Predicting Brand Sales with Twitter Text Analysis

Thomas J. Weinandy

Ph.D. Candidate in Applied Economics

Western Michigan University

International Conference on Big Data Analytics and Data Science

November 11-12, 2019

Overview

Part One: Mining tweet data using the Twitter API

- Searching and streaming publicly available tweets
- Feature engineering and data management

Part Two: Techniques in natural language processing

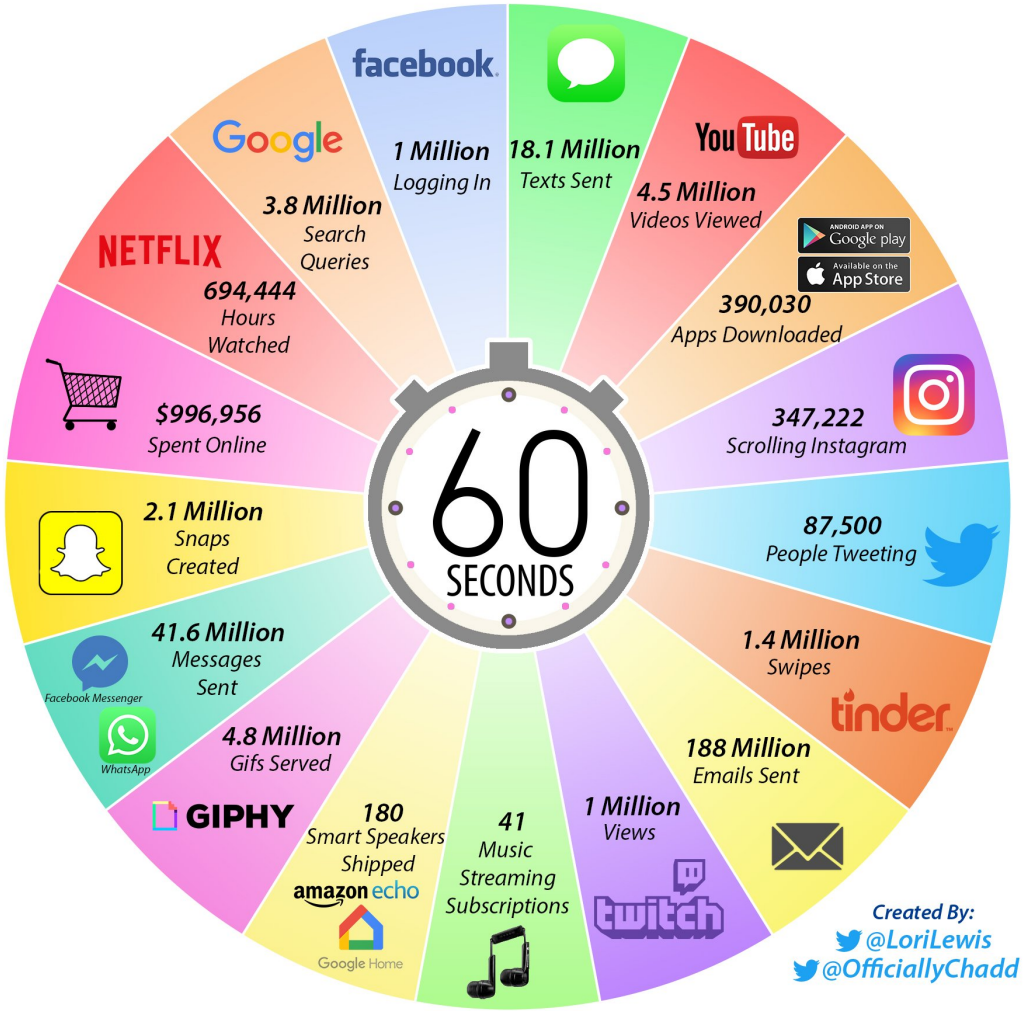
- Challenges with text analysis and solutions
- Sentiment detection tools

