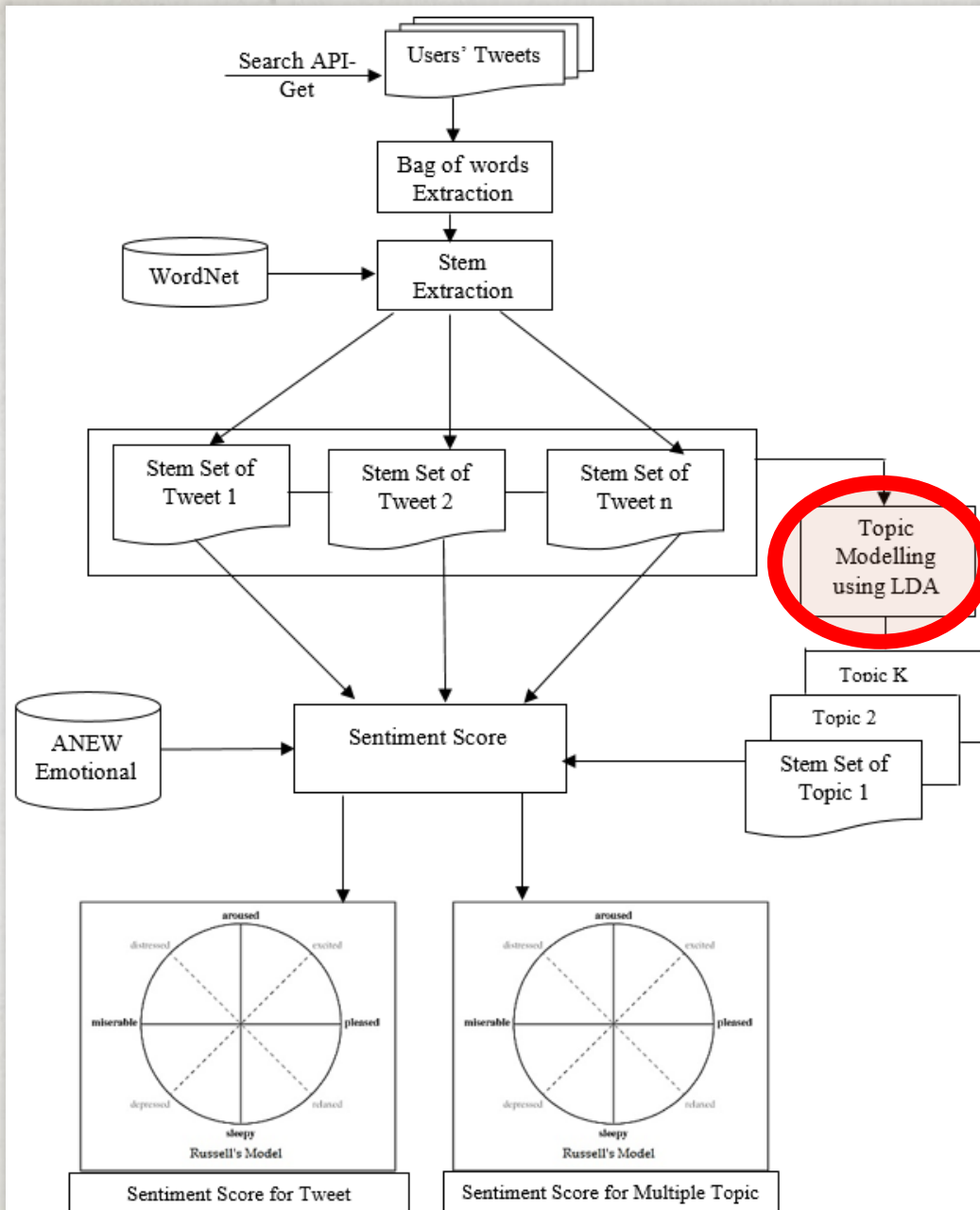


WORDNET

Introduced by Miller on 1995, provides English dictionary for computer programs, for computer to understand natural English words, useful for stem extraction

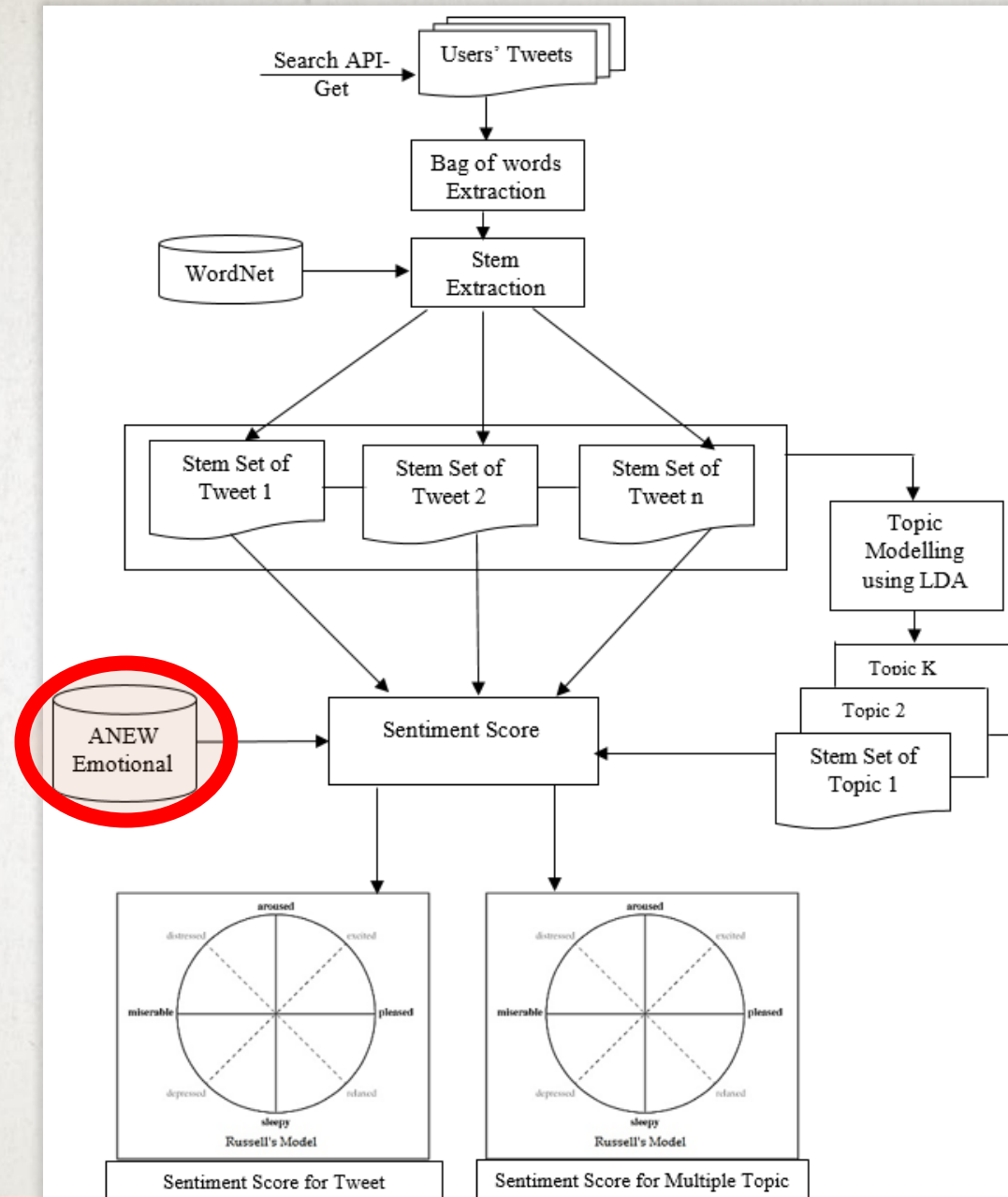
LATENT DIRICHLET ALLOCATION

- A generative statistical model Introduced in 2003 for improved topic modelling strategy
- A self-learning model with corpus only, based on variety models by Michael et. al in 1999
- Each word is treated as part of completed topic and generate probability for each word.
- Being widely used in natural language processing as its ease of use



