# Challenges Presented in Social Media Mining

**Yoong Pei Xian 100334770**
**Data Mining**
**Word Count: 1940 words**

## 1 Introduction

Social media mining is an emerging field that its application has expanded across various industries to allow organizations to embark on solving complex problems in an innovative way. It extracts data to analyze the trends, patterns, or correlations from the raw data on social media platforms such as Twitter, Reddit, and Facebook where the public express their opinions and views. Thus, public sentiment can be easily captured to forecast the future event, resulting in the adoption of sentimental analysis to have skyrocketed in recent years. One decent example is the use of sentimental analysis in algorithm trading by investors to predict the stock price. Looking at the robustness of social media mining, researchers have developed various data mining models to maximize the performance of social media mining.

Social media data is characterized by its high-volume, high-variety and high-velocity (Laney 2001, GartnerResearch 2012) that requires innovative forms of processing to discover useful insights. Due to its nature, several challenges can be imposed on its mining process. In this paper, a literature review is conducted by studying several of the most well-presented papers to explore the challenges presented in social media data mining in three aspects: data collection, processing, and evaluation.

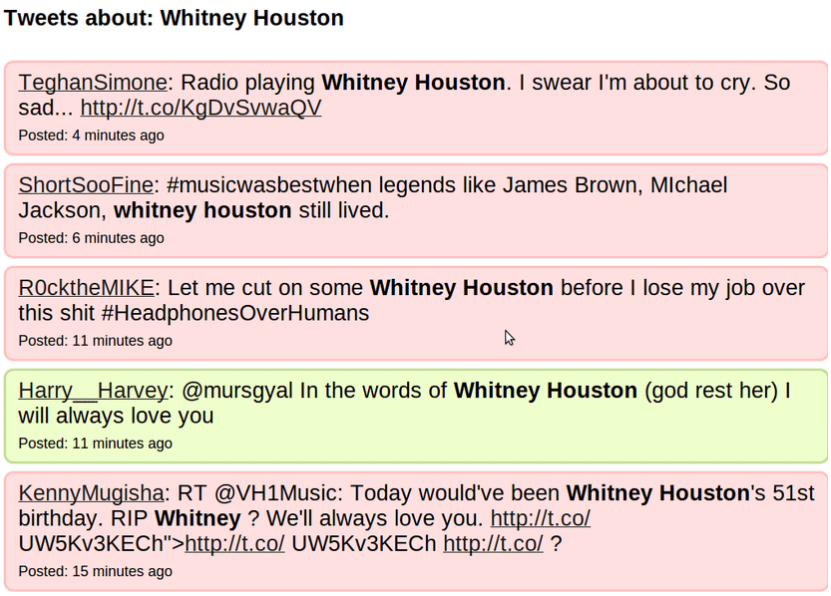
## 2 Challenges Presented in Social Media Mining

### 2.1 Data Collection

#### 2.1.1 Relevancy

The divergence of posts into unrelated topics is a common issue on social media platforms, where an irrelevant post, such as an American football match can appear on the U.S. election threads just because of the hashtag #american (Maynard et al. 2012). Retrieving data according to the topics or pages does not necessarily guarantee the relevance of the data obtained. To resolve this challenge, Maynard et al. (2012) suggested training the classifiers for the relevant tweets and then adopt clustering algorithms to group the relevant tweets together and they found the result to be promising.

The correlation between the search keywords and the sentiment of the posts might also be ambiguous despite the correlation test. For instance, the sentiment analysis appeared to be negative the following day after Whitney Houston's death, but it did not mean that the users dislike Whitney Houston but being sad for her death (Maynard et al. 2014).




By adopting an object-centric approach, which identifies the object first before looking for the posts semantically relevant to the object, it reduces the chances of misinterpretation (Maynard et al. 2012). This is the limitation presented in a substantial proportion of the existing papers where the algorithms are built to only identify the sentence as "sentiment-containing".

