University of Nottingham Ningbo China

Faculty of Business

Academic Year 2019/20 Semester

Analytics Specializations and Applications

(BUSI4392 UNNC)

Lecturer: Dr Jenny Xiaodie PU

Students: Mingyuan Zhu

20219292

1、 Executive Summary for Business Task

It is well known that the stock market driven by the psychological and financial factors, as it's a tremendously complex and volatility processing system across the world. With many socioeconomic factors, hidden behaviour and complex models for the stock price, consequently, is hard to predict and evaluate. But such an informationally efficient is also an ethical judgment resource for investors and predictors take those into account and act accordingly with profitable market returns over a long-term, for the stock price tend to reflect all available information fully.

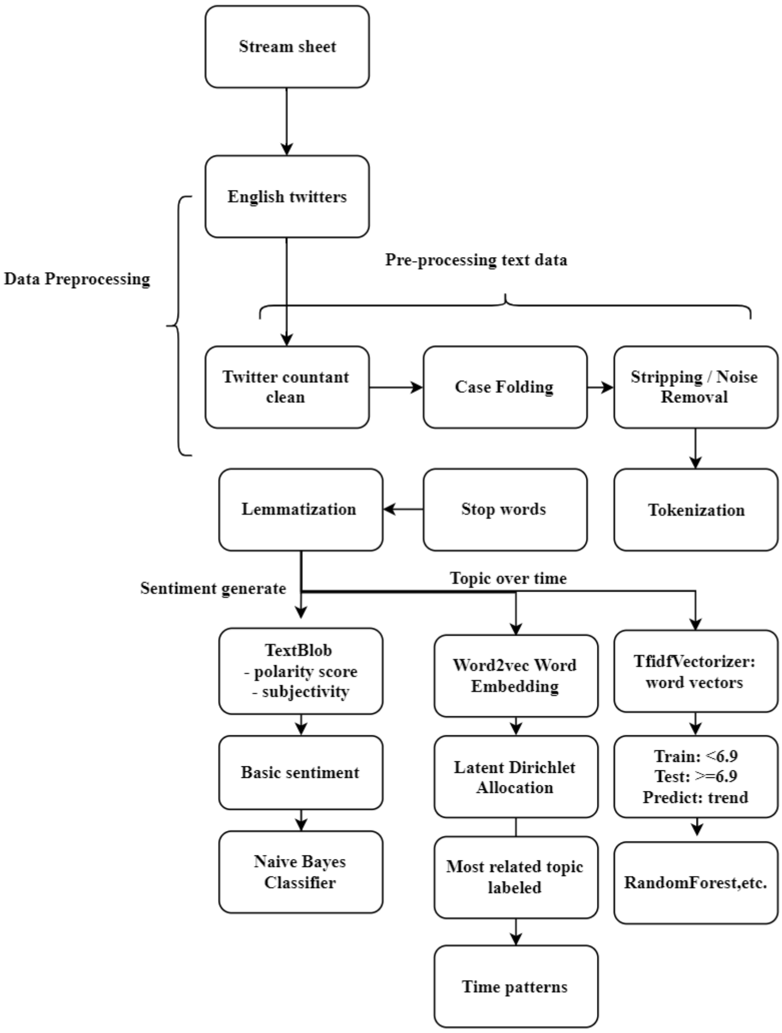


Figure 1: Structure of our research

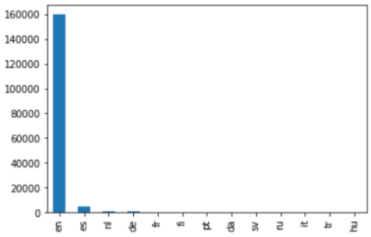
APPLE Inc. aimed to do an exploratory analysis of the social networking environments and public mood influence on its stock, like twitter. As behavioural finance emphases, investors behaviours, as a direct factors influence of stock price, is subject to particular emotions and psychological biases. Consequently, in this research, we aimed at exploring the twitters data from April 2, 2016, to June 15, 2016, to predict and understand the factors that shape investors' individual and collective behaviour.

We establish a model-based sentiment-retrained trackers and large-scale Twitter contents analyzer at short time intervals. Then we applied the resulting of sentiments and word vectors to predict stock values trends, in particular the individual's contribution for daily, with surprising results. We are surprising find that sentiments and words vectors performed well in stock prediction.