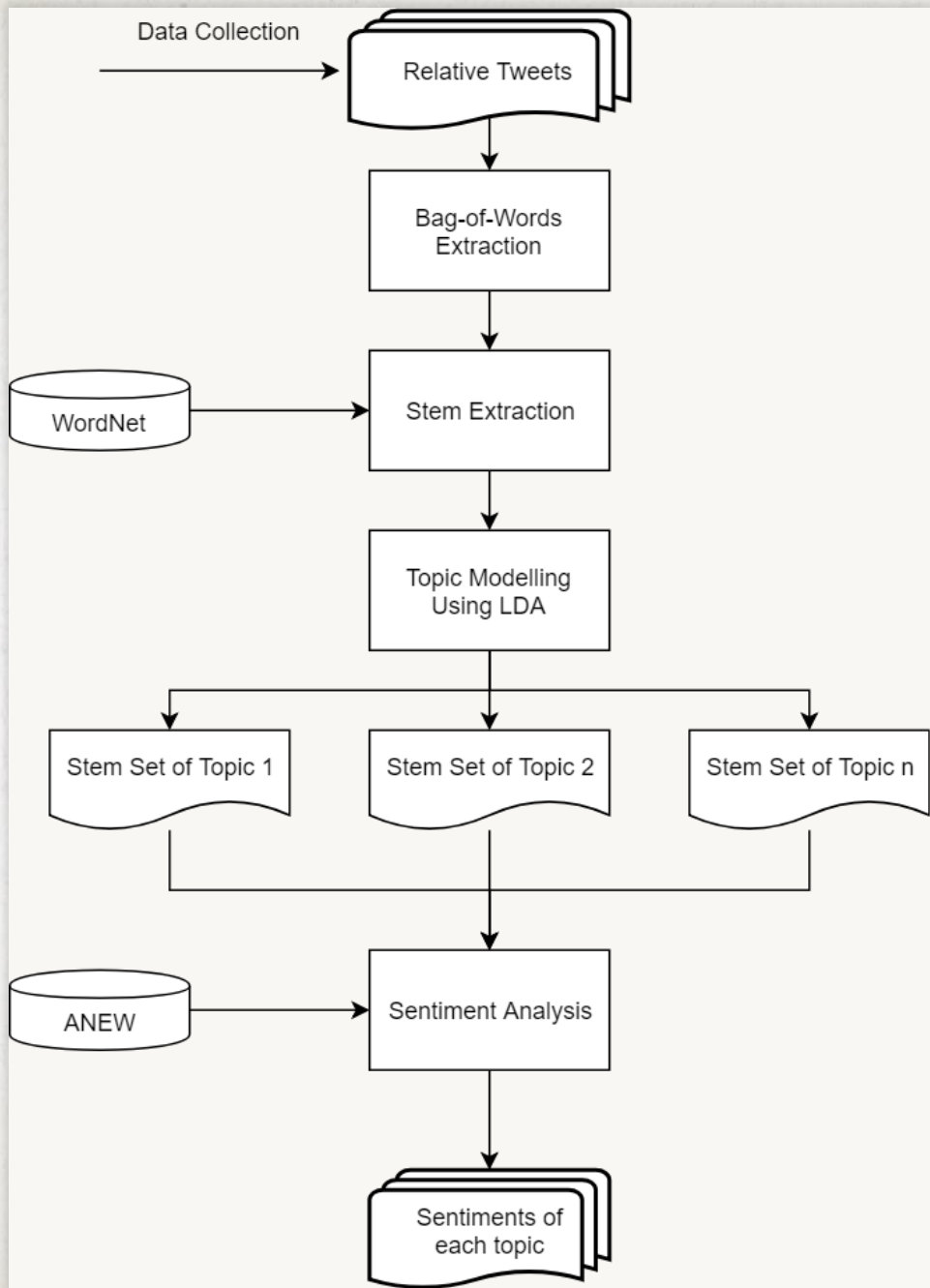
AFFECTIVE NORMS FOR ENGLISH WORDS

- Provided by Bradley in 1999, gives referential sentiment score for generally used words assessed by Self-Assessment Manikin (SAM) in a scale of 1-9, where 5 means neutral, 1 means negative polarity and 9 means positive polarity
- ANEW treats word in 2 dimension: Valence and Arousal. Feelings could represents with both dimension
- Evolved ANEW: Introduced in 2013, based on origin version, it provides sentiment attachment for 13,915 words



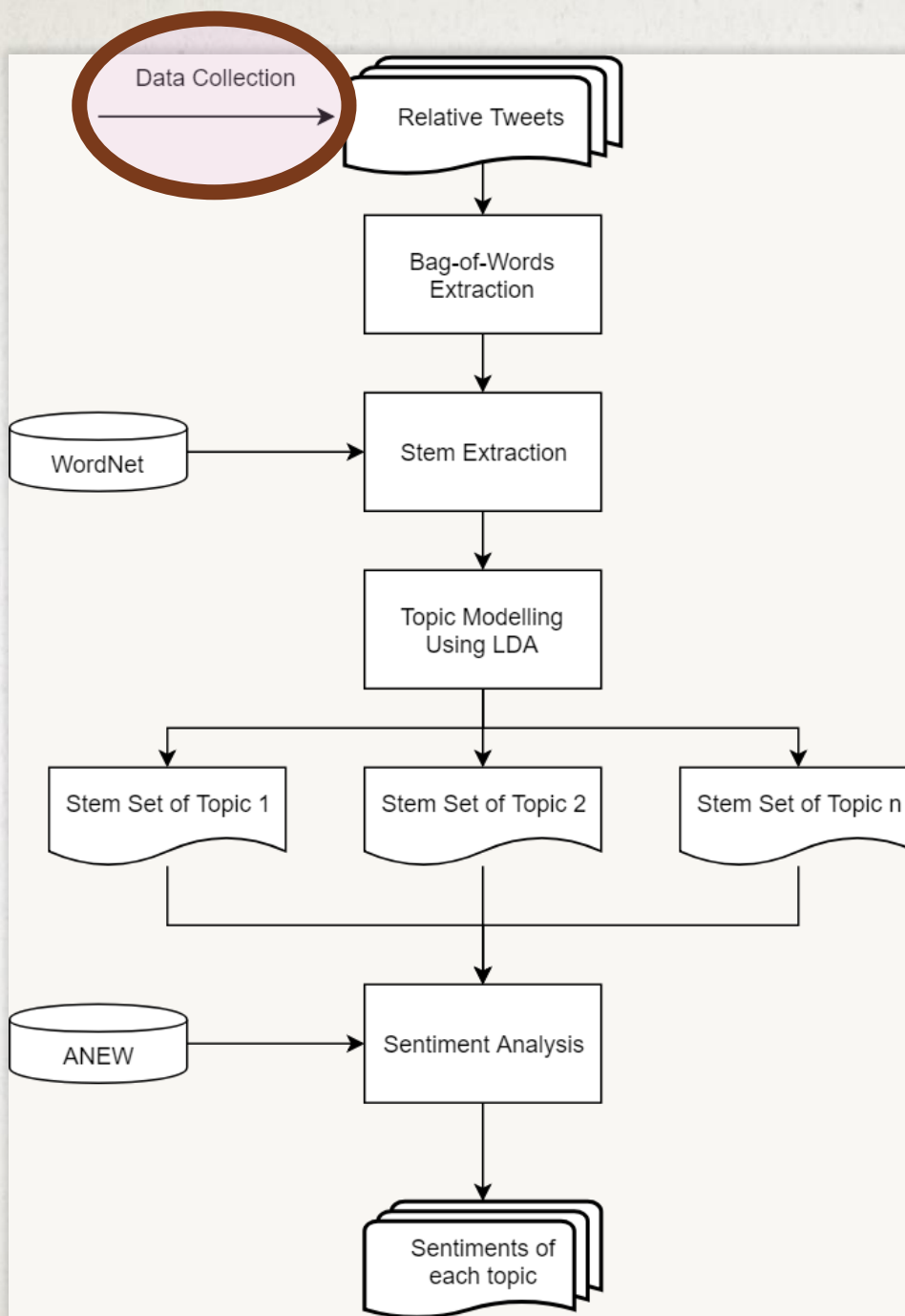
CHANGES NEEDS TO MADE

- Sentiment based on every twitter cost long time and so much resources
 - Using topic only is sufficient for wide-ranged amount of twitter
-



SENTIMENT EXTRACTION WITH TOPIC ONLY

- A simplified version based on sent_LDA
- Removes single twitter analysis and keep topic modelling only
- In this research, a sentiment correlation discovery is conduct on company Apple Inc. using public data on October 2017

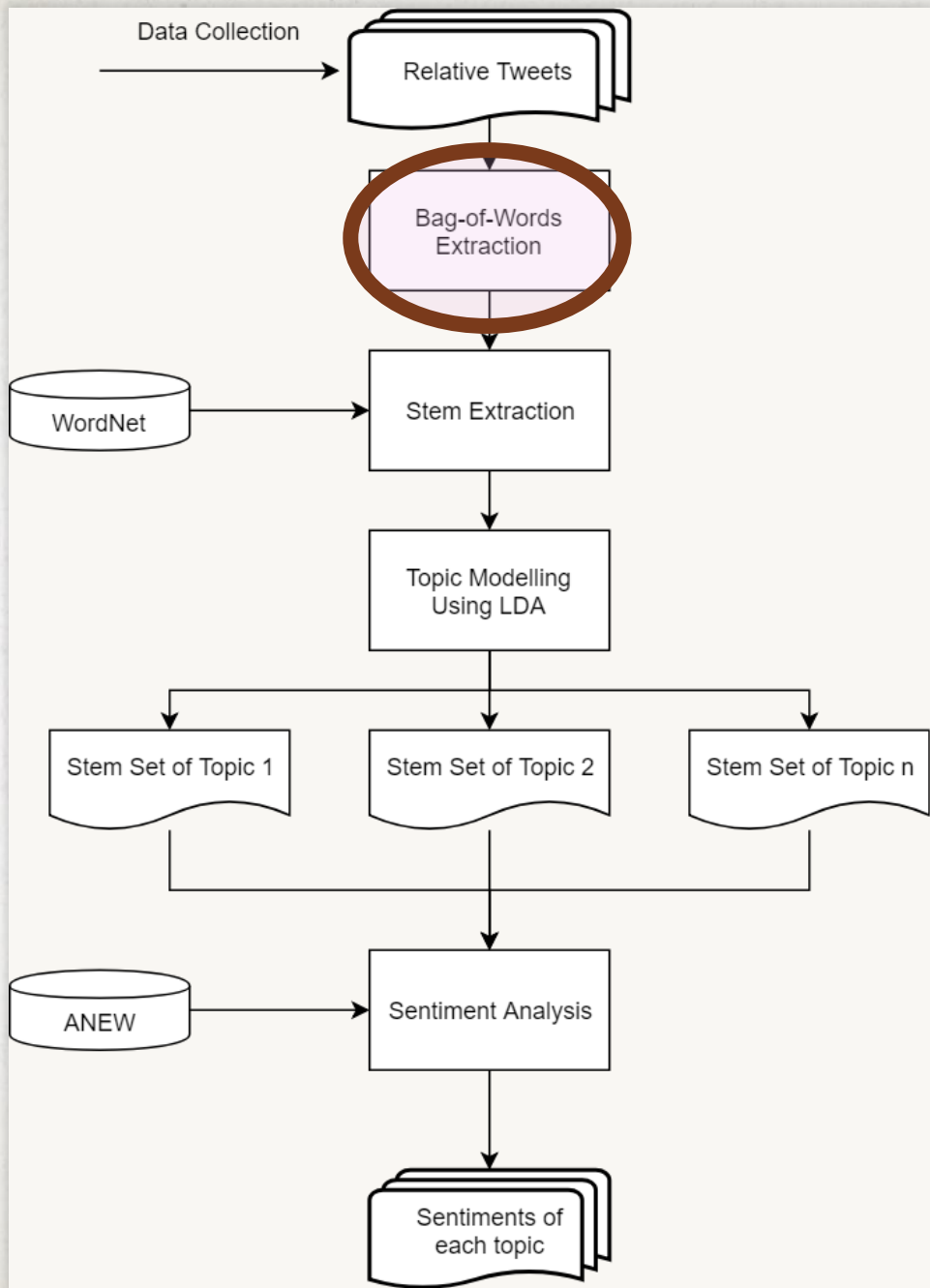


DATA COLLECTION

- Using Twitter tweets as data as easy to extract
- Raw Twitter data are gathered from the Internet Archive
- Extract relative tweets based on keywords associated to desired company
 - Using Google Trends and WordStream Free Keyword Tool for common topic extraction
 - Manually pick representative keyword from list provided from above
- Or Extract relative tweets based on direct mention (“@”) to company’s twitter account