### 2.1.2 Usability and Reliability

Given the massive and continuous streams of social media data from multiple sources, it is difficult to identify the usability of the data that can provide values for decision making process. Using multiple sources of data can also complicate the overall performance in a similar fashion. Xiang et al. (2017) conducted research into reliability of social media data by examining the accuracy of text classification of hotel reviews collected from TripAdvisor, revealing issues that travelers tend to misclassify their travel purposes even on the highly credible sources, leading to unwarranted conclusions. Taking Banerjee & Chua (2016)'s research for instance, their studies of travelers' hotel ratings pattern that requires the data indicating the travel purpose (business, couple, family, friend and solo) might not yield a valid result as there was not a verification step taken to ensure that all travelers labeled their trips correctly.

Apart from this, the population's distribution such as the demographic data is always unknown for social media data, highlighting the questions whether they are reliable representatives of the full data to identify their relevancies in certain subjects. To emphasize the importance of these information, Ikeda et al. (2013) have developed a demographic estimation algorithm to infer the demographic of Twitter users based on variables such as their tweets, followers and following.

## 2.2 Data Processing

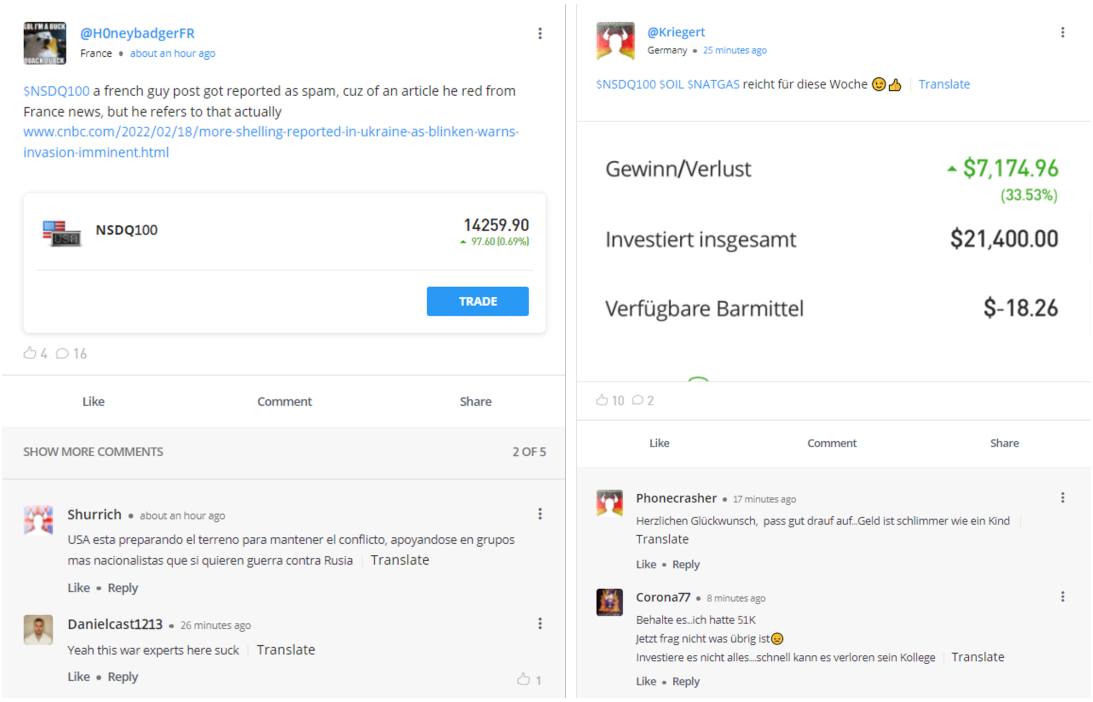
### 2.2.1 Noise Removal Fallacy

Data preprocessing is required in any data mining application to improve overall accuracy. Due to all the clutter, social media data is usually unstructured and noisy with the occurrence of repeating letters, exclamation marks, or capital letters to emphasize the words (Chen et al. 2014). Studies conducted without clearly stating the data-processing steps often prompt doubt on credibility. This makes data-processing exceptionally important as "garbage in and garbage out." Nonetheless, the handling of noise can also be challenging as it is a subjective quantity and can always be confusing, so blindly removing noise can also eliminate valuable information and introduce more errors in pattern recognition. For

instance, case standardization might alter the emotion intensity using case-sensitive tools such as VADER (Valence Aware Dictionary for Sentiment Reasoning).

