DS501 Case Study 1 Team 5: TWITTER Mining

By Harsh Pathak, Jidapa Thadajarassiri, Krushika Tapedia,
Pitchaya Wiratchotisatian, Prince Shiva Chaudhary

Overview

In this study, we aim to do analysis on “Acer” using twitter data. Approximately 2,000 real-time tweets with #acer are collected and converted into a json file. Some frequency analysis are applied so as to study tweet entities in this data such as popular words, popular tweets, popular hashtags and popular user mentions. Then, Laurie Leshin is selected as our role model for being our new local brand ambassador. Her friends and followers lists on twitter are fetched and studied. Lastly, some interesting business questions are raised for further analysis; for instance,

1. How good is the brand awareness for Acer comparing to competitors?
2. What is the geographic segmentation for Acer’s customers?
3. How do people feel about Acer?
4. What time period in a day should we release an advertisement on twitter?

For these answers acquirement, an extra number of tweets from some competitors are also collected including HP, Lenovo, Dell and Apple.

Topic Selection

In this digital era, computers play an important role in many aspects especially for business and education. In college, although a number of personal computers are provided for students, most students still own at least one laptop. Most laptops typically last for 3-5 years. This implies that laptop owners usually buy a new laptop for every 3-5 years. When the event of buying a new laptop comes, people always come up with many aspects to consider such as specification, price, design or service. Most buyers habitually look for a higher specification with a lower price. Comparing the same specification among various brands, Acer tends to offer the lower price; however its unit sales are still lower than other brands such as HP, Lenovo, Dell and Apple.

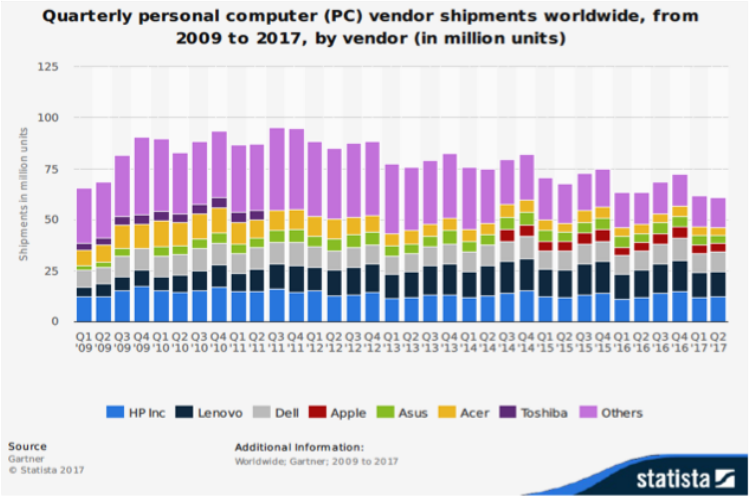


Figure 1: Quarterly personal computer (PC) vendor shipments worldwide, from 2009 to 2017, by vendor (in million units)[1]

[1] https://www.statista.com/statistics/263393/global-pc-shipments-since-1st-quarter-2009-by-vendor/

This fact motivates us to think of what happens to Acer. It is obvious that Acer offers good products with attractive price. What exactly are the points that Acer has missed? How could Acer do to compete that market? Lastly, what could we do if we are data scientists in Acer company? Therefore, our interest is doing twitter mining on Acer.

