Twitter Sentiment Analysis Using VADER

Nurul Ariessa Binti Norramli

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak

Abstract

Twitter is known as a platform for opinion sharing through posting tweets, retweeting, or liking tweets. As such, I used VADER (Valence Aware Dictionary for sEntiment Reasoning) to perform sentiment analysis on texts gathered from Twitter. VADER is a simple rule-based model that was created specifically for sentiment analysis on social media. This is because it takes slangs, short words and emoticons into account.

Introduction

In a data-driven world, sentiment analysis is extremely useful as it allows a wider public opinion about certain topics. Sentiment analysis or opinion mining is the process of computing and categorizing texts into their respective polarity: positive, negative or neutral.

For this project, I used a dataset of tweets taken from public Twitter profiles. The tweets extracted were those that contains the word 'HelloGold' with the month and year posted in September 2019. The total number of tweets taken were capped at 100, which is the maximum number of tweets allowed for Twitter Premium 30 Days API. After searching for tweets, the data were structured in a pandas DataFrame for easier manipulation.

	Author	Content	Date	Url	Followers
0	miha_miharu	ดาวน์โหลดแอพ HelloGold แล้วซื้อทองค่า มูลค่า 500...	Wed Sep 18 02:18:33 +0000 2019	[{'url': 'https://t.co/deeUEgXbNw', 'expanded_...	27
44	14ottobre2010	Molto interessati? A casa loro, le stesse macc...	Wed Sep 11 11:24:55 +0000 2019	[{'url': 'https://t.co/8UqTaeynpA', 'expanded_...	673
45	hazimdace	@myhellogold Muat turun aplikasi HelloGold. Bu...	Wed Sep 11 00:23:33 +0000 2019	[{'url': 'https://t.co/5dYr33ySbh', 'expanded_...	17

Figure 1. Example of dataset

Preprocessing

The tweets contain texts that has slangs, short words, and emoticons. Therefore, it needs to be preprocessed into a more understandable format. In this project, the column 'Content' was preprocessed as follows:

- Remove twitter handles
- Remove links
- Remove special characters, numbers, punctuations (except exclamation mark)
- Remove short words
- Tokenize tweets
- Stem tweets

Analyze Tweets

The first part of analyzation is making a Word Cloud, which is a visual representation of the words in tweets. The more frequent the word appears in the tweets, the more prominence the word.

Figure 2. Word Cloud

The second part involves the calculation of hashtag mentions inside tweets. This part is done by extracting the hashtags from column 'Content' and calculate the frequency distribution using NLTK's frequency distribution function. After calculation, the data is then presented using Seaborn's barplot function.

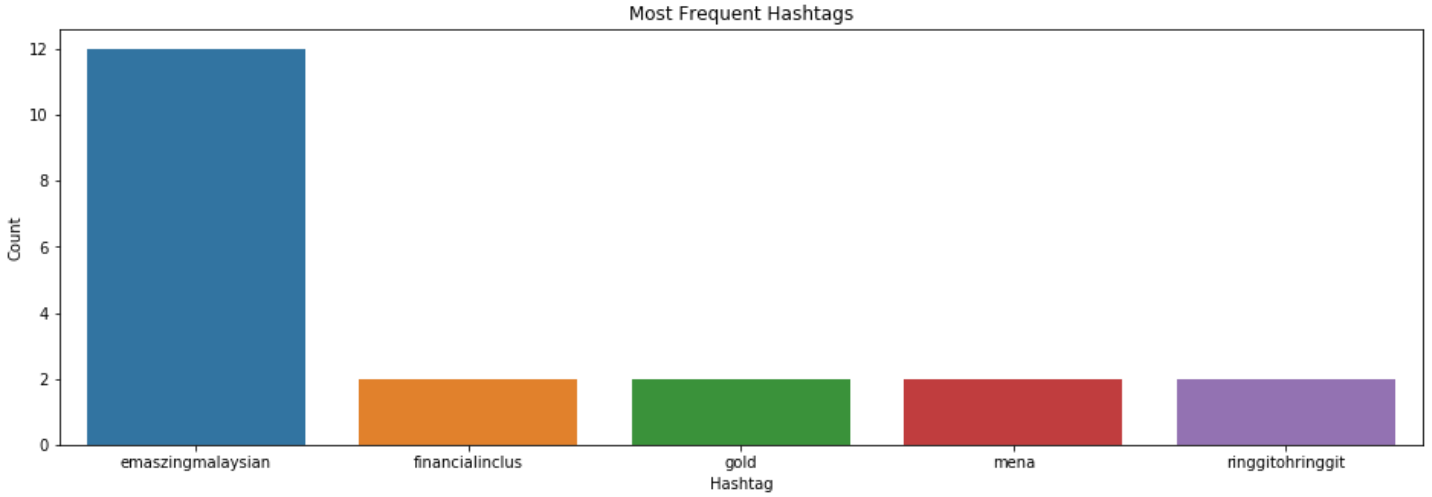


Figure 3. Frequency Distribution of Hashtag Mentions

For the third part, I used VADER's Sentiment Intensity Analyzer function to calculate polarity. Even though there are other libraries that can be used for sentiment analysis, I chose VADER because it uses 4 metrics for polarity: compound, positive, negative, and neutral. After calculation, the tweets are then categorized into 3 polarity types according to its compound value:

- If the compound value is less than zero, the tweet will be labelled as 'Negative'
- If the compound value is more than zero, the tweet will be labelled as 'Positive'
- If the compound value equals to zero, the tweet will be labelled as 'Positive'

	Author	Content	Date	Url	Followers	Tidy_Content	Compound	Negative	Neutral	Positive	Polarity
0	miha_miharu	ดาวน์โหลดแอพ HelloGold แล้วซื้อทองคำมูลค่า 500...	Wed Sep 18 02:18:33 +0000 2019	[{'url': 'https://t.co/deeUEgXbNw', 'expanded_...	27	hellogold xxxx	0.0000	0.0	1.000	0.000	Neutral
1	14ottobre2010	Molto interessati? A casa loro, le stesse macc...	Wed Sep 11 11:24:55 +0000 2019	[{'url': 'https://t.co/8UqTaeynpA', 'expanded_...	673	molto interessati casa loro stess macchinett q...	0.0000	0.0	1.000	0.000	Neutral
2	hazimdace	@myhellogold Muat turun aplikasi HelloGold. Bu...	Wed Sep 11 00:23:33 +0000 2019	[{'url': 'https://t.co/5dYr33ySbh', 'expanded_...	17	muat turun aplikasi hellogold buat pembelian e...	0.0000	0.0	1.000	0.000	Neutral
3	rcfl1967	RT @myhellogold: We sat down with our friends ...	Wed Sep 04 14:50:08 +0000 2019	[]	364	down with friend from mypf talk about gold hel...	0.4939	0.0	0.802	0.198	Positive
4	AlayyubMahazan	Muat turun aplikasi HelloGold sekarang. Buat p...	Wed Sep 04 14:37:26 +0000 2019	[{'url': 'https://t.co/0KZfgV9qeL', 'expanded_...	452	muat turun aplikasi hellogold sekarang buat pe...	0.0000	0.0	1.000	0.000	Neutral

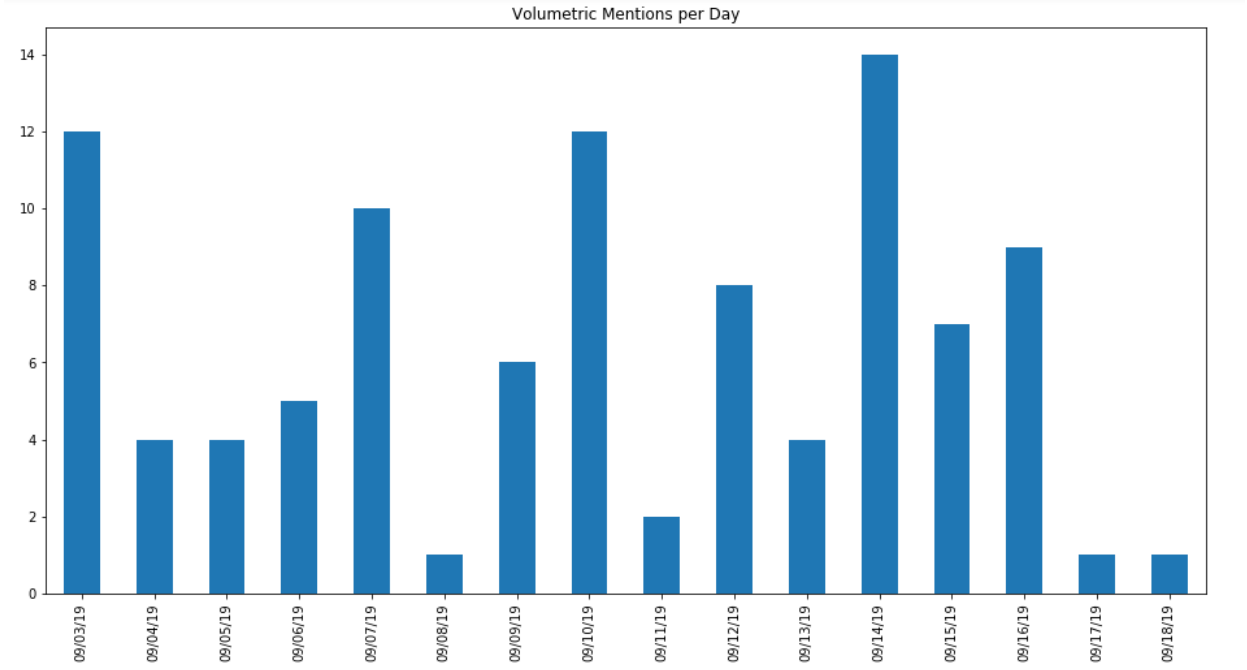
Figure 4. Example of dataset with polarity

Results

The goal for this project is to perform sentiment analysis on data taken from Twitter by answering the following questions:

- What are the volumetric mentions of the word 'HelloGold' per day?
- What are the users' sentiments towards the word 'HelloGold'?
- What are the net sentiment scores per users?