Additional resources for beginner and advanced coders

A fire hose of information

In the digital era there is an (over) abundance of data

- 500 million tweets/day[1]
- The challenge is not a lack of data but how to manage the volume, variety and velocity of that data



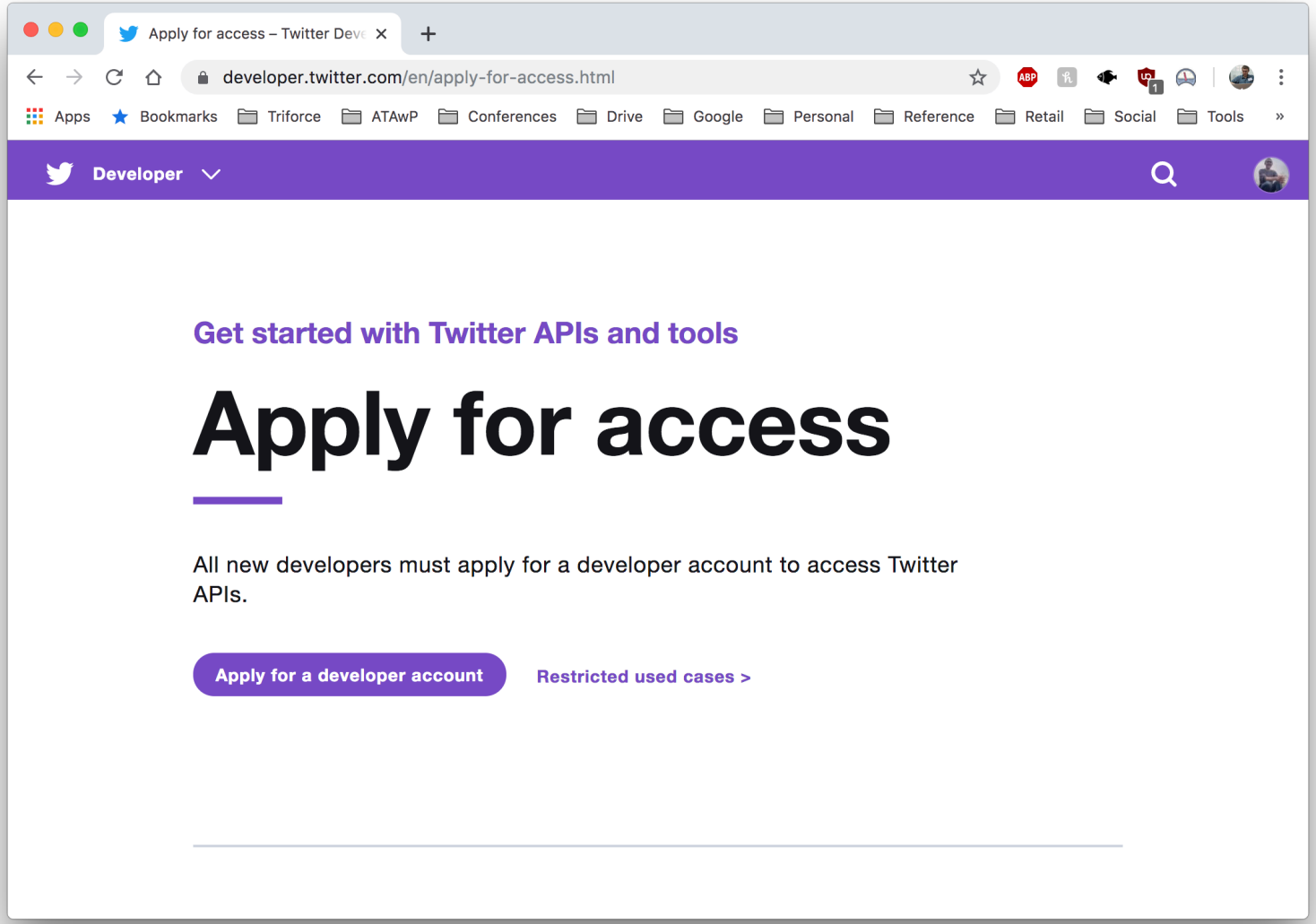
1. http://business.twitter.com (accessed 10/15/19)

What's in a tweet?