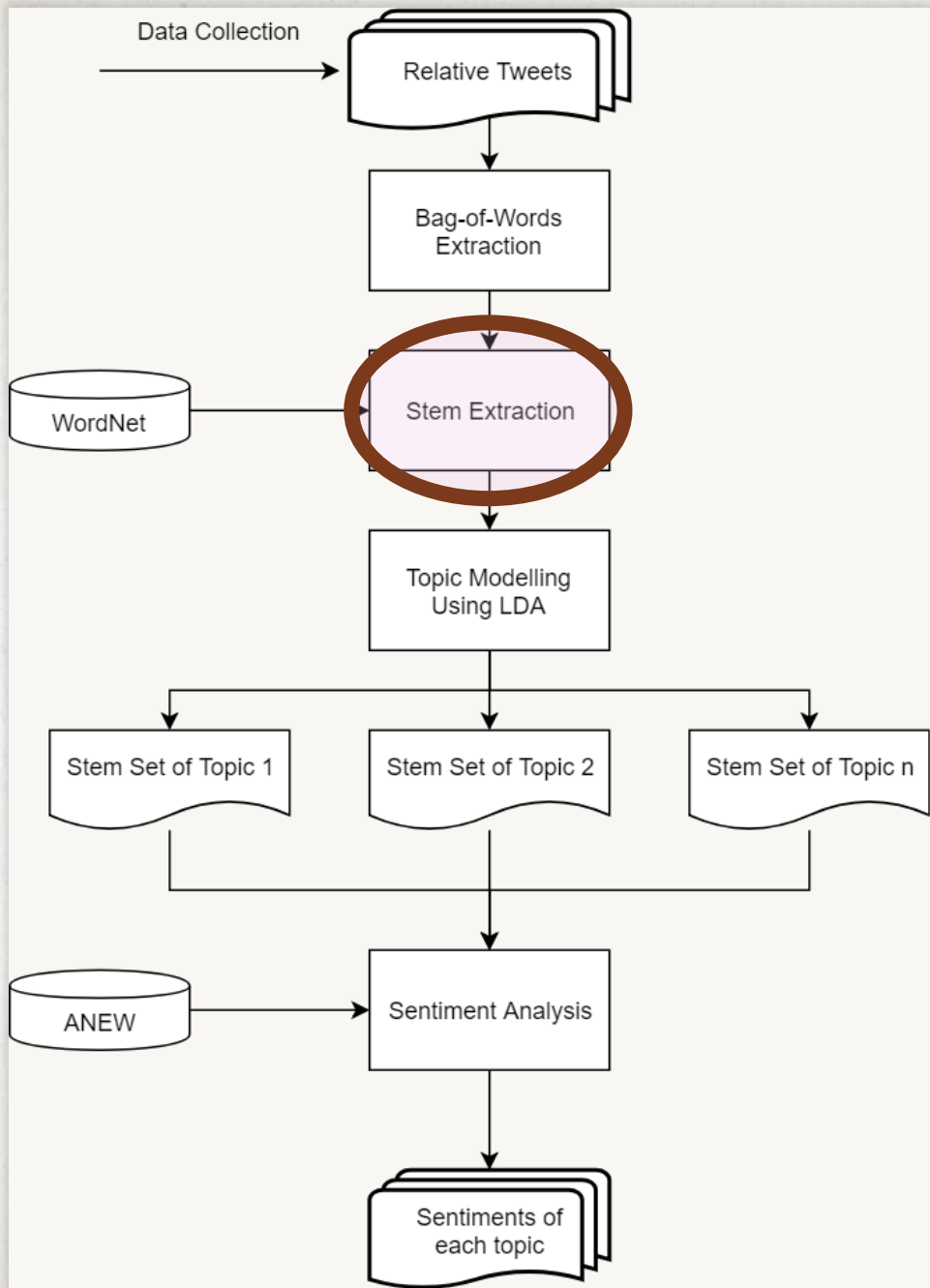
TWITTER EXTRACTION EXAMPLE

- Examples uses Apple Inc.
 - Keyword:
 - iPhone
 - iPad
 - Apple Watch
 - Mac
 - iOS
 - AppleCare
 - Apple Music
 - Direct Nomination
 - @Apple
 - @AppleSupport
-



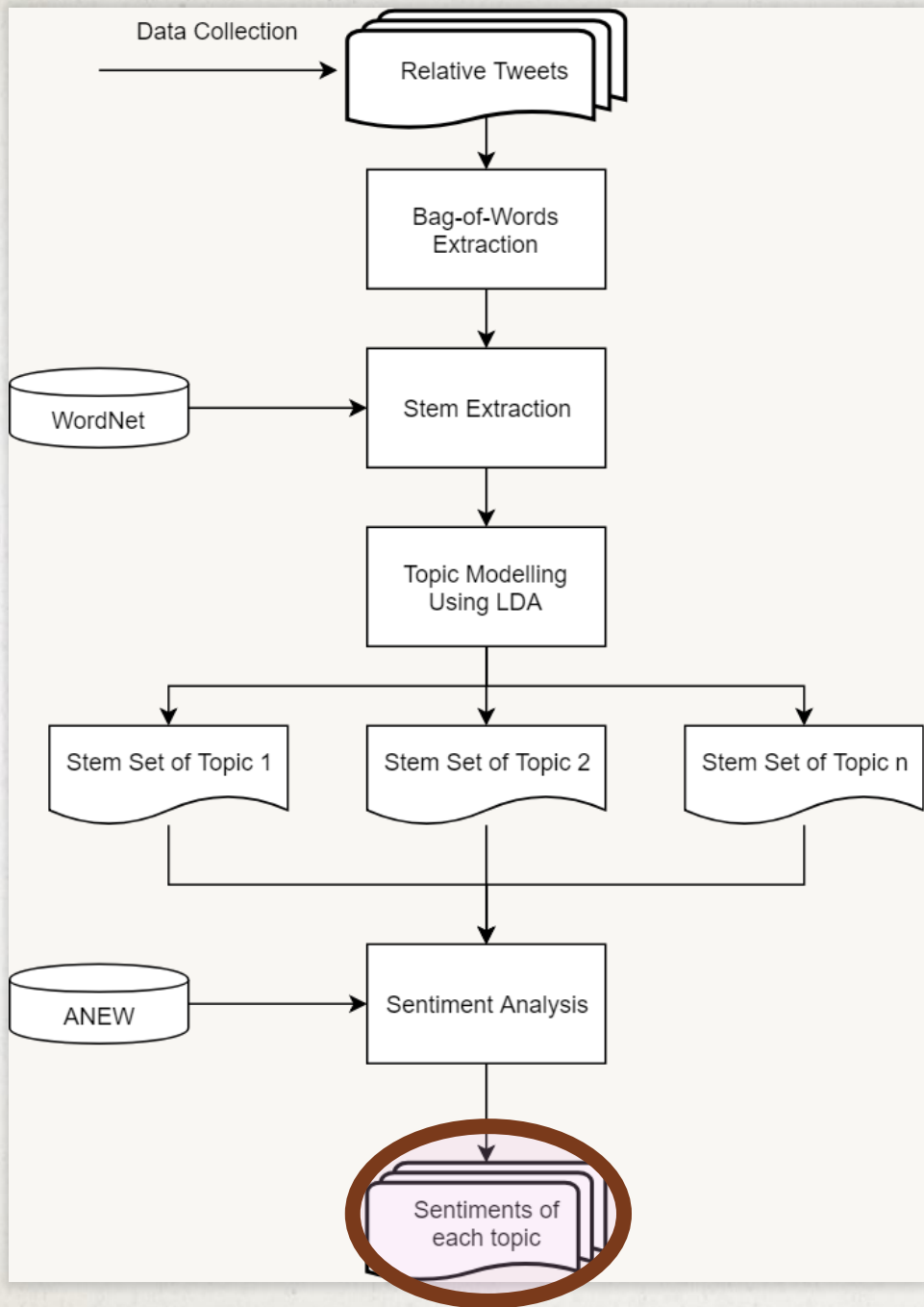
BAG-OF-WORDS EXTRACTION

- removes stopwords and take each word into object
- Stopwords: common used word which does not contain meaning or too widely used with different meanings
 - Afterwards
 - Again
 - Although
 - Computer
 - Could
 - Either
 - Except
 - (and many more)



STEM EXTRACTION

- Looking up words in dictionary (WordNet) and find the stem of each word in the dictionary
- Stem: The origin form of word, without any transformation.
 - Word “Tall” is stem of list “Tall, Taller, Tallest”
 - Word “she” is stem of list “she, her, hers”



