Word2Vec for Semi-supervised Sentiment Polling on Social Media

Pitipat “Mac” Kongsomjit

Description

The intent of this project is to explore the use of NLP techniques, in particular Word2Vec and K-Means Clustering as an experimental methodology of conducting sentiment polling.

Twitter API provides a way of retrieving information at the speed of thought. For the purposes of this project, we will be looking at recent Tweets mentioning Elon Musk as a case study of how Tweets can be used to derive sentiment polling insights.

Procedure/Techniques

Dataset—Gathering

Data is retrieved from twitter through Twitter API V2 in conjunction with Tweepy. In particular because of access limitations, we are using the “Search recent tweets” endpoints.

The query is for any tweet which mentions any variation of “elon” “musk” or “elon musk.”

Since the query limit for Tweepy is at 100 tweets, we are using Paginator to bypass this limit. In the end the dataset contains roughly 59,000 Tweets from the past 7 days mentioning Elon Musk.

Dataset—Example

	edit_history_tweet_ids	id	text	withheld.copyright	withheld.country_codes
0	[1594326317633503234]	1594326317633503234	It is after all only part of the bigger plan o...	NaN	NaN
1	[1594326317008551937]	1594326317008551937	@tomselliott @elonmusk It's brilliant marketin...	NaN	NaN
2	[1594326316576276482]	1594326316576276482	@AdamKinzinger Money protects money! Musk isn...	NaN	NaN
3	[1594326314953097219]	1594326314953097219	As pointed out by others, what Elon posted was...	NaN	NaN
4	[1594326314575712256]	1594326314575712256	@nealasher It's really difficult to say. Am su...	NaN	NaN

Dataset—Preprocessing

Most Tweets are very messy and need to be preprocessed, often containing links to attachments or other websites which needs to be removed. Interestingly enough many of these Tweets seem to be bot accounts advertising ‘altcoins’ or otherwise related to cryptocurrency in some form and only tangentially related to Elon Musk at all. This is likely due to Elon Musk’s presence in the cryptocurrency community due to his involvement in both Bitcoin and Dogecoin.

Some consideration was given to whether hashtags or mentions should be removed. In the end, it was decided that both should stay as they carry meta information about the Tweet at the expense of an acceptable level of noise.

Word2Vec—Brief Overview Based on Original 2013 Paper

Generally speaking, linguistic theory tells us that words with similar meanings will have similar surrounding words. An intuitive explanation is this: if two words have similar meanings, then you can use them interchangeably without changing the rest of the sentence, so for example:

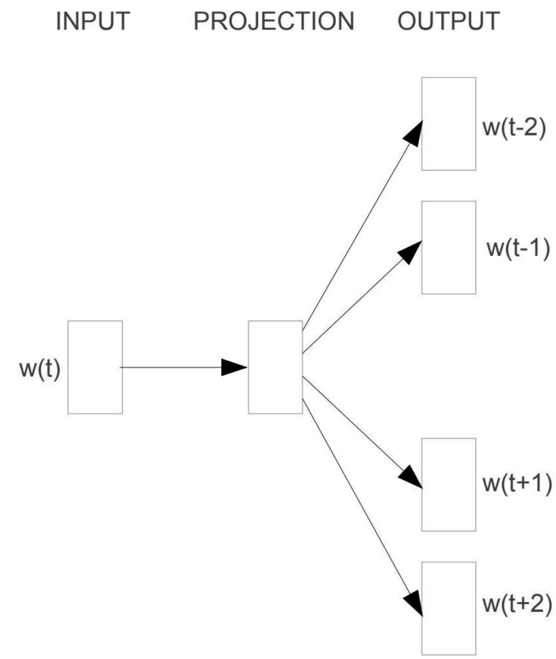
Center word
Surrounding words/Context words

“The smell of **rotten** meat is rancid”
“The smell of **spoiled** meat is rancid”

So if one reads through many documents, one might find that generally these words will have similar surrounding words.