Carl Hovland's (1948) four elements of Social Communication:

1. Communicator sending the message (Twitter user)
2. Stimulus transmitted (tweet)
3. Individuals receiving the message (followers)
4. Response to the message (like/ retweet/comment, offline action)

The Twitter API: Authentication



The Twitter API: Search

Twitter offers a free, standard search API with the following restrictions:

- Up to 100 tweets per request
- Up to 180 requests per 15-minute window
- Up to the previous 7 days of tweets
- With a variety of search parameters, including:
    - Search query
    - Language
    - “recent” or “popular” results

Resource: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

The Twitter API: Streaming

Twitter also offers free, streaming API with the following restrictions:

- 400 keywords
- 5,000 user IDs
- 25 locations

Resource: https://developer.twitter.com/en/docs/tweets/filter-realtime

The Twitter API: Twython Wrapper

The Twython wrapper provides access to the Twitter API through some simple Python code

- Can search/stream tweets and output in JSON


```
{'statuses': [{'created_at': 'Wed Oct 23 00:14:01 +0000 2019',
  'id': 1186797891257397248,
  'id_str': '1186797891257397248',
  'text': '@shwood I picked the ace because I saw the movie "The Princess Bride"□',
  'truncated': False,
  'entities': {'hashtags': [],
   'symbols': [],
   'user_mentions': [{'screen_name': 'shwood',
     'name': 'Brian Brushwood',
     'id': 14645160,
     'id_str': '14645160',
     'indices': [0, 7]}],
   'urls': []},
  'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
  'source': '<a href="http://twitter.com/download/android" rel="nofollow">Twitter for
Android</a>',
  'in_reply_to_status_id': 1186793271571431426,
  'in_reply_to_status_id_str': '1186793271571431426',
  'in_reply_to_user_id': 14645160,
  'in_reply_to_user_id_str': '14645160',
  'in_reply_to_screen_name': 'shwood',
  'user': {'id': 14481210,
   'id_str': '14481210',
   'name': 'Theoiv',
   'screen_name': 'theoiv',
   'location': 'Austin, TX',
   'description': 'loving life after the military. loving each day i earned through service to my
country while in the US NAVY. Proudly Retired. A husband of 20+ years and a Dad',
   'url': None,
   'entities': {'description': {'urls': []}},
   'protected': False,
   'followers_count': 112,
   'friends_count': 386,
   'listed_count': 3,
   'created_at': 'Tue Apr 22 22:01:43 +0000 2008',
   'favourites_count': 3152,
   'utc_offset': None,
   'time_zone': None,
   'geo_enabled': True,
   'verified': False,
   'statuses_count': 926,
   'lang': None,
   'contributors_enabled': False,
   'is_translator': False,
   'is_translation_enabled': False,
   'profile_background_color': '000000',
   'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme5/bg.gif',
   'profile_background_image_url_https':
'https://abs.twimg.com/images/themes/theme5/bg.gif',
   'profile_background_tile': False,
   'profile_image_url':
'http://pbs.twimg.com/profile_images/1084638901791346688/errTiEjB_normal.jpg',
   'profile_image_url_https':
'https://pbs.twimg.com/profile_images/1084638901791346688/errTiEjB_normal.jpg',
   'profile_banner_url': 'https://pbs.twimg.com/profile_banners/14481210/1487056356',
   'profile_link_color': '3B94D9',
   'profile_sidebar_border_color': '000000',
   'profile_sidebar_fill_color': '000000',
   'profile_text_color': '000000',
   'profile_use_background_image': False,
   'has_extended_profile': True,
   'default_profile': False,
   'default_profile_image': False,
   'following': None,
   'follow_request_sent': None,
   'notifications': None,
   'translator_type': 'none'},
  'geo': None,
  'coordinates': None,
  'place': None,
  'contributors': None,
  'is_quote_status': False,
  'retweet_count': 0,
  'favorite_count': 0,
  'favorited': False,
  'retweeted': False,
  'lang': 'en'}],
 'search_metadata': {'completed_in': 0.016,
  'max_id': 1186797891257397248,
  'max_id_str': '1186797891257397248',
  'next_results':
'?max_id=1186797804817006591&q=The%20Princess%20Bride&count=2&include_entities
=1',
  'query': 'The+Princess+Bride',
  'refresh_url':
'?since_id=1186797891257397248&q=The%20Princess%20Bride&include_entities=1',
  'count': 2,
  'since_id': 0,
  'since_id_str': '0'}}
```