OVERALL SENTIMENT SCORING

- Calculate overall sentiment score based on words and probability extracted from LDA model
- Average: summarize sentiment score of every word extracted from LDA topic model and divide word count of it
- Weighted: calculates the overall score by involving weight of every word from LDA model with equation:

$$S_{Overall} = \frac{\sum (S_{word} \times W_{word})}{\sum W_{word}}$$

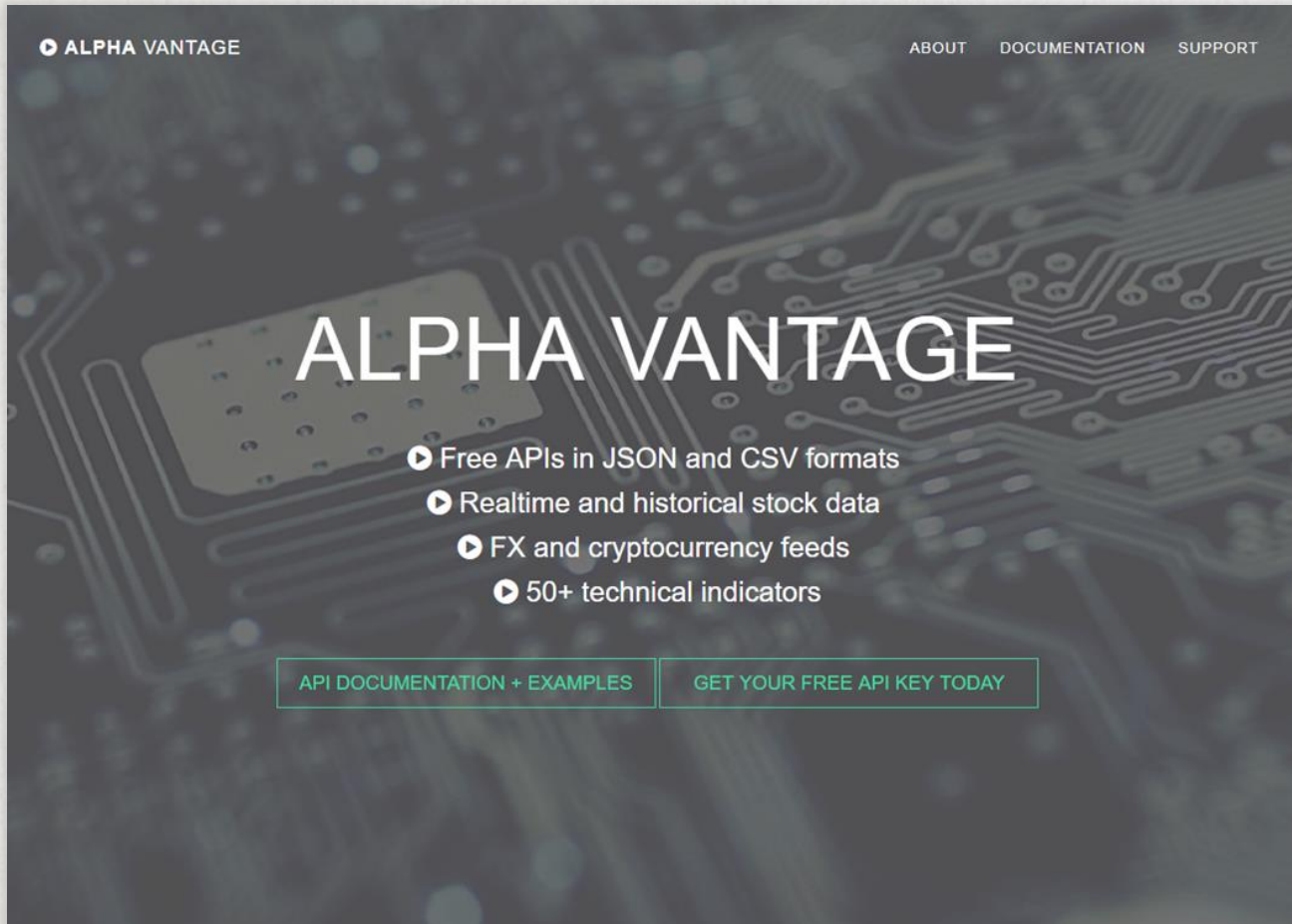
$S_{Overall}$ - overall sentiment, S_{word} - sentiment of word, W_{word} - weight of word

SENTIMENT SCORING EXAMPLE

- Word: Open
 - Valence: 6.14
 - Arousal: 4.43
- Word: Sad
 - Valence: 2.1
 - Arousal: 3.49



- Assuming word Open has probability 0.4 and word Sad has probability 0.6
- Average: Valence: 4.12, Arousal: 3.96
- Weighted: Valence: 2.7336, Arousal: 3.1572



FINANCIAL INVOLVEMENT

- Uses sentiment score to compare with share price of desired company (Apple)
- Gather historical share market data through Alpha Vantage: A free API for collecting JSON and CSV format share market data for up to 20 years

STRATEGY

- Taking Sentiment score of all topic generated by LDA model as an overall sentiment score
 - Run sentiment extraction on twitter related to Apple Inc, separated daily
 - That is, each day in October 2017 is one set of corpus
 - There are total 31 runs to generate 31 set of sentiment score
 - To be compare with daily share price change
-