2、 Data Preprocessing

After loading the "stream" sheet from original excel files, we apply data preprocessing steps including data cleaning, editing, reduction and selections. Here we used all the English twitters data of 159864 and its date, text contents as our input.

- Data collection



Keep **English twitters**: As left figure shows, most Twitter language is English. The small part of none-English words will increase the model complex and interpretation problems. So first steps we only keep the English twitters as our analysis row input.

- **Data editing** step: After checking the data types of each feature, this step revised data formats into the correct data types. We also create the time interval and weekly time for next steps analysis by setting a reference data of April 2 and changed its format.

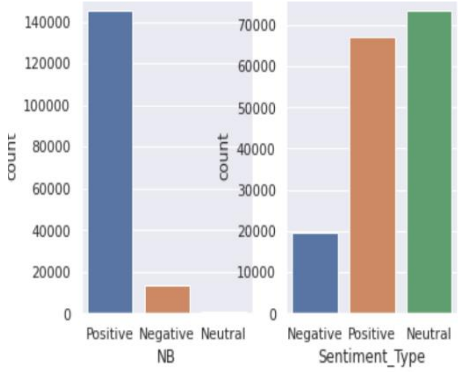
- **Text preprocessing**:

  When building twitter data related Machine Learning systems, preprocessing for its text data is essential for the subsequent analysis. The “preprocessor” is the preprocessing library for tweet data in Python programming environments and makes it easy to clean, parse or tokenize the tweets. So first Twitter contents cleaning steps using "tweet-preprocessor" package to drop the inappropriate twitter contents, like URL's, Mentions and Hashtags.

  Following steps apply case folding, stripping with noise removal, stop words dropping, lemmatization and tokenization steps. To get the tokenized text with any stopwords removed, the stop words collected from NLTK corpus must be filtered out before we processed natural language data. We only adopt the lemmatization instead of stemming for its original converting format are more friendly for human interpretation.

- **Sentiment generates:**

There could be many different methods to get sentiment judgment data, and some of the libraries provide out-of-the-box sentiment analysis function that we can directly use on the text. Still, the accuracy we manually checked is so distinguishing. We here firstly generate a rough results of our sentiments type. For "TextBlob", its sentiment property returns the tuple of polarity score[-1.0, 1.0] and subjectivity of [0.0, 1.0]. Based on these basic sentiments, here we categorize the description column into Positive(>0), Neutral(=0) and Negative sentiments(<0). For subjective, the 0.0 manes very objective and 1.0 means very subjective, so we only keep the scores above 0.6 as train dataset to pursuit better judgement results.



- **Naive Bayes for Sentiment Analysis:**

Naïve Bayes classifier based on Bayes' theorem, find the probability of how likely data attributes are associated with each class on prior knowledge of conditions, like sentiments of positive and negative. This algorithm will be performed well to predict whether a review is negative or positive, just based on the text of twitter contents.

We applied the Naive Bayes Classifier as sentiment analyzer to reset a better sentiment. As small data can get a better result, here we use 2000 items as our training set and 300 as our test set, and finally got the accuracy of 77.6%. The raining results performed much better in more sensitive to the positive and negative(most twitter is to share the daily happy, and bad emotions may be dropped by twitter automatically).

3、Data Exploration

- **The stock of APPLE:**

The APPLE stock data was collected from the application programming interface of yahoo finance. As candlestick charts shows, for its total trend, its shares slumped about 22% from 109 on April 16 to 90 on May 13, and growing trend until May 30. Then it fluctuated to June 13, which shows this is quite a common trend and its stock get steady.

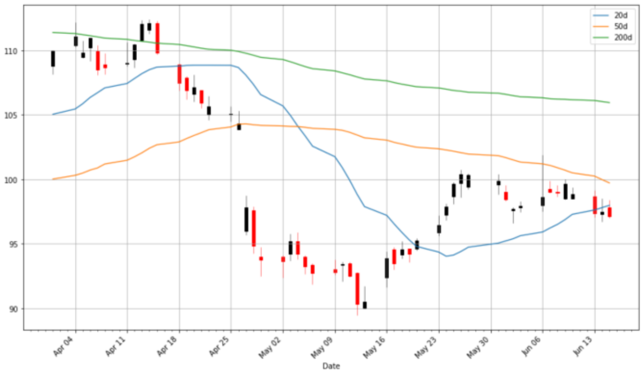
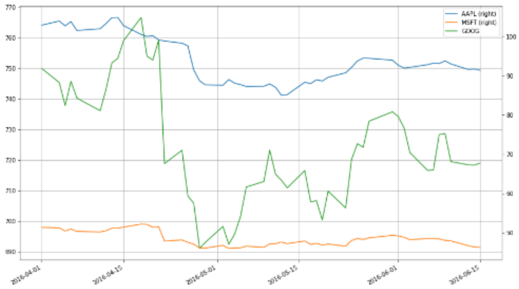