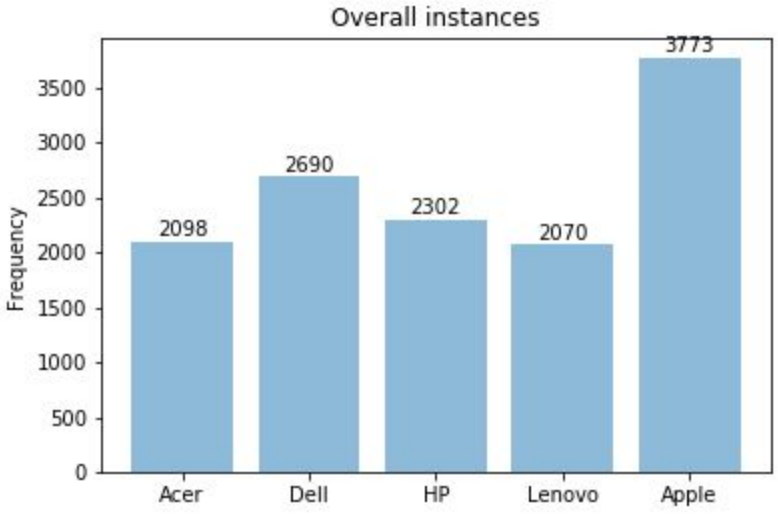
Data Gathering/Crawling Twitter Data

In this project, we deal with one of popular data sources, Twitter. The tweets contain a huge amount of real time information. There are millions of Twitter users all around the world. Anyone with a Twitter account can collect some tweets through public APIs at no cost.

Twitter APIs can be accessed only via authenticated requests. The authentication of API requests is carried out using Open Authentication (OAuth), an open standard for authentication adopted by Twitter to provide access to protected information. Each request must be signed with valid Twitter user credentials.[2] There is a rate limit to a specific number of requests with a rate limit window of size 15 minutes. Requests to the APIs contain some key parameters, for example, hashtags, keywords, and geographic regions. Responses from the APIs is Twitter user IDs in JavaScript Object Notation (JSON) format, which is a format that is widely used as an object notation on the web.

Twitter allows several features. Users can respond to another user's tweet by clicking on a reply button on that user's tweet. Users can mention other users in their tweets by adding '@' to the username of another user in a tweet. Users can share someone else's tweet to their followers. In addition, users can add a hashtag (#) before relevant keywords in their tweets to make it easier to search tweets. Very popular hashtags become trending topics on Twitter.

We crawl tweets from Twitter using Twitter Streaming API. We periodically collect tweets of five well-known laptop brands: Acer, Dell, HP, Lenovo, and Apple using hashtags. We obtain 12,933 tweets in total with the amount for each brand displayed below.



Retrieving data into Python, Pandas is used to read tweets from a JSON file. Data is manipulated in order to get rid of non-English words, English stopwords, links, and some special characters.

[2] https://dev.twitter.com

Data Analysis and Results

Part 1: Frequency Analysis

- *Word counts*
  - *Top 30 words*

    The number of each word is used to identify how popular that word is. Then the 30 largest number of word counts are selected and presented in the below table - top 30 words table.

acer	3219	alpha	1033	acerjapan	80	/ω/	71
aspire	2146	staffpick	1031	de	77	kae	70
notebook	1078	touchscree...	1030	ヘアサロン改装終了！	71	2353	70
antonlinecom	1066	laptop	141	ato	71	エイサー	68
switch	1048	gaming	140	ミスト７区４６号室	71	pc	66
lcd	1038	predator	122	レも完備したのでいつでも遊びに来てね	71	vr	53
touchscreen	1036	read	97	有頂天7区荘	71	10	52
						ebay	52
						dell	51

- *The most popular tweets*
  - *Top 10 tweets*

    The number of retweet of each tweet is used to identify how popular that tweet is. Then the tweets with 10 largest number of retweet counts are selected and presented in the below table - top 10 tweets table.

text	retweet_count
RT @antonlinecom: Acer Aspire Switch Alpha 12"...	1279
Acer Aspire Switch Alpha 12" Touchscreen LCD 2...	1279
RT @antonlinecom: ACER Gaming\nhttps://t.co/4v...	1271
ACER Gaming\nhttps://t.co/4vva53USXR\n#acer #p...	1271
RT @RECITONERS: Por que necesitas un "gran #...	103
...شاهد خصومات ضخمة, وأسعار تبدأ من 699 ريال, عل	83
RT @kae__ato: ミスト７区４６号室\nヘアサロン改装終了！\nトイレも完備したの...	82
ミスト７区４６号室\nヘアサロン改装終了！\nトイレも完備したのでいつでも遊びに来てね(*ノ...	82
RT @ayatokura: 【チラ見せ】Windows Mixed Reality対応アプ...	30
【チラ見せ】Windows Mixed Reality対応アプリ開発に関するご質問の多い内容...	30

- *The most popular Tweet Entities*
  - *Top 10 hashtags*

    The number of each hashtag is used to identify how popular that hashtag is. Then the 10 largest number of hashtag counts are selected and presented in the below table - top 10 hashtags table.