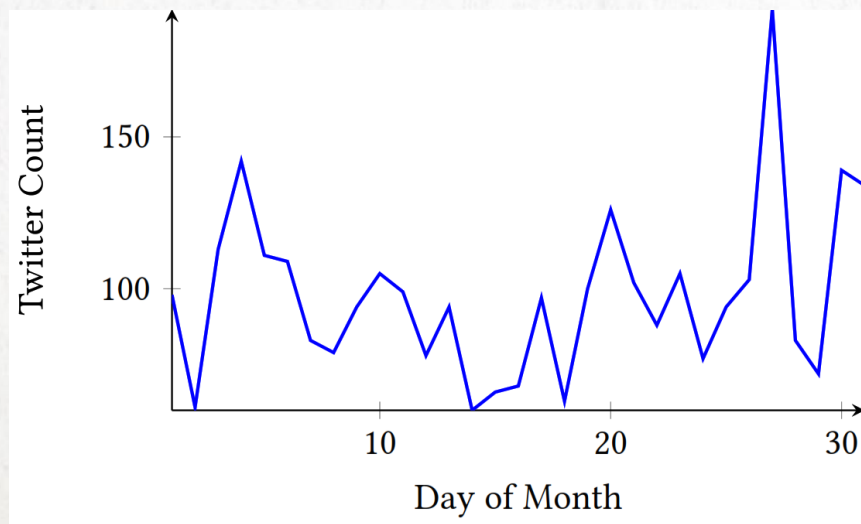
DISCOVERY

USING APPLE INC ON OCTOBER 2017

TOPIC POPULARITY

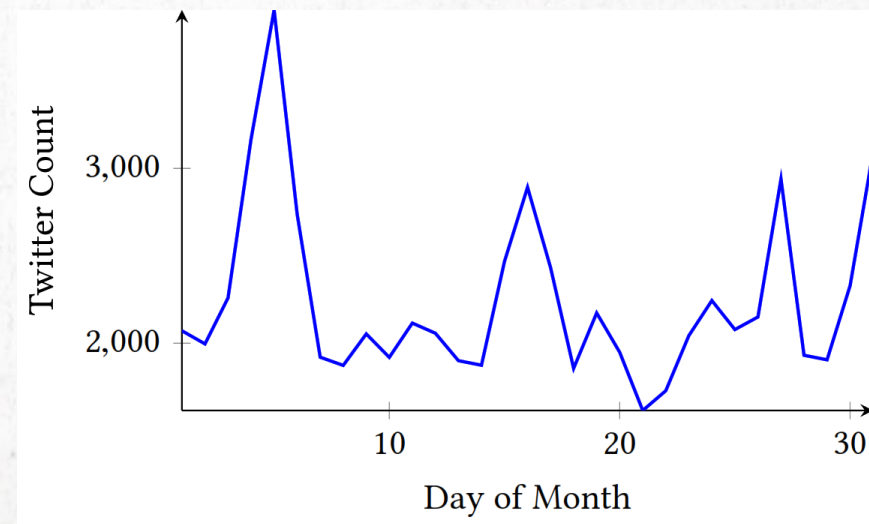
DIRECT NOMINATION

- Less than 200 tweets over 1 million English tweets daily



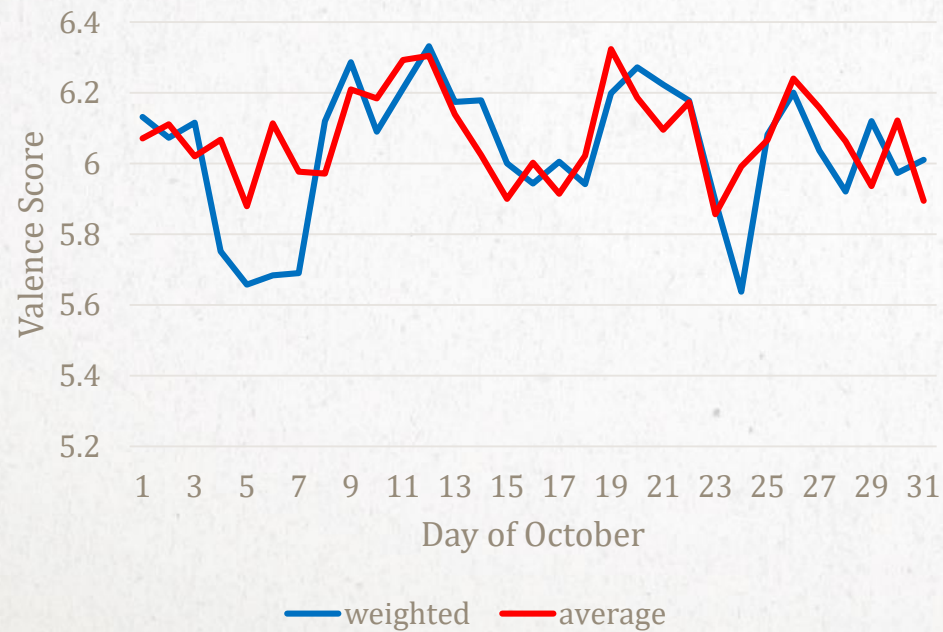
KEYWORD MATCHING

- Up to 4,000 tweets over 1 million English tweets



WEIGHTED AND AVERAGE SCORE COMPARE

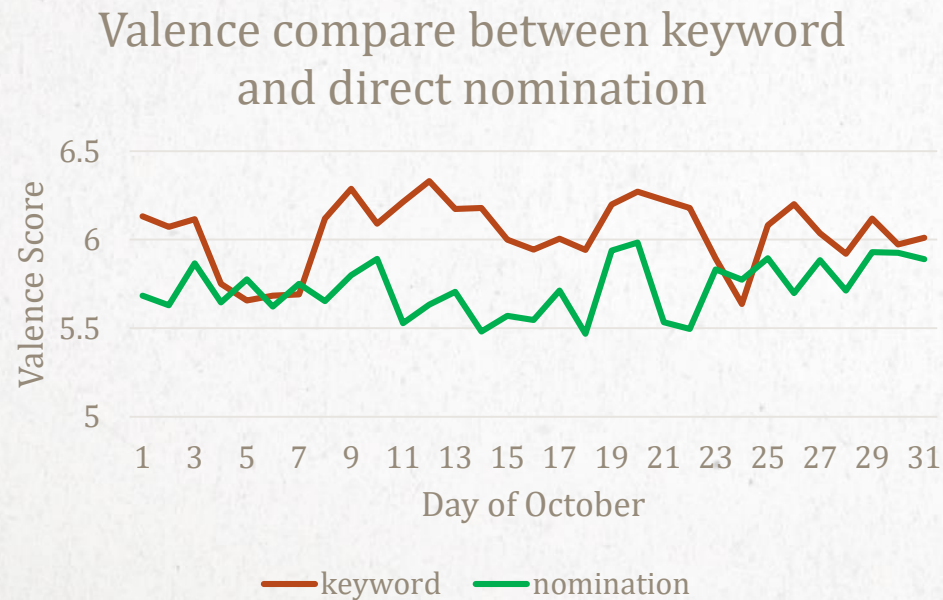
Valence Sentiment Score compare
between Weighted and Average
calculation



- Valence Score: higher means happier
- Weighted calculation is more stable than using average, and it is more reasonable with conjunction to topic modelling

SENTIMENT COMPARE BETWEEN KEYWORD AND NOMINATION - VALENCE

VALENCE

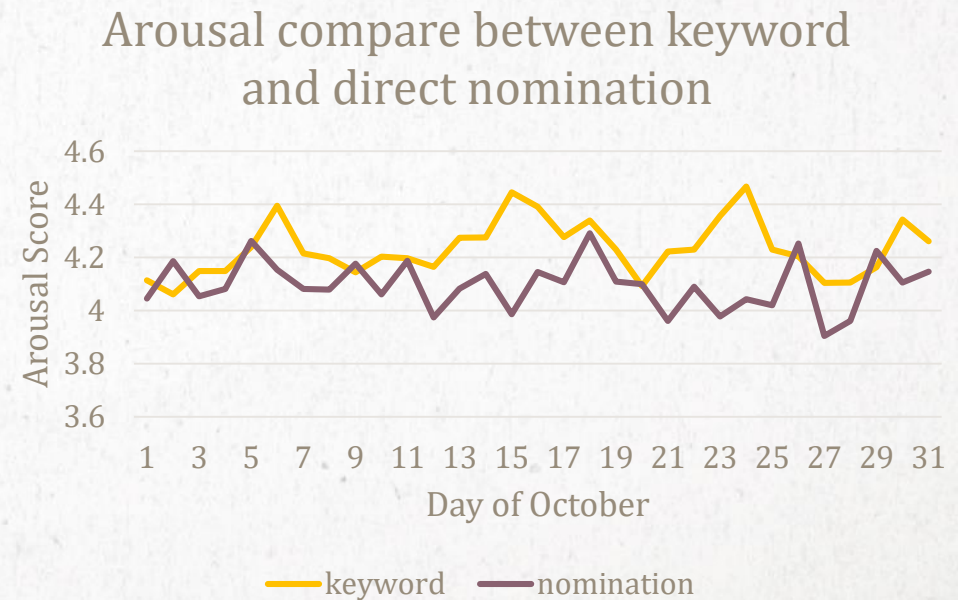


- Valence Score: Higher means happier
- Using direct nomination seems less stable than keyword finding, and in real life it is less possible for people to feeling happy become unhappy and back quickly

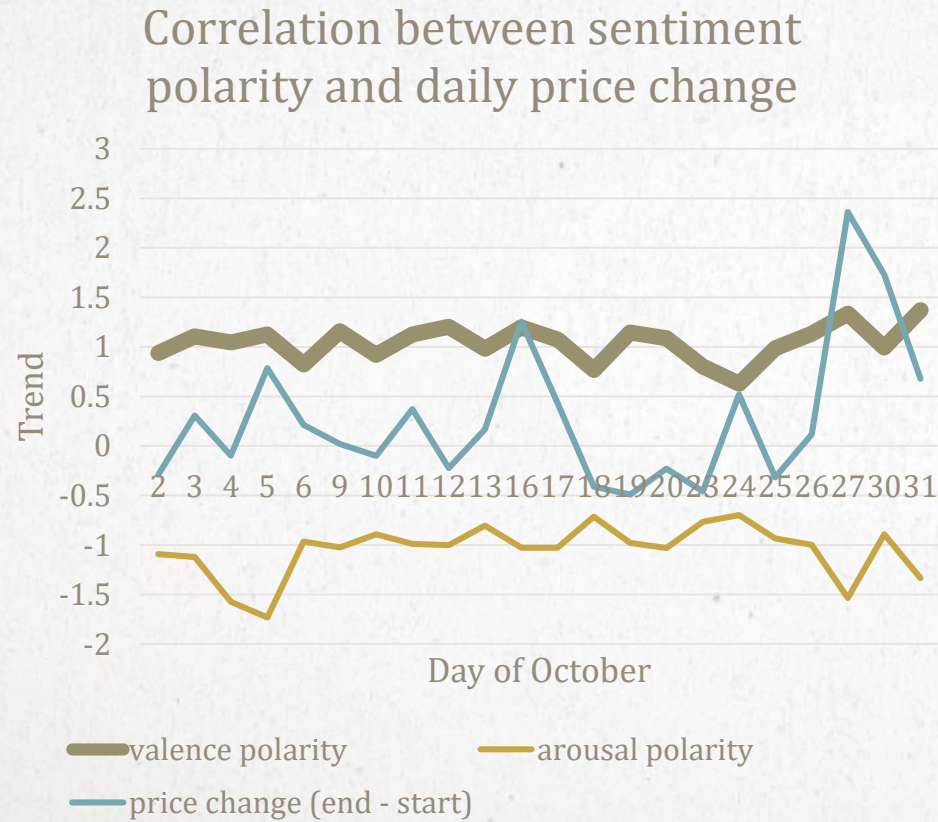
SENTIMENT COMPARE BETWEEN KEYWORD AND NOMINATION - AROUSAL

AROUSAL

- Arousal score: higher mean more active, or more forceful
- Similar to problem on Valence, less stable than using keyword