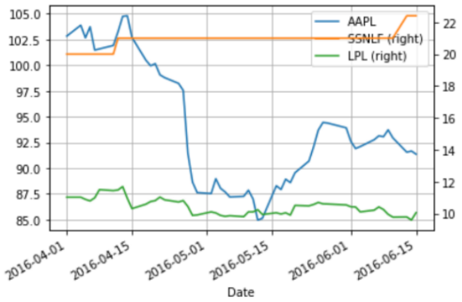
Figure 2: Candlestick charts for apple stock

- **The stock of other companies**

We also compared other high-technique companies like GOOGLE and MICROSOFT and evaluated whether the total market of high add-valued is dropped. It shows that GOOGLE performed the same drop trend during this period, which may indicate that most people may not be favourable to high-technique at this time.

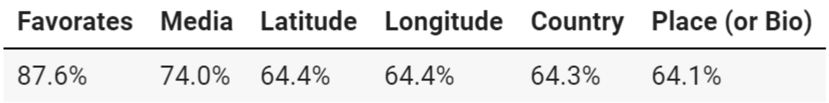


How about other smartphone companies? We pick the top 3 American phone companies of Samsung and LG and find both of them is steady for its stock. It seems that for those top smartphone companies, only apple company suffer a huge cliff during this time.



- **Null values**:

The features null value count will influence the analysis features we picked, for lack of too many things will bring little values in analysis. Drop the null values may be the not wise decision here. Although map information like longitude has more than 64% null values, gapping by model or neighbours is not wise for its hard to inferred and meanless.



Favorates	Media	Latitude	Longitude	Country	Place (or Bio)
87.6%	74.0%	64.4%	64.4%	64.3%	64.1%

Figure 1: Top count for null values features

- **Date:**

When something of importance happened, people are more willing to talk about it on social media. The topic of hot and daily active users can be inferred by the count of twitter each day. The more active, the more exposure for search engines and twitters followers the APPLE will get. As figure 3 shows, twitter counts seem to performed a seasonality pattern periods of two weeks as a hidden behind time series.

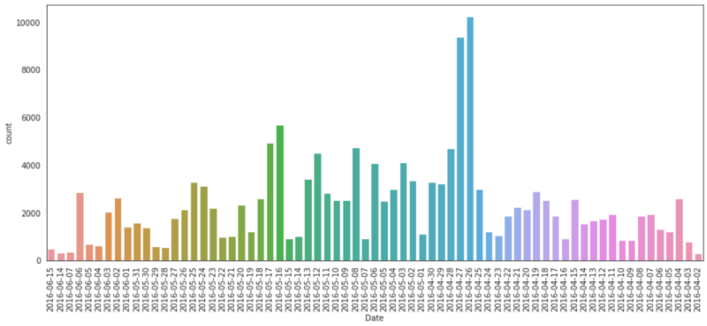
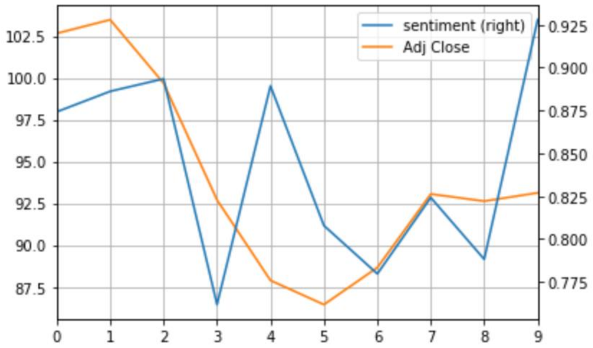


Figure 3: Twitter counts over time

- **Lexicon Sentiment** Over Time

We also compared the lexicon sentiments for each week with the stock price. The weights for sentiments using the common senses each week trained by classifier before. The influence of emotions has the hysteresis effect for postponed one week, and less positive attitudes will be observed the stock price cliff soon like week 3(4-11) shows.