Hashtag	Count
acer	1267
aspire	1036
staffpick	1031
Acer	683
有頂天7区荘	71
エイサー	68
ACER	56
gaming	47
VR	47
FC東京	46

  - *Top 10 user mentions*

    The number of each user mention is used to identify how popular that user mention is. Then the 10 largest number of user mention counts are selected and presented in the below table - top 10 user mentions table.

User Mentions	Count
@antonlinecom:	1066
@AcerJapan:	80
@kae__ato:	70
@ayatokura:	33
@eBay	24
@IntelRussia:	23
@blackdragonsBR:	15
@IntelJapan:	9
@reparamoscompus:	9
@KaliMarcum:	7

Part 2: A popular twitter user

We access friends and followers of 'LaurieofMars', the twitter of Laurie Leshin, the president of WPI since we are thinking of her as a local brand ambassador. She has 1,100 friends and 5,570 followers.

We use tweepy.Cursor function from tweepy package. The function accesses the API to get the user_id (with id option). We then use get_user(<user_id>).screen_name to get the screen_name of a user. A list of mutual friends and followers is the intersection of the set of friend_IDs and the set of follower_IDs. Below are the top 20 friends, followers, and mutual friends from friend list and follower list.

- 20 *Friends*

ID	Screen name
16099390	dgoodtree
789123160578854913	Mass_STEMHub
594847411	keskrivan
15655383	mickuvirk
125857356	bethlogic
77764733	marty_walsh
337503376	PHKoules
835727932953763840	ofo_bicycle
301491540	HLuceFdn
19568591	ananavarro
289342771	laurenduca
846335241778352128	lrob0043
2624357553	k_deux_v
13691782	b0yle
2282176736	NorthLightAlert
856958594452946944	KJohnsonSUNY
903037667767492608	II_ASU
15584374	vorlon
114680171	cornerrobot
4164997882	EshipRPI

- 20 *Followers*

ID	Screen name
16099390	dgoodtree
1237109474	HSLavoie
15655383	mickuvirk
546377461	pinette14
910296270358683648	SoonRam76273422
17544399	JamesARay
910199889438691328	ANGELLRestrepo4
910198559001149441	William10407767
784337324515614720	AndrewAJJordan
3044774477	JMakl225
389602055	MikeBouso
337503376	PHKoules
909973166503718912	Jennife25439434
169526321	tamifite6
894792805343801344	pedrovotefor5
142781128	Jobelephant
417433623	RyanCanuel
909853130355068928	TammyMc82482740
701908237394845698	ATEKAssetScan
902877270468988928	AshishY88869744

- 20 *Mutual friends from friend list and follower list*

ID	Screen name
846335241778352128	lrob0043
789123160578854913	Mass_STEMHub
708683962827190274	Moho_Disco
722429857989271552	k8writesWPI
795630741497532420	Teixeira_Lab
710873031967514628	MishOnMars
755227401513275392	marnibhall
555024391	BolekNY
828655962529599488	Jiminy_Kirket
753663650641088512	wpi_ck
49891338	TMMCC
746046557548584965	Bogdan4Research
18309133	CaseyDreier
390440974	SGurska
1161545748	LeshinStephen
69644318	WPIWSoccer
3447795743	WPI_KEEN
16287777	jcwiley
2730410022	GGCatWPI
607272998	Mass_Tech

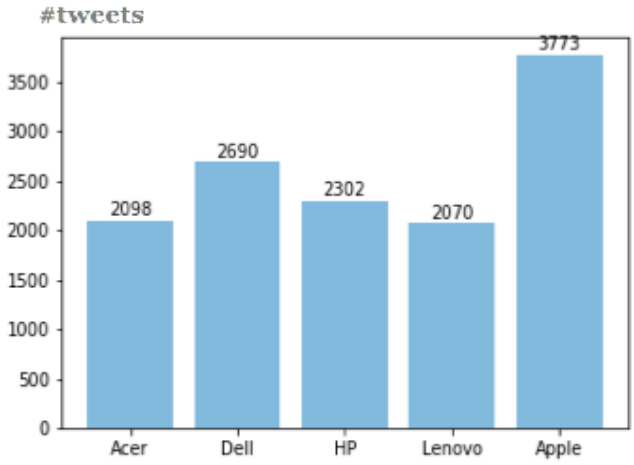
Part 3: Business questions

- ***Brand Awareness***

  Objective: to identify How good our brand awareness is when comparing to competitors.