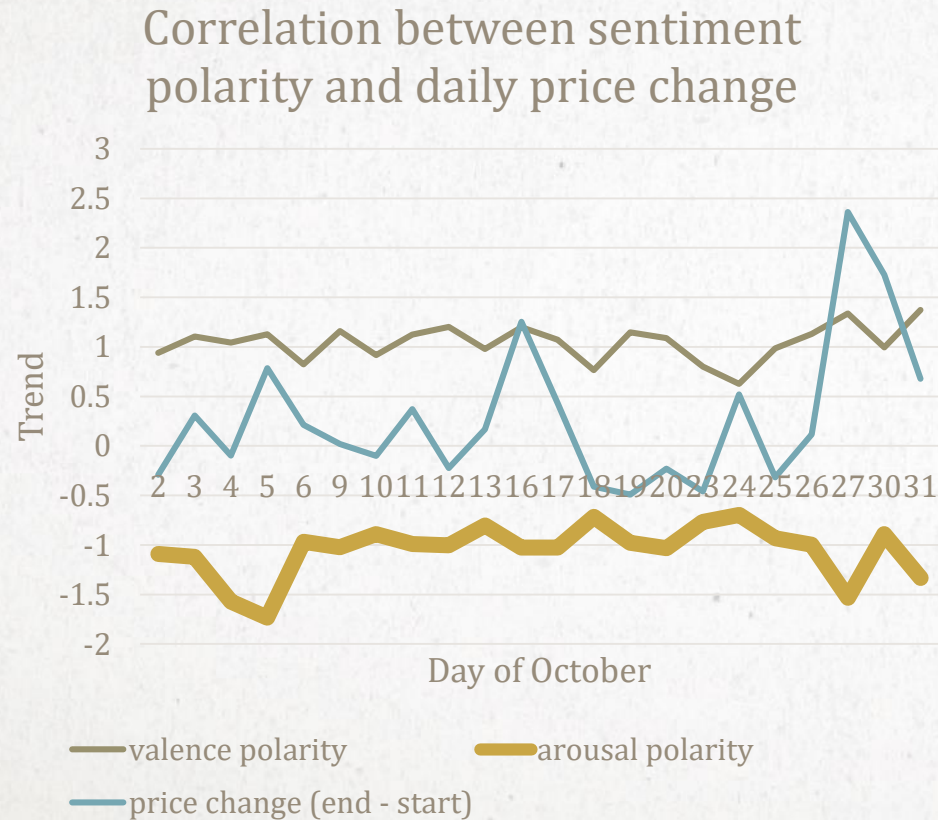


SHARE PRICE CORRELATION – SENTIMENT AND DAILY PRICE CHANGE



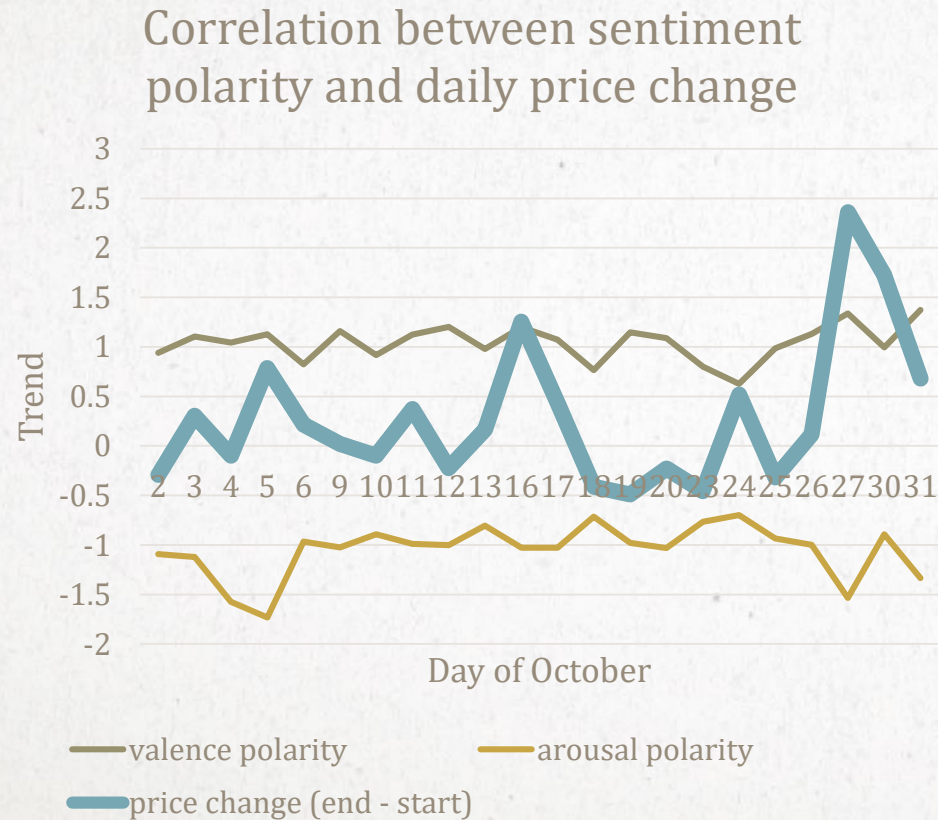
- Valence Polarity: Valence Score take away 5 (neutral feeling score)

SHARE PRICE CORRELATION – SENTIMENT AND DAILY PRICE CHANGE



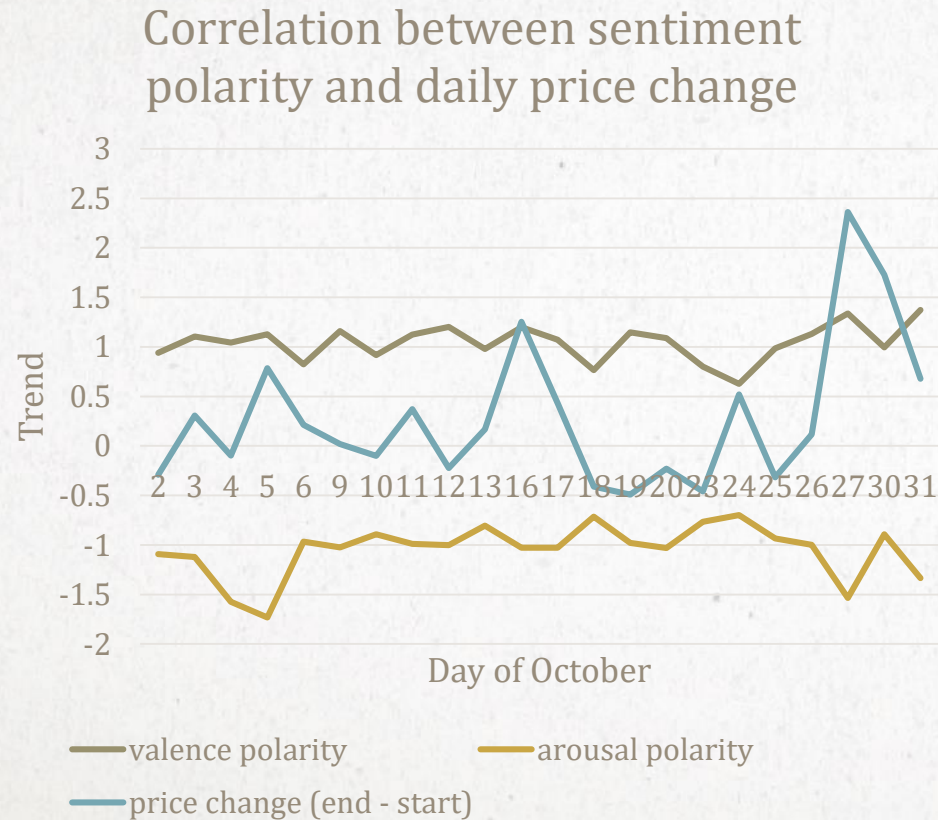
- Valence Polarity: Valence Score take away 5 (neutral feeling score)
- Arousal Polarity: Arousal Score take away 5 (neutral feeling score)

SHARE PRICE CORRELATION – SENTIMENT AND DAILY PRICE CHANGE



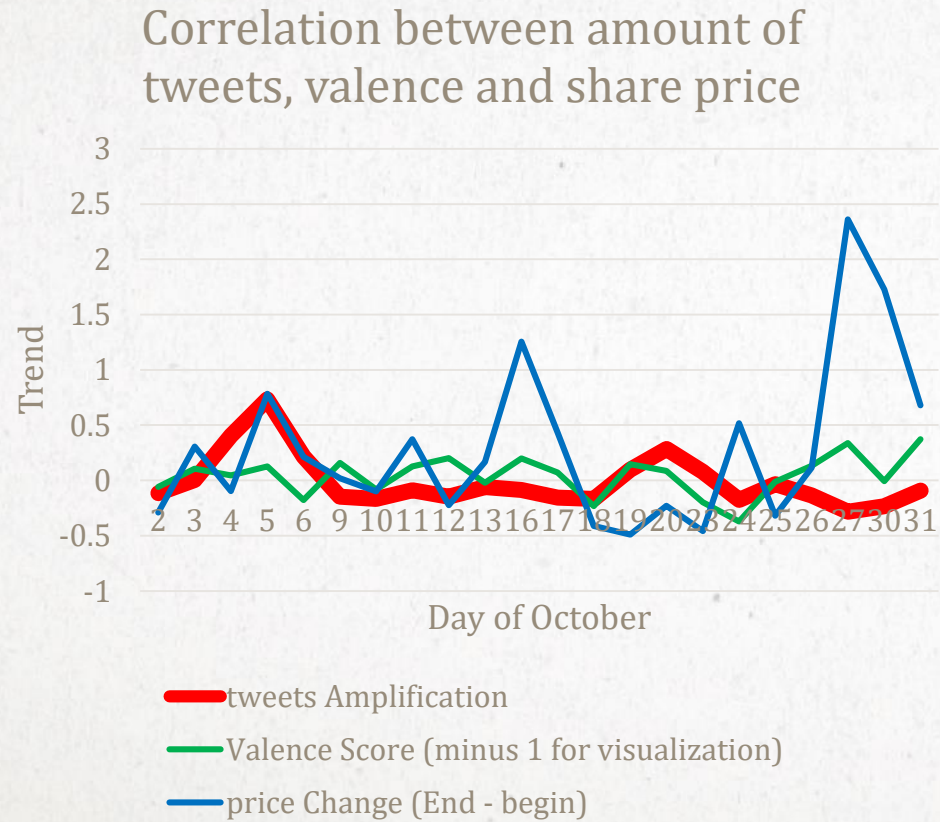
- Valence Polarity: Valence Score take away 5 (neutral feeling score)
- Arousal Polarity: Arousal Score take away 5 (neutral feeling score)
- Daily Price Change: closing price minus opening price (for example, if closing at 105.5 and opening at 103.2, daily price change is $105.5 - 103.2 = 2.3$)

SHARE PRICE CORRELATION – SENTIMENT AND DAILY PRICE CHANGE



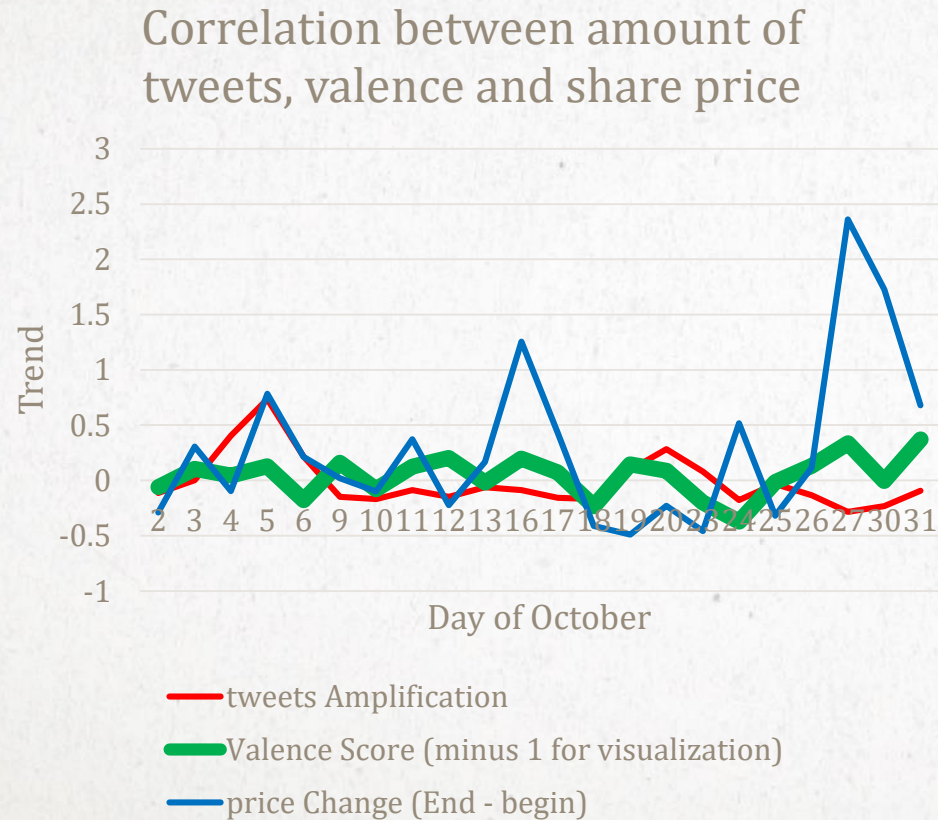
- An increase of share price could correlate to increase of valence compare to monthly average.
- Higher valence means higher increase of share price
- A long term teasing (high arousal) could also build up valence and leads to increase of share price
- That shows a feeling change from “feeling interesting” to “feeling happy”