-

Figure 4: Lexicon sentiment and stock price relationship

4、Topic modelling overtime

This steps we combing **LDA** and **Word Embeddings** for topic modelling to compare the different latent variables for each time. Word2vec, as a highly popular word embedding methods, is applied to get semantically words correlated and better LDA model performance. Its pre-training steps capture the unsupervised from unsupervised twitters by word vectors.

Then the topic vectors are learned in this word space and put the constraints on its weights vectors. All twitters contents trained the LDA model with the topic number of 20.

We also evaluate topic models manually by changes the topic numbers in **perplexity**, as it indicates the probability model predicts a sample. The minimal perplexity score of -13.3 indicates an excellent generalization performance among different parameters.

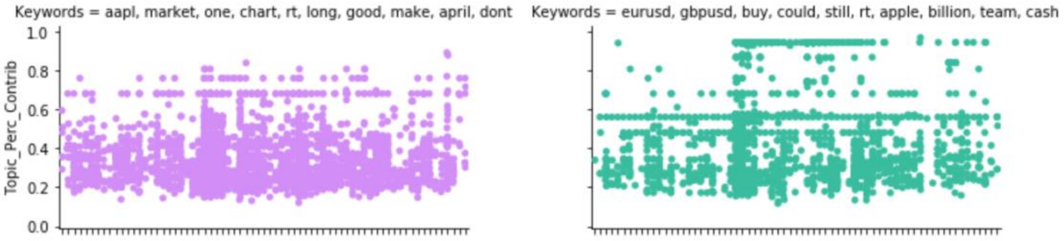
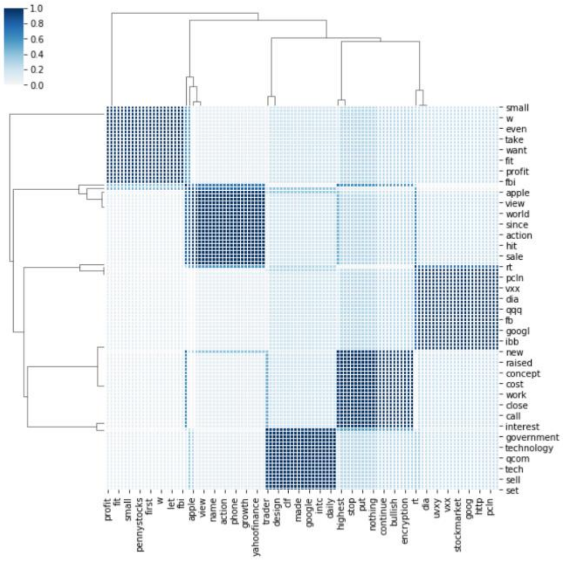


Figure 5: Subplots for topic 17 and 18



After labelled the most related topic and weights for each data, for twitter's topic weights more than zero columns, we plot the topic count for each time. We find that of them performed strong time-related patterns, like topic 18 about "APPLE" stock growing trend and drop intervention policy has strong seasonality. During the stock cliff week, the topic shows strong fluctuations in adjacent days

How about the average of topic weights per year？ We plot the weekly topic trend and stock to evaluate its direction. The most representative document also helps us to understand each topic further while just the topic keywords may not be enough to make sense. The first trend we can see is its complaints about the stock drop. Like topic 3 and 4 of "Accused of over-predicting Apple's stock price" on April 25, which is also the massively drop one week in 4.18, they increasing a lot and indicates disappointment for apple stock. The earns and its shares drop is far away from the analyst expectations, while more and more investors choose to stock selling. The large quantities of stocks selling results the more excellent supply than demand and the price would fall. That gap between APPLE's public prediction and real revenue is wide enough to scare away many investors.