Resource: https://twython.readthedocs.io/en/latest/

What's in a tweet?

{'statuses': [{'created_at': 'Wed Oct 23 00:14:01 +0000 2019',
'id': 1186797891257397248,
'id_str': '1186797891257397248',
'text': '@shwood I picked the ace because I saw the movie "The Princess Bride"😅',
'truncated': False,
'entities': {'hashtags': [],
'symbols': [],
'user_mentions': [{'screen_name': 'shwood',
'name': 'Brian Brushwood',
'id': 14645160,
'id_str': '14645160',
'indices': [0, 7]}],
'urls': []},
'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
'source': '
Twitter for Android',
'in_reply_to_status_id': 1186793271571431426,
…}

Parsing

I run my Python code from Jupyter Notebooks to parse the data by:

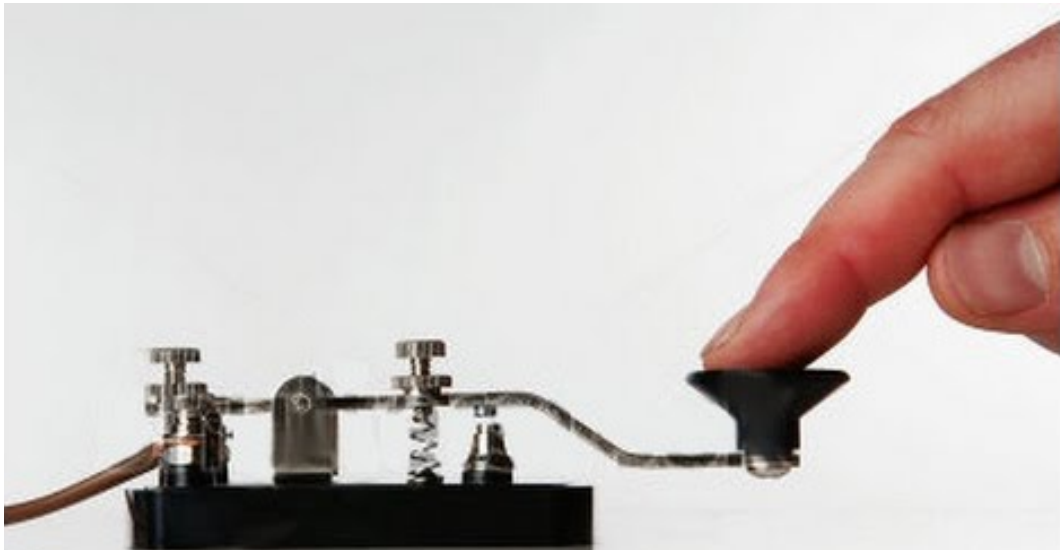
- Tweet data: tweet id, timestamp, text, hashtags, mentions, number of likes, number of retweets, device source
- User data: user id, user creation timestamp, number of followers, number of friends, bio, location