Firstly, I found out about the volumetric mentions per day by counting the number of tweets. Since the data was already inside a DataFrame, I used the pandas plot function to make a bar chart.



Secondly, I expressed the polarity of tweets per day by grouping. For this visualisation, I used stacked bar chart to show the highlight the different types of polarity measured each day.

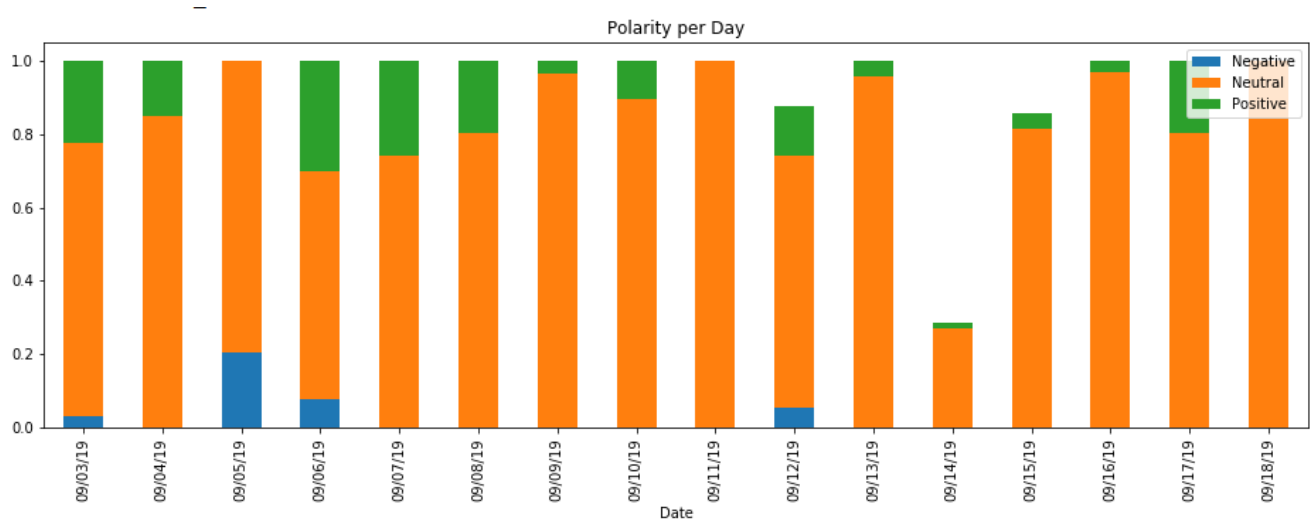


Figure 5. Stacked Bar Chart of Polarity

Lastly, I visualised the net sentiment score per user using Matplotlib's scatter plot. Due to size constraint, users with the same number of records and sentiment score were removed from the scatter plot.

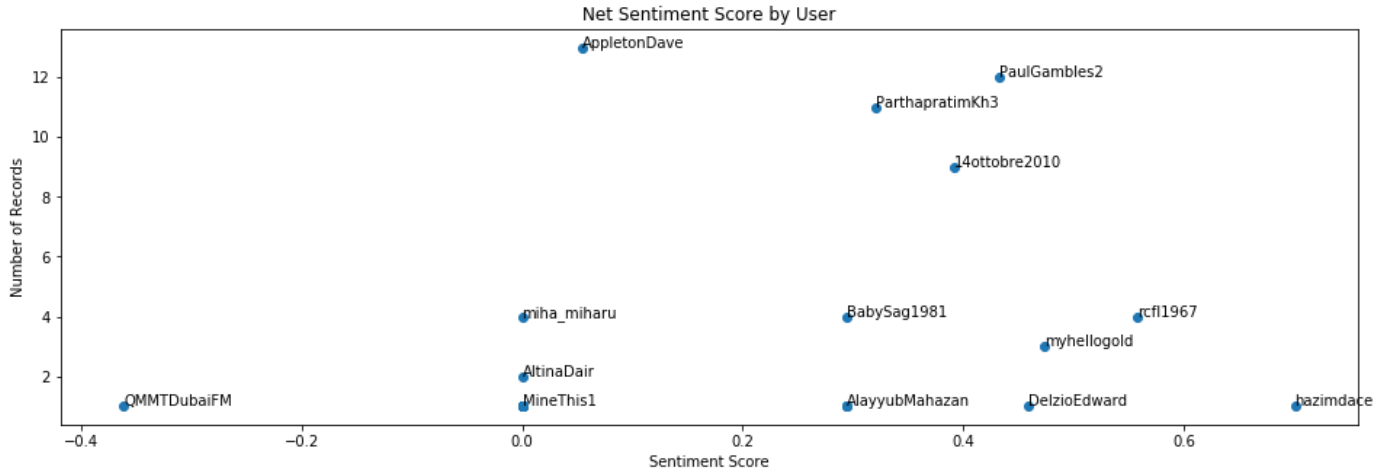


Figure 6. Scatter Plot of Net Sentiment per User

References

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30-38).

[2] Hutto, C., & Gilbert, E. (2014). . In *International AAAI Conference on Web and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122

[3] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).