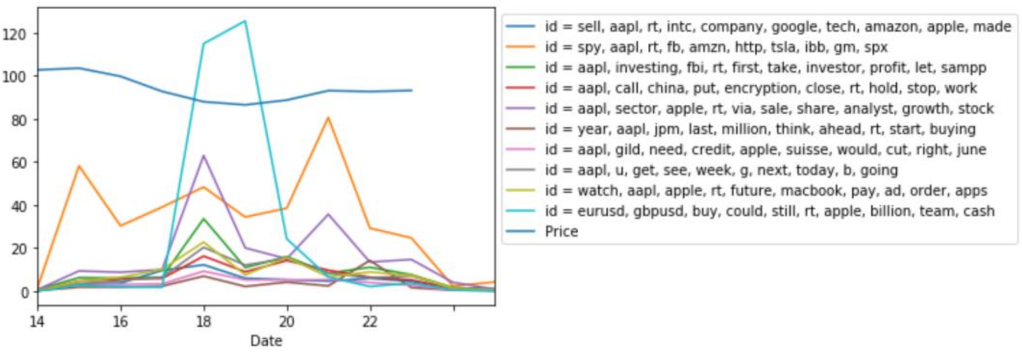


Figure 6: the topic trend from 1 to 9

Meanwhile, there is another trend that consists of stock steady growing after cliff descent in week 20, and it would be the attitude increase and its policies. Like topic 2, people talk more about the others stock's drop trend and hold the optimize for apple and hope their share buybacks soften the blow on a per-share basis and diluted the stock drop.

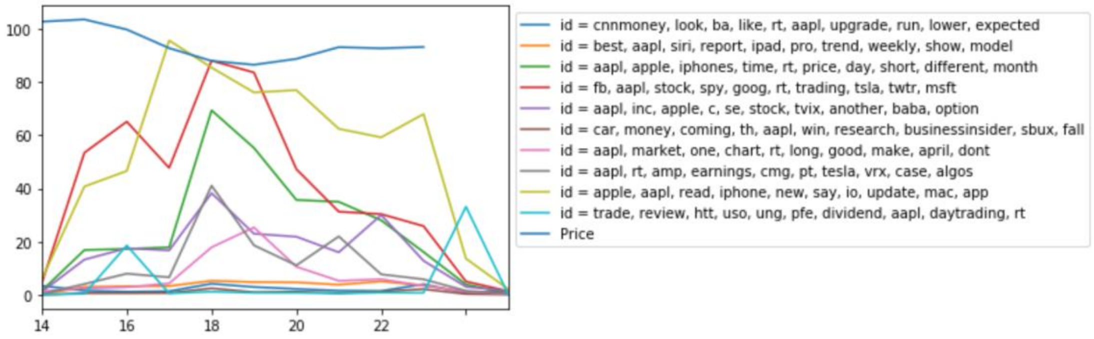


Figure 7: the topic trend from 10 to 20

How's the topic for iPhones? The stock drop has many reasons. As APPLE's second-quarter report shows, the sales drop in the iPhone may be the mainly driven of steep revenue decline and weak guidance. The decline in unit price in the second quarter from 691 to 642 may be the critical point of its lower revenue. Sales of iPods and Macs have also performed a weak pattern during the second quarter. This trend indicates us to more focus on the customers' attitude for its products. Even, the wireless carriers, such as Verizon, are killing off their hardware subsidies, which increase the cost of iPhones. Apple applied for some price promotions and tried to fight this not updating new phones effect in customers.

For topic 13 and 19, which is related to "iPhone and iPad, Mac". Topic 13 also performed growing pattern during the stock drop, but never increased after that. They mostly talk about how bad the new iPhone, why they are not good enough. This lousy mood can explain the problem of low sales of new phones. In contrast, topic 19 mainly talk about the suitable environments of its apple software also increase with the stock trend—the confidence of apple and its future developments may be the critical point for more investments.

5、 Text models

- Tfidfvectorizer model trained:
  The word dictionary generated from Tfidfvectorizer after prepared by all twitter content. The vocabulary building steps ignore the words have a document frequency strictly higher than 0.5 and lower than 3. We only keep the 22181 words in our dictionary.
- Generate the words vectors: For the words in a word dictionary, we assigned the index in the text. Our train data is twittered before June 9, while other data as our test data.
- The predicted label was the trend compared with the last date, for "-1" represents the dropping trend, "1" for growth and "0" for stock market stopped and steady.
- GBDT: The features which we extracted and used from the text vector data as in input was from gradient boosted decision trees (GBDT) selections. This feature selection models performed well and not prone to overfitting. Here we only keep the features with GBDT important scores more than 0.0001 and drop more than ten words vectors features with low variance
- Model training: Here, we pick the models of RandomForestClassifier, Adaptive Boosting Classifier and Gradient Boosting. The Random forest performed well in acc. 0.6102117061021171 Recall: 0.6102117061021171 Specificity: 0.6102117061021171 Balanced: 0.5031746031746032