### 2.2.2 Multilingual Issues

The presence of language variation among the social media posts can be problematic when performing the data processing as different models have to be created to deal with the different languages (Maynard et al. 2012). The image below illustrates two post extracted from Etoro, a trading platform, shows 2 different languages in a single post.




Currently, most of the existing studies focuses on English words and the analytic tools adopted are also in English. Some researchers discovered that the performance of algorithms varies on different languages and the accuracy tends to decrease significantly when there is a huge variation between the language distribution of the training and testing set, depending on the algorithms being used. Gamallo et al. (2014). Therefore, it is crucial to handle the language processing before feeding them into the model. For instance, Maynard et al. (2012) proposed the use of GATE plugin for the TextCat language identifier to identify the language of every sentence. However, they found that the short length of tweets (less than 140 characters) makes the language identification tough, especially special characters such as @, , emoticons, URLs are presented in the tweets.

## 2.3 Data Evaluation

### 2.3.1 Degree of Result Validity

The evaluation of social media context is often conducted in the missing of ground truth to guarantee the validity of the patterns. There are no existing patterns or trends that can be followed, thus presenting insurmountable challenges to organization. Zafarani et al. (2014) proposed the use of semantics or clustering quality measures when evaluating results without ground truth. Semantics refers to looking for coherency from attributes such as demographics and content generated among the target communities while clustering quality measures computes the values such as sum of squared errors to compare the performances between two algorithms with the same target communities.

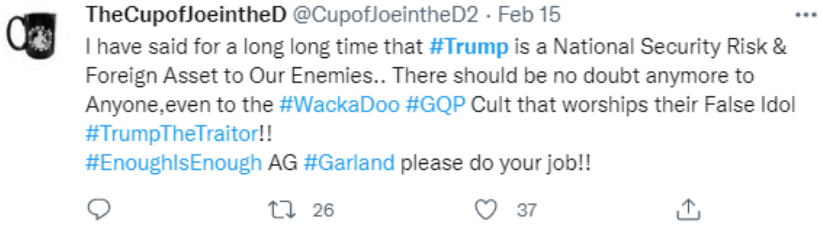
Besides, (Chen et al. 2014) criticized that the meaning of social media content can often be ambiguous and subjective due to the presence of vast number of acronyms, misspellings, informal language and sarcasm. This might result in faulty assumptions, supported by Rost et al. (2013). This is particularly true for microposts such as tweets that do not contain much contextual information as they do not follow any conversation thread except for the use of hashtag, forming an isolation among tweets of the same topics (Maynard et al. 2012). Algorithms still struggle to understand these ambiguities. For instance, Poria et al. (2016) identified that sarcasm is topic-dependent and contextual. Looking at the image below, the post implies a positively happy emotion when Bitcoin just hit $10000 4 years ago, but it might be a sarcastic statement if the post were posted in 2021 when the Bitcoin hit $60000, showing the investors' negative outlook on the price.


Algorithms will require much more additional information, and also pre-trained word embedding and personality models to achieve this. They proposed the use of convolutional neural networks to address the issue.

### 2.3.2 Representation of the Sentiment Score

Sentiment analysis that assigns posts with polarity score between –1 and 1 does not imply how accurate the score is. For example, it is impossible to tell whether a score of 0.701 awarded to a tweet in Donald Trump forum indicates that the person has a higher inclination of voting him in the next election compared to a tweet with score of 0.453. What can be drawn from the scores is only the confidence level, or the strength of emotions.

Furthermore, it is worth mentioning that most of the studies on social media data analyzing focused on classifying words into broad categories, for example, positive or negative, rather than slicing them into a narrower sub-category such as emotion-based anger, fear, joy, sadness. Junianto & Rachman (2019) criticized that a 'positive' or 'negative' could represent several different emotions and it is crucial to capture the exact emotions. Taking the tweet trump below, it is possible to identify it as a negative tweet, but machine learning will have challenges implying whether the author is showing confident that his statement is that Trump is an enemy is finally verified, or he is expressing his anger or disappointment towards people's admiration of Trump.