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	created_at	screen_name	followers_	friends_	statuses_	user_created_	location	source	in_reply_to	retweeted_	favorite_	retweet_	hashtags	bio	text
2	1.1898E+18	Thu Oct 31 0	Baasubaie7	57	60	3448	Sat Nov 03 20:	24:13 +0	Other			0	0	[]	#abdu	I'm at Starl
3	1.1898E+18	Thu Oct 31 0	Irzwianikaa	8	44	1485	Tue May 28 11:	24:16 +(	iPhone		0x1081bc7	0	733	[]	comeback	Awkarin w
4	1.1898E+18	Thu Oct 31 0	KallieAReynold	16	236	6382	Sun Jan 18 18:	Iowa	Android		0x108251e	0	236	[]		Iâ€™m rea
5	1.1898E+18	Thu Oct 31 0	haafizaaaaah	366	321	33363	Fri Feb 24 12:	Selangor	iPhone		0x1080d52	0	26	[]	i have trust i	Starbucks I
6	1.1898E+18	Thu Oct 31 0	nnforyh	4	61	2769	Sat Oct 12 11:	â˜…yose	iPhone		0xe38370a	0	74	['Johnny', 'Ja	â™¡ ðŸ˜˜❺˜ð	(Trans) 180
7	1.1898E+18	Thu Oct 31 0	kimmyye	179	319	162789	Thu Mar 11 17	Bangkok	Web App		0x1082d15	0	417	['à¹ⓘà¸ˆà¸ⓘ	à¸-à¸¹à¸ˆà¸´à¸	à¸‰à¸¥à¸-à
8	1.1898E+18	Thu Oct 31 0	erlinefelicia	446	215	6791	Wed Dec 22 0	Bandung	iPhone		0x1080ce9	0	1898	[]	a virgo woma	awkarin's p
9	1.1898E+18	Thu Oct 31 0	Atku_tta	691	83	76069	Sat Dec 09 16:	49:02 +0(	iPhone		0x1082d15	0	417	['à¹ⓘà¸ˆà¸ⓘ	Credit on Ph	à¸‰à¸¥à¸-à
10	1.1898E+18	Thu Oct 31 0	chaerulean	440	445	34514	Tue Nov 19 04:	57:15 +0	Web App		0x1082da1	0	1	[]		@chaerule
11	1.1898E+18	Thu Oct 31 0	deadlybarbieX	643	1805	49324	Sun Jul 07 01:	Selena| `	iPhone		0x1082a23	0	284	[]	ðŸ’»: bombe	Starbucks i
12	1.1898E+18	Thu Oct 31 0	saldhferiii	59	53	5149	Mon May 22 0	R.alm	Other			0	0	[]	Ø§Ù„Ø¹Ù❹	I'm at Starl
13	1.1898E+18	Thu Oct 31 0	azlynazman	128	154	101026	Fri Feb 19 10:	Kuala Lu	iPhone		0x1082cf49	0	42	['PromoJoeM	"if everyone	Promo Sta
14	1.1898E+18	Thu Oct 31 0	IcezingMerqry	232	273	76147	Fri Jan 24 02:	Thailand	iPhone		0x1082d15	0	417	['à¹ⓘà¸ˆà¸ⓘ	ã€ŒCU102 â	à¸‰à¸¥à¸-à
15	1.1898E+18	Thu Oct 31 0	tgarrett98	243	312	12206	Thu Aug 20 16	appalach	iPhone			0	0	[]	ig: tracieolivi	I will never
16	1.1898E+18	Thu Oct 31 0	githaanandha	612	247	5153	Sun May 20 1	Palopo	iPhone		0x1080c60	0	422	[]	ig: gsa._____	Ini untuk y
17	1.1898E+18	Thu Oct 31 0	kimberlorena	35	61	345	Sun Apr 28 23:	53:14 +0	iPhone	0x1082da543	d57b002	0	0	[]	w	I really crie
18	1.1898E+18	Thu Oct 31 0	capdemocrata	1	69	92	Sun Oct 27 23:	36:05 +0	iPhone		0x1082d91	0	2	[]		@melnicks
19	1.1898E+18	Thu Oct 31 0	auuelsa	31	93	1806	Wed Aug 17 2	à¸ⓘà¸£à	iPhone		0x1082d15	0	414	['à¹ⓘà¸ˆà¸ⓘ	Actor and en	à¸‰à¸¥à¸-à