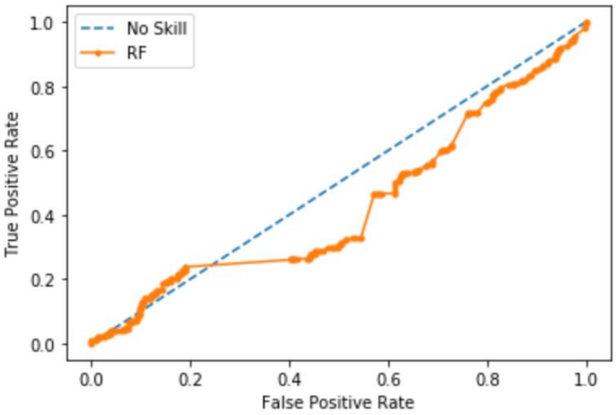
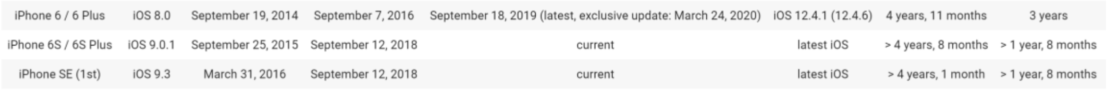


Figure 8: ROC curve for random forest

6、 Conclusion

By applying the intraday perspective on our twitters data and stock price, our research analysis the dynamics stock changes, Twitter sentiment and its contents activity. After accounting for the adjusted closing price, we find some statistically significant co-movements of intraday volatility and information from stock-related and apple-related Tweets for all constituents of the APPLE stock. However, economically, the special effects are of negligible magnitude for analysis features in stock. From a practical point of view, we find that high-frequency topic is not particularly useful for assessment and forecasting for the trend of the capital.

What caused the cliff of APPLE stock? The far below the expectation of iPhone SE^{1st} is the critical point. As the most popular products, iPhone 6 and its plus unconsciously set a high bar for iPhone sales, which first time apply the more minor upgrades. The iPhone 6s and also match the same magic in the upper minors and more robust software. But the innovation just can be workable at one time, while the iPhone SE^{1st} have trouble in matching predecessor's success. Only the larger-than-ever forms can satisfy their need



iPhone 6 / 6 Plus	iOS 8.0	September 19, 2014	September 7, 2016	September 18, 2019 (latest, exclusive update: March 24, 2020)	iOS 12.4.1 (12.4.6)	4 years, 11 months	3 years
iPhone 6S / 6S Plus	iOS 9.0.1	September 25, 2015	September 12, 2018	current	latest iOS	> 4 years, 8 months	> 1 year, 8 months
iPhone SE (1st)	iOS 9.3	March 31, 2016	September 12, 2018	current	latest iOS	> 4 years, 1 month	> 1 year, 8 months

Table 2: History and availability of iPhone from 2014 to 2016 from Wikipedia

7 、Further analysis direction:

We notice that our word vector model is not robust to explain the stock, even the daily changes of stock(three class). We ignored too many information's.

How to make it better? First, we should carry on the **social network analysis** instead focus on each individual contributes. Except for celebrities., the micro-influencers may be more relatable and trustworthy to influence the audience. Micro-influencers like a food blogger and fitness guru, are those individuals who have more followers and are considered experts in their respective niche, like the data of top follower and following given in the table. As for influencer marketing, business and high-tech companies will benefit more from this.

Second, we should add **other language** information, like trending in the world with more than 70 mainstream languages. It's because except for English native speaking countries, Asia accounts more and more great sales propositions, especially for China. For this cliff of stocks, growth in China is slowing down big time. China is Apple's second-largest market and a source of massive potential growth. The company's revenue grew 14% in the quarter before this drop and a staggering 99% in the quarter before that. This cliff time around; however, Apple's growth in China was negative, dropping 26%.

The third is **personal information**. When you consider someone's talks, it's hard to ignore their background. What's the tag for this person, social freak or cat liker? Does his work have enormous social influence, like a writer or Newspaper reporter? The right tag and personal info will help us add more values on identifying who is real influencers, or at least different weights for them.

Fourth is more robust **sentiments models**: the more advanced models and divided dimensions is the key point to success. For GPOMS, it tracks the mood in six aspects (calm, alert, sure, vital, kind, and happy), while OpinionFinder can classify texts in terms of their positive versus negative sentiment.