This is particularly important for organizations, such as those in the retail industry, to recognize how satisfied their customers are, whether they are simply being positive, happy, or love the products.

# 3 Summary

The purpose of performing data mining using social media data is to extract accurate and reliable information to help organizations achieve their objectives. Based on the literature review, social media data, which remains to be an untapped knowledge source, presenting multiple challenges to be addressed:

- Uncertainties about the relevance, usability and reliability of the data extracted to correctly represent the target community.
- Informal environments encourage people to use colloquial and personal elements, which makes it hard for data cleaning, as there is no proven method to perform it the right way.
- Missing ground truth to ensure the result validity and the representation of sentiment score.

With the advancement of machine learning, researchers are improving the algorithms to handle the complexity of social media data, as several methods have been discussed in this paper to address the challenges. Personally speaking, the greatest challenges still lie in the identification of the right source of data as it is the human's decision on which source of data to be extracted. Facebook might be a more reliable option due to the availability of the demographic, but there come the questions of ethical issues. Future works can focus on the following aspects: I) to research into the demographic estimation algorithm to sample the data correctly, ii) to improve the social intelligence to detect the sarcasm part, and iii) to develop analytical tools that comes with language variations.

## References


Banerjee, S. & Chua, A. Y.-K. (2016), 'In search of patterns among travellers' hotel ratings in tripadvisor', *Tourism Management* **53**, 125–131.

Chen, X., Vorvoreanu, M. & Madhavan, K. (2014), 'Mining social media data for understanding students' learning experiences', *IEEE Transactions on Learning Technologies* **7**(3), 246–259.

Gamallo, P., Garcia, M., Sotelo, S. & Campos, J. (2014), 'Comparing ranking-based and naive bayes approaches to language detection on tweets', *CEUR Workshop Proceedings* **1228**, 12–16.

GartnerResearch (2012), 'The importance of 'big data': A definition', https://www.gartner.com/en/documents/2057415. [Accessed on 8 February 2022].

Ikeda, K., Hattori, G., Ono, C., Asoh, H. & Higashino, T. (2013), 'Twitter user profiling based on text and community mining for market analysis', *Knowl. Based Syst.* **51**, 35–47.

Junianto, E. & Rachman, R. (2019), Implementation of text mining model to emotions detection on social media comments using particle swarm optimization and naive bayes classifier, *in* '2019 7th International Conference on Cyber and IT Service Management (CITSM)', Vol. 7, pp. 1–6.

Laney, D. (2001), '3d data management controlling data volume velocity and variety', https://idoc.pub/documents/3d-data-management-controlling-data-volume-velocity-and-variety-546g5mg3ywn8. [Accessed on 8 February 2022].

Maynard, D., Bontcheva, K. & Rout, D. (2012), 'Challenges in developing opinion mining tools for social media', *Proceedings of @NLP Can U Tag User-generated-content?! Workshop at LREC 2012* .

Maynard, D., Gossen, G., Funk, A. & Fisichella, M. (2014), 'Should i care about your opinion? detection of opinion interestingness and dynamics in social media', *Future Internet* **6**, 457–481.

Poria, S., Cambria, E., Hazarika, D. & Vij, P. (2016), 'A deeper look into sarcastic tweets using deep convolutional neural networks', *Proceedings of COLING* .

Rost, M., Barkhuus, L., Cramer, H. & Brown, B. A. T. (2013), 'Representation and communication: challenges in interpreting large social media datasets', *Proceedings of the 2013 conference on Computer supported cooperative work* .

Xiang, Z., Du, Q., Ma, Y. & Fan, W. (2017), *Assessing Reliability of Social Media Data: Lessons from Mining TripAdvisor Hotel Reviews*, pp. 625–638.

Zafarani, R., Abbasi, M. A. & Liu, H. (2014), *Social Media Mining: An Introduction*, Cambridge University Press.