Part 2: Natural Language Processing

Natural Language Processing (NLP) is the conversion of human language into a format that can be processed by computers

- Put differently, NLP is simply changing words into numbers



https://images.app.goo.gl/cKxfcvot2AB5tQTP6

The challenge of high dimensionality

Text data is inherently high dimensional

- Suppose we have a sample of messages that are each w words long and selected from p possible words
    - These documents will have a dimension of p^w
    - If the messages are 30 words long and selected from 1000 possible words, then this will have a dimension greater than the number of atoms in the universe (Gentzkow et. al. 2019)

Dimensionality reduction

There are several types of techniques used to reduce the dimension of the problem:

- Removing stop words (e.g. an, the, she, at, with, if)
    - Sample list: http://www.ranks.nl/stopwords
- Linear methods (e.g. Lasso/Ridge/Elastic Net regression)
- Non-linear methods (e.g. SVM regression)
- Scoring methods
    - Training an algorithm to assign a numerical score to text, e.g. sentiment

Why sentiment matters online

Bickart and Schindler (2001) show that subjective, user-generated reviews are considered more reliable than vendor information

Chong et. al. (2015) estimate on how an individual online review of product's attribute on Amazon can change its sales rank

- They also find the sentiment and number of reviews can help predict the sales of consumer electronics

Since language on Twitter has a distinct style, some authors recommend using a sentiment analyzer that is specific to the platform (Go et. al. 2009, Ghiassi et. al. 2013)