  Methodology: the number of tweets of Acer and some competitors (HP, Lenovo, Dell, Apple) were collected.

  Results: the results of each company were plotted in below bar graph. Acer and Lenovo have the lowest number of tweets. This shows that Acer need to implement more activities to acquire more awareness.

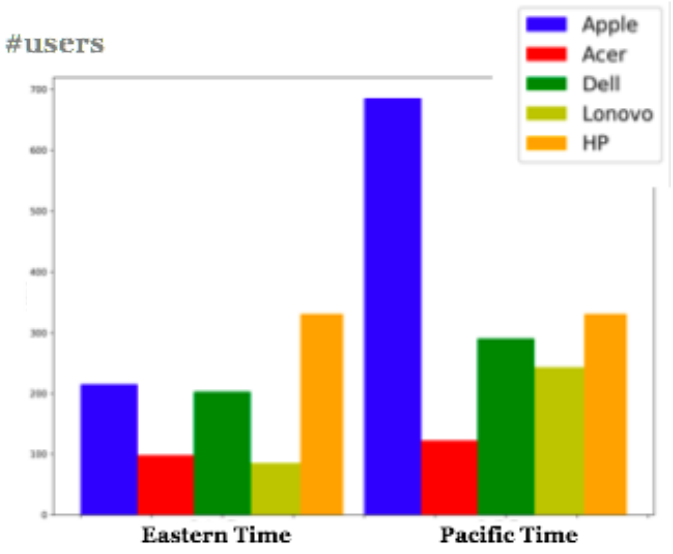


- ***Customers Distribution***

  Objective: to identify the distribution of Acer customers in the US. market

  Methodology: the number of twitter users in each location were collected.

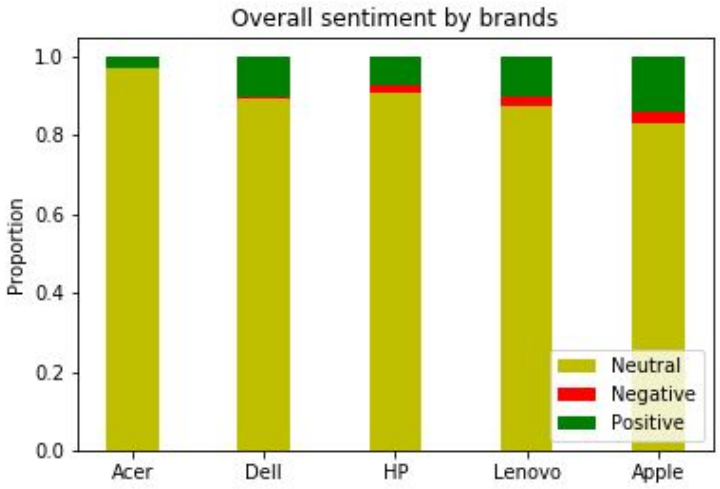
  Results: the results of number of users in eastern and western side were plotted in below bar graph. It shows that Acer's customers is still low in either side of US.



- ***Service improvement***

Sentimental analysis is performed to improve customer service. We use the 'nltk' package to categorize sentiment of tweets. SentimentIntensityAnalyzer.polarity_scores function in nltk.sentiment.vader is applied to cleaned texts. The result of each text consists of compound, positive, neutral, and negative scores. The compound score is computed by summing the valence scores of each word in the lexicon.[3] The positive, neutral, and negative scores are ratios for proportions of text that fall in each category, hence they all add up to 1. According to the following rules, we categorize sentiment of a tweet by comparing to a specified threshold p (usually p = 0.5) as following:

- Positive sentiment: compound score ≥ p
- Neutral sentiment: (compound score > -p) and (compound score < p)
- Negative sentiment: compound score ≤ -p



	Acer	Dell	HP	Lenovo	Apple
Positive	0.027	0.101	0.072	0.103	0.138
Negative	0.003	0.007	0.020	0.023	0.032
Neutral	0.969	0.891	0.907	0.872	0.829