Word2Vec—Brief Overview Based on Original 2013 Paper

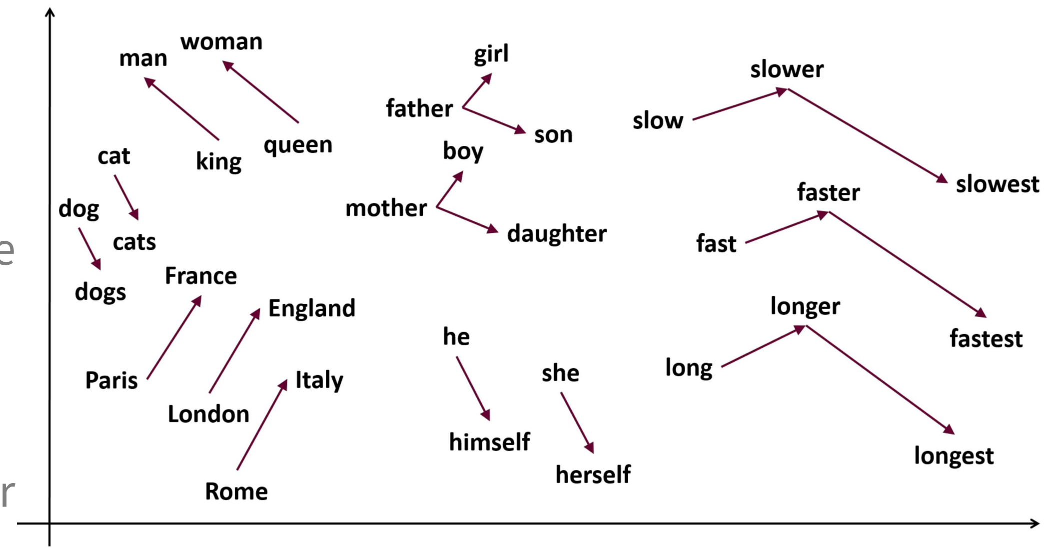
By leveraging the notion that surrounding words can define the semantic meaning of the center word, Word2Vec is able to capture the semantic meaning of words in its word vectorizations. By training a neural network where the goal is to predict the words surrounding the input center word, the neural network naturally learns vector representations of words such that the semantic meaning is captured.



Skip-gram

Word2Vec—Brief Overview Based on Original 2013 Paper

In practice, what this means is that the meaning of words are represented as vectors in euclidean space. Closer words have more similar meanings and certain euclidean operations have useful interpretations. For example, one Word2Vec model was able to capture the meaning of the word “king” such that the vector for the word “king” - “man” + “woman” resulted in the vector for the word “queen.”



Implementation

The python Gensim library implementation of word2vec is used. A corpus of 59,000 Tweets is parsed into a list-of-list structure where each sub-list contains strings of words. The structure is passed into the gensim word2vec implementation which then trains the model on the corpus.

```
In [16]:  #df.loc[0,['text']]
          tweets_list

Out[16]: [['It is after all only part of the bigger plan of Elon and his buddies: To destroy truth (bury it under lies and conspir
         acy theories), and ultimately destroy democracy.\nhttps://t.co/VO0FEn2oeO'],
          ['@tomselliott @elonmusk It’s brilliant marketing on Musk’s part, Tom.'],
          ['@AdamKinzinger Money protects money!  Musk isn’t much better than Trump!  Just not as crooked yet!'],
          ["As pointed out by others, what Elon posted wasn't really a *question*, it was an autocrat's *proclamation*.\n\nTrump's
         use of Twitter for inciting his violent insurrection, where people died, is no problem at all to Elon, obviously. To him,
         it's the *point*.\nhttps://t.co/0N82bL5Z7o"],
          ["@nealasher It's really difficult to say. Am sure there were those types but am also sure there were top notch sw engin
         eers who left. \nJack (founder) once said the need was to take Twitter back to its roots. Ie a start-up environment cut t
         o the bone. Elon is doing that and ensuring buy-in."],
```

Implementation

The word2vec model transforms a one-hot-encoding of a word into a 100x1 vector representation for a skip-gram network.

```
In [76]:  from gensim.models import Word2Vec

In [77]:  model_w2v = Word2Vec(corpus, min_count=1,vector_size = 100,workers=3, window =3, sg = 1)

In [78]:  model_w2v.wv["elon"]

Out[78]: array([ 0.28060898, -0.2292149 ,  0.0686238 ,  1.1552439 ,  0.6549879 ,
                -0.74889535, -0.7409937 ,  1.2111602 ,  0.315926  , -0.7267212 ,
                -0.02627701, -0.9344227 ,  0.8863182 ,  0.62607163, -0.2050012 ,
                -0.6154148 ,  0.50320935, -0.39041388,  0.25501612, -0.21944682,
                 0.6300483 ,  0.0386956 ,  1.2825071 , -0.8034544 , -0.14353912,
                -0.33927703, -0.66300464,  0.37030283, -0.02883749,  0.8101913 ,
                 0.02958682,  0.40976647, -0.16786988, -1.3916967 , -0.4013426 ,
                 1.3304071 ,  0.03851149, -0.45677668,  0.05272586, -0.59913075,
                 0.34484807, -0.62842065,  0.71499443,  0.16608554,  0.33693564,
                 0.60767114, -0.2369431 , -0.03435935,  0.52529573,  0.23274577,
                 0.29340056,  0.15942036,  0.96519876, -0.02837575, -0.5095847 ,
                 0.11199093, -0.33045217,  0.72958106, -0.05689446,  0.96552855,
                 0.5758763 ,  0.55037475,  0.55885756, -0.2440363 ,  0.07672374,
                -0.04062217, -0.06094014,  0.08778096, -0.54217196,  1.0055485 ,
                -0.48957363,  0.26732442,  0.23967932,  0.06380723,  1.4532294 ,
                 0.27245843,  0.2542462 , -0.36378333,  0.38480833,  0.3350726 ,
                -0.21267666, -0.12494438, -0.39337665,  0.42831618, -0.95005333,
                -0.5162044 ,  0.3024477 , -0.43668497