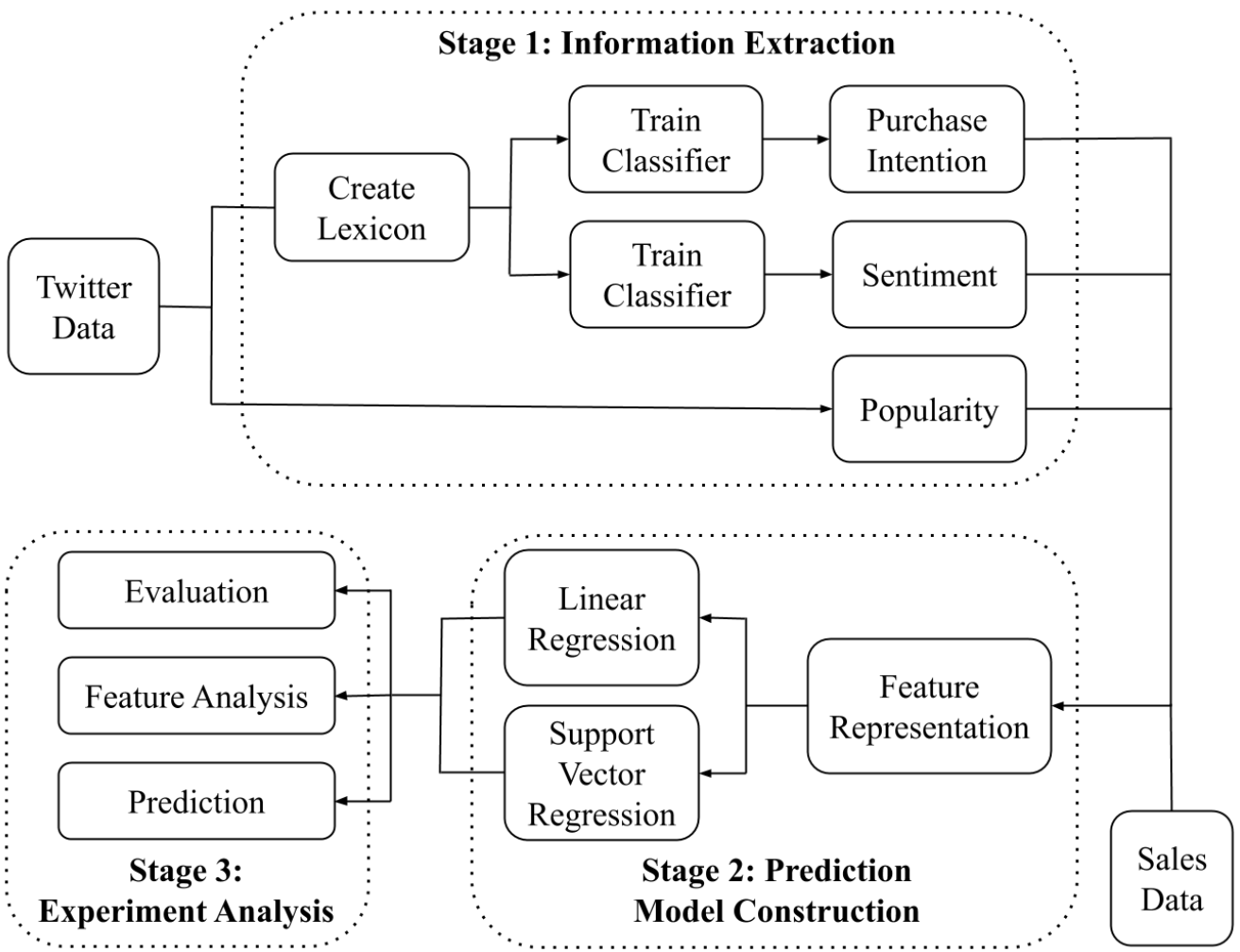
Why sentiment matters on Twitter

Bollen et. al. (2011) find tweets about company stocks that are “calm” and “happy” predict future price movements of those stocks

Ranco et. al. (2015) conclude that Titter sentiment *does not* help predict company stock price returns during normal times, but *does* predict returns during special events when the volume of tweets spike

Liu et. al. (2014) finds that movie box-office sales in China can be better predicted when measuring sentiment, popularity and purchase intention from the micro-blog Sina Weibo

Predicting brand sales with Twitter



Flow chart of methodology, adapted from Liu et. al. (2014)

Sentiment classifiers

You can build your own sentiment classifier using common machine learning techniques

- Go et. al. (2009) use “distant” learning to classify tweets with positive or negative emojis to assign sentiment

There are also some off-the-shelf Python libraries that can assign sentiment to text

- SentimentAnalyzer from nltk
- TextBlob

Conclusion

1. There is a treasure trove of information available on Twitter
2. The API and some simple Twython wrapper code makes this data accessible
3. Assigning sentiment scores is one way to solve the problem of high-dimensionality challenge with Twitter text
4. Apply for developer status and use my Python code to begin mining Twitter data

Works cited

Bickart, Barbara, and Robert M. Schindler. "Internet forums as influential sources of consumer information." *Journal of interactive marketing* 15.3 (2001): 31-40.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." *Journal of computational science* 2.1 (2011): 1-8.

Chong, Alain Yee Loong, Boying Li, Eric WT Ngai, Eugene Ch'ng, and Filbert Lee. "Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach." *International Journal of Operations & Production Management* 36, no. 4 (2016): 358-383.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. "Text as data." *Journal of Economic Literature* 57.3 (2019): 535-74.

Ghiassi, Manoochehr, James Skinner, and David Zimbra. "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network." *Expert Systems with applications* 40, no. 16 (2013): 6266-6282.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

Hovland, Carl I. "Psychology of the communication process." *Communications in modern society* (1948): 58-65.

Liu, Ting, Xiao Ding, Yiheng Chen, Haochen Chen, Maosheng Guo. "Predicting movie Box-office revenues by exploiting large-scale social media content." *Multimedia Tools and Applications* 75.3 (2014): 1509-1528.

Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. "The effects of Twitter sentiment on stock price returns." *PloS one* 10, no. 9 (2015): e0138441.

I'm here to network. Let's talk!

Thanks.

 @TomWeinandy

https://github.com/tomweinandy/twitter_data_mining