SHARE PRICE CORRELATION – VALENCE, TWEET AMPLIFICATION AND PRICE CHANGE



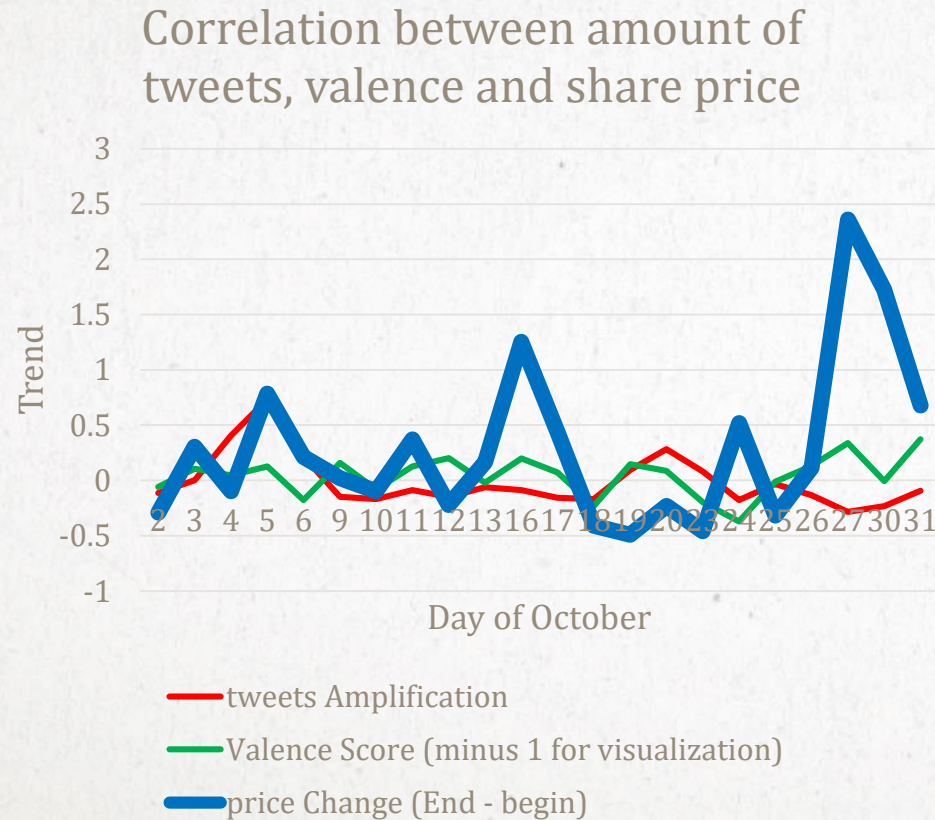
- Tweets amplification: Daily number of tweet in sample compare with monthly average. (Daily / average - 1)

SHARE PRICE CORRELATION – VALENCE, TWEET AMPLIFICATION AND PRICE CHANGE



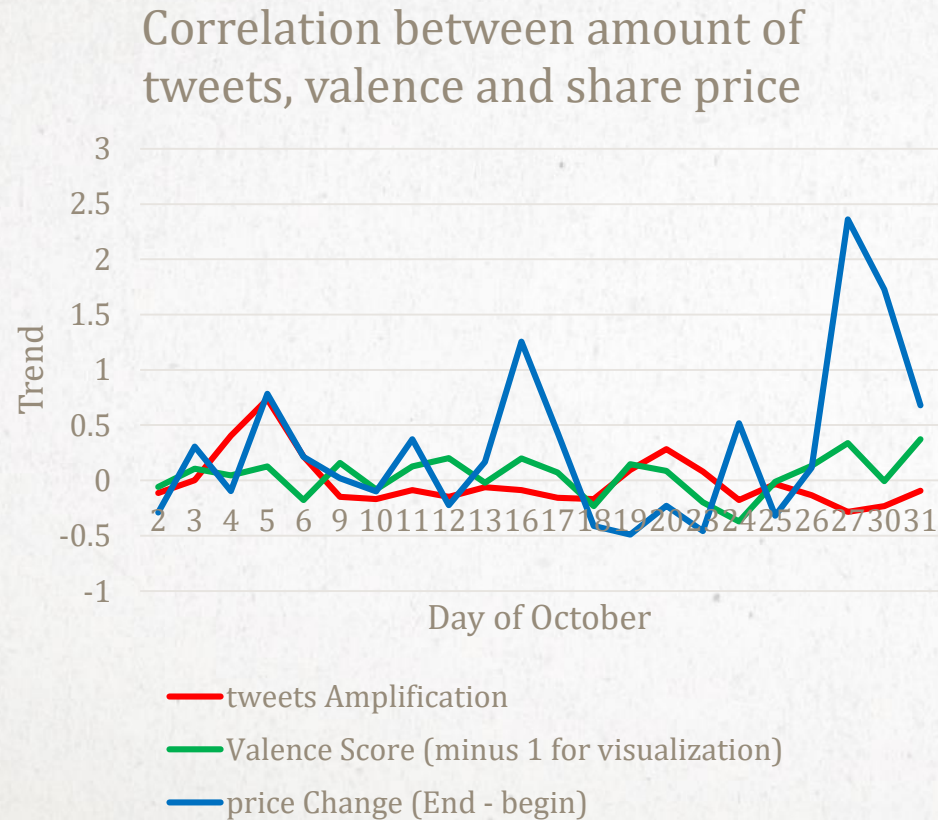
- Tweets amplification: Daily number of tweet in sample compare with monthly average. (Daily / average - 1)
- In this graph, Valence polarity has been shift down 1 for visualization

SHARE PRICE CORRELATION – VALENCE, TWEET AMPLIFICATION AND PRICE CHANGE



- Tweets amplification: Daily number of tweet in sample compare with monthly average. (Daily / average - 1)
- In this graph, Valence polarity has been shift down 1 for visualization
- The price change has remain unchanged

SHARE PRICE CORRELATION – VALENCE, TWEET AMPLIFICATION AND PRICE CHANGE



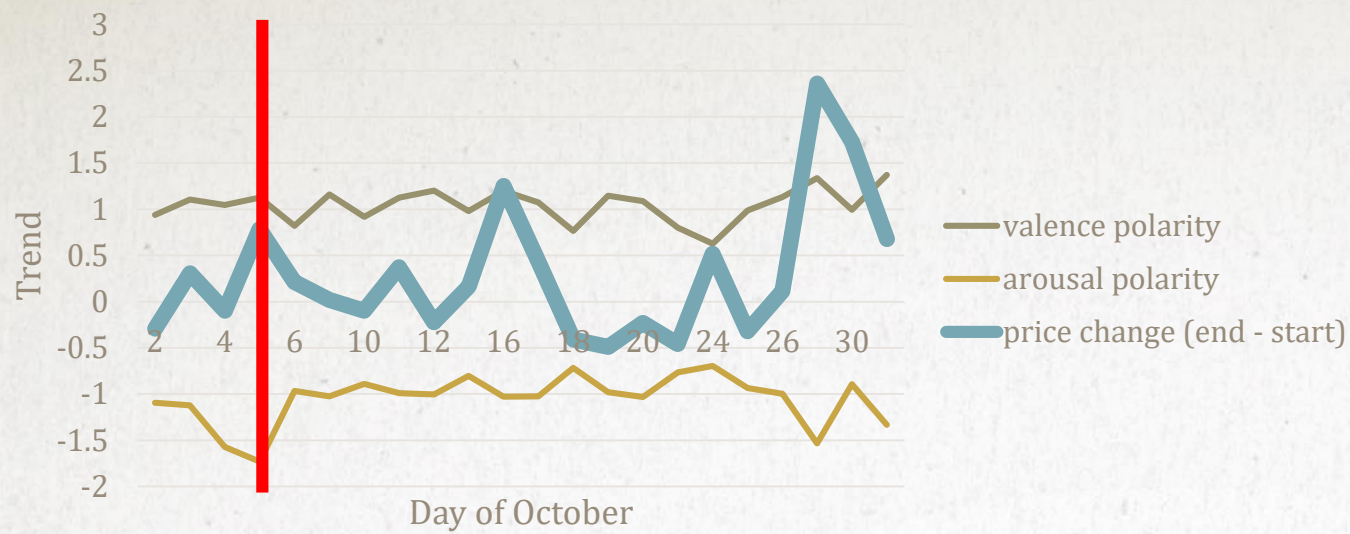
- Tweets amplification: Daily number of tweet in sample compare with monthly average. (Daily / average - 1)
- In this graph, Valence polarity has been shift down 1 for visualization
- The price change has remain unchanged
- Shows a huge discussion and high sentiment could push share price upward