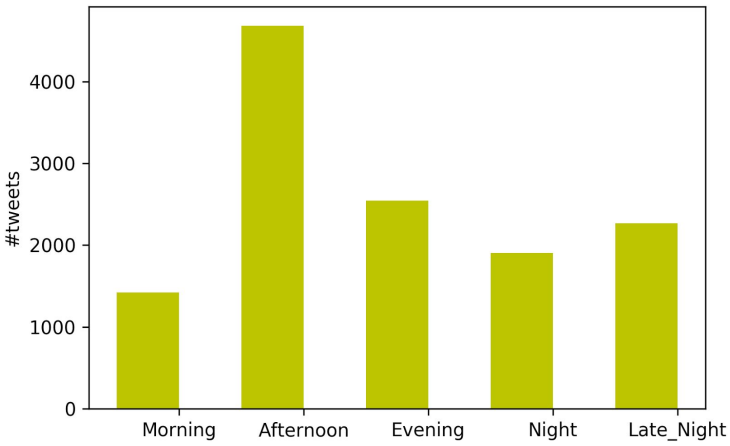
In this project, we set a classifying threshold to be 0.5. The proportion of 'neutral' sentiments are over 80 percent of the data. This is our concern because the neutral tweets might be valuable to study and might give us pros and cons of the products. Given more time in the future, we should look more closely to these twitters with neutral sentiment. The proportion of positive sentiment twitters is approximately five times of the negative sentiment twitters. This means people talk more positive than negative on Twitter. Apple has the largest proportion of positive tweets. Lenovo and Dell, respectively, have the second and the third largest proportion of positive tweets. Then HP, and Acer, respectively, have comparatively smaller proportion of positive tweets. This result agrees our assumption that Acer gets the less positive feedback from customers. We look into words with high frequency in negative sentiment texts, we find some relevant words, such as core, and i7. However, we are unable to answer the exact reasons why customers have negative sentiment towards Acer due to the small amount of negative sentiment texts and a lot of noise in the data.

[3] https://github.com/cjhutto/vaderSentiment

- ***Advertisement***

We analyzed data collected from twitter based on time slots i.e. Morning, Afternoon, Evening, Night and Late_Night. We observed that most of the tweets fall under slot categorized as Afternoon (1PM to 5PM) and Evening (5PM to 8PM). This infers that most people tweet about computers in the afternoon and continue until late night.

In conclusion, we can apply this trend to set timing of release of an advertisement around noon wherein users will be talking about it on twitter in the afternoon and continue up until late night. The proper release time of an advertisement will be able to draw highest attention from most customers and potentially impact product sales.



Summary

In this study, the results from frequency analysis of twitter data on Acer show top 30 words by ordering highest 30 word count, top 10 tweets by ordering highest 10 retweet count, top 10 hashtags by ordering highest 10 hashtags count and top 10 user mentions by ordering highest 10 user mentions count. Friend list and follower list of Laurie Leshin's twitter account are studied. We found that, from her total 1,100 friends and 5,569 followers, she has 550 mutual friends from both lists.

For business perspective, some interesting questions for Acer are raised and then answered by doing further analysis on twitter data. To answer how good brand awareness for Acer is, a number of tweet of Acer versus other brands (Apple, Lenovo, HP and Dell) are plotted and show that Acer has low awareness comparing to other brands.

One of the basic key in business that we need to realise before implementing other strategies is to know what is the geographic segmentation for our customers. To acquire this information from twitter data, the plot between the number of users in each location is used. Based on our data, we focus on how the Acer customers distribute around the United State and we found that Acer has the lowest customer share in both western and eastern side of the US.

Another interesting question for Acer is to know how people think about our business. For this topic, we implement sentimental analysis and found that among 5 brands - Acer, Apple, Lenovo, HP and Dell - Acer has the less positive feedback which is not good for the business. Acer might do deeper analysis to seek the reasons behind this and improve in some aspects.

Lastly, since how to make advertising being the most impact is always the issue for any business and also for Acer. We scope down this issue in how to make it most efficient in twitter.

We explain this by the plot of number of tweets versus time period in a day - morning, afternoon, evening, night and late night. The plot shows that the highest time of a day that people get involve in twitter is in the afternoon and continue until late night. Therefore, in order to get the most attention from twitter users an advertisement should be released since noon.