FURTHER ANALYSIS

**INSPECTION OF WHAT COULD PUSH SHARE
PRICE UPWARD**

KEYWORD INFLUENCE EXAMINATION

- Attempt to extract which keyword leads to high sentiment in order to push share price forward
- By calculating product of valence polarity and weight as “Influence”
 - For example, a word with sentiment 7 and probability 0.3, its influence is 2.1

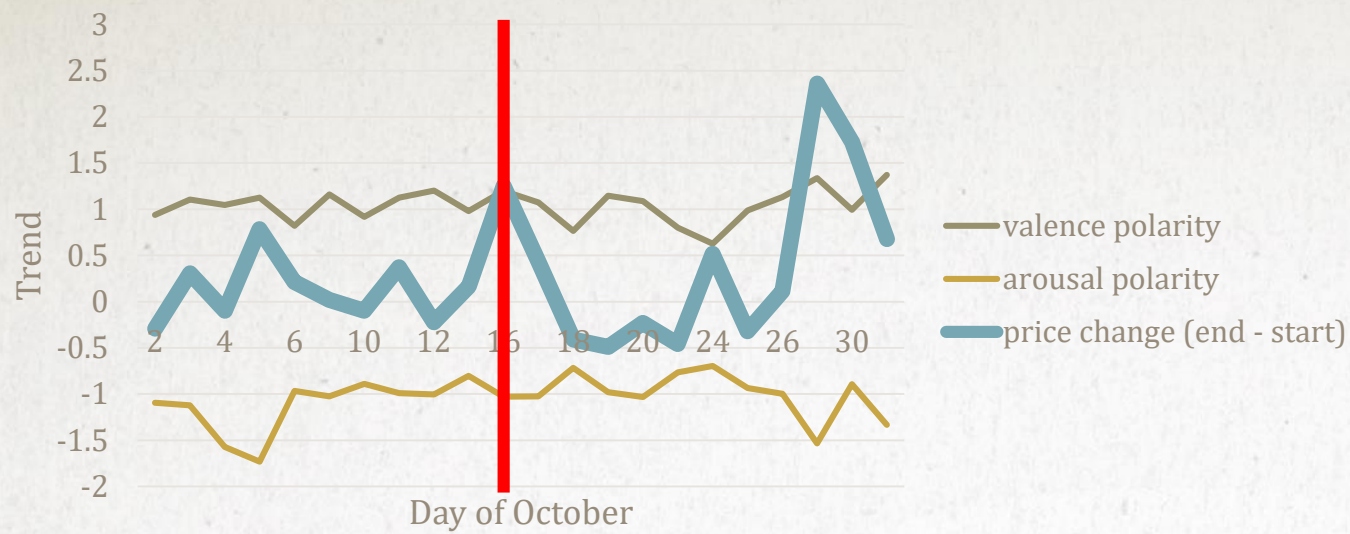


| word | weight | valence | valence polarity | Influence |
|-----------|--------|---------|------------------|-----------|
| like | 0.32 | 7.44 | 2.44 | 0.7808 |
| new | 0.294 | 7.68 | 2.68 | 0.78792 |
| know | 0.249 | 6.82 | 1.82 | 0.45318 |
| cheese | 0.194 | 6.81 | 1.81 | 0.35114 |
| sure | 0.187 | 6.64 | 1.64 | 0.30668 |
| perfect | 0.184 | 7.19 | 2.19 | 0.40296 |
| paragraph | 0.179 | 5.4 | 0.4 | 0.0716 |
| multiple | 0.178 | 5 | 0 | 0 |
| polite | 0.177 | 6.57 | 1.57 | 0.27789 |
| grammar | 0.177 | 6.25 | 1.25 | 0.22125 |

Work output from LDA model perform by stem extracted from twitter set filtered by keyword searching, on 5 October, sorted by Influence (Weight * Valence Polarity)

INFLUENCE OF WORD ON 5 OCTOBER

- Highest word: like, new
- Provides a potential new technology release information

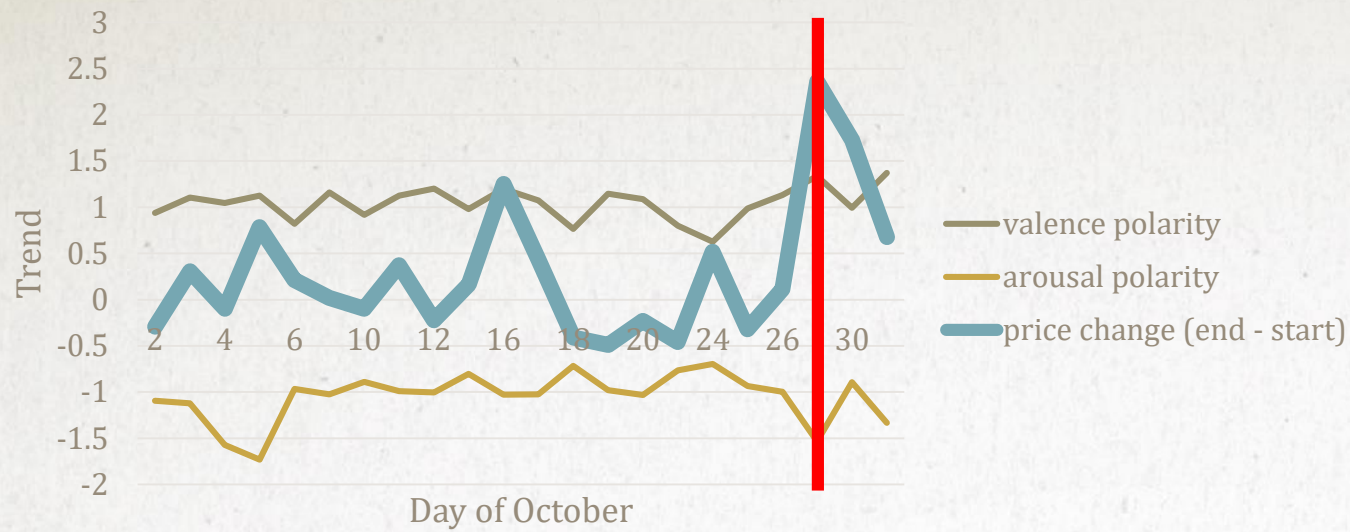


| Word | Weight | Valence | Valence Polarity | Influence |
|---------|--------|---------|------------------|-----------|
| money | 0.832 | 7.1 | 2.1 | 1.7472 |
| get | 0.687 | 6.09 | 1.09 | 0.74883 |
| love | 0.23 | 8 | 3 | 0.69 |
| one | 0.613 | 6.09 | 1.09 | 0.66817 |
| new | 0.204 | 7.68 | 2.68 | 0.54672 |
| picture | 0.241 | 6.73 | 1.73 | 0.41693 |
| keep | 0.251 | 6.32 | 1.32 | 0.33132 |
| video | 0.171 | 6.64 | 1.64 | 0.28044 |
| buy | 0.15 | 6.82 | 1.82 | 0.273 |
| know | 0.139 | 6.82 | 1.82 | 0.25298 |

Work output from LDA model perform by stem extracted from twitter set filtered by keyword searching, on 16 October, sorted by Influence (Weight * Valence Polarity)

INFLUENCE OF WORD ON 16 OCTOBER

- Highest word: money, get
- Potential exposure of financial information
- Also proves influential of investor to share price change



| Word | Weight | Valence | Valence Polarity | Influence |
|--------|--------|---------|------------------|-----------|
| new | 0.237 | 7.68 | 2.68 | 0.63516 |
| music | 0.231 | 7.67 | 2.67 | 0.61677 |
| apple | 0.233 | 6.62 | 1.62 | 0.37746 |
| free | 0.094 | 8.25 | 3.25 | 0.3055 |
| phone | 0.275 | 6.09 | 1.09 | 0.29975 |
| cheese | 0.163 | 6.81 | 1.81 | 0.29503 |
| laptop | 0.131 | 6.95 | 1.95 | 0.25545 |
| get | 0.229 | 6.09 | 1.09 | 0.24961 |
| video | 0.143 | 6.64 | 1.64 | 0.23452 |
| win | 0.119 | 6.97 | 1.97 | 0.23443 |

Work output from LDA model perform by stem extracted from twitter set filtered by keyword searching, on 27 October, sorted by Influence (Weight * Valence Polarity)

INFLUENCE OF WORD ON 27 OCTOBER

- Highest word: new, music
- Potential discussion of services provided by company

RESOLUTION

- Any discussion drags satisfaction of product, new product release, or discussion with financial enclosure has high influence due to high valence
- The word “cheese” in each list shows an improvement for keyword finding: Keyword Mac could be MacBook, could also be Big Mac, or Mac ‘n’ Cheese.

DISCUSSION

WHAT THIS RESEARCH COULD DO

- Provides a creative way to predict business model or share market trends
 - Provides another practical application to sentiment analysis, together with natural language processing
 - Provides industries/business inspirations to promote their products and strategy planning for promotion
-

VALIDITY OF RESEARCH

- Due to insufficient research resource data gathered, a further validation is required to prove applicability to multiple businesses
 - Twitter data other than October 2017 from the time research commenced (April 2018) provided by the Internet Archive are either incomplete, corrupted or out-of-date.
 - Apple Inc. has been used in this research as it is an international company which would have social media discussion globally and hence provide large amount of sample data
 - A more accurate result could be provided if there are self-analysis model for finding business-specific keyword and use for twitter extraction
-

FURTHER STEP FROM THIS RESEARCH

- Using machine learning to build models as share price prediction
 - Further analysis on different dimension of share price, for better understanding correlation
 - Trying to discover correlation involving dominance sentiment to build up 3D model
-

ETHICS PROBLEM REVEALED FROM RESEARCH

- Influential person has higher chance of share market manipulation:
 - Billionaire
 - News Media
 - Pop Star
 - Potential misleading news, fake news, or malicious offence from influential person could destroy a business
 - People should take responsibility and think twice before they expose their comment on business
-

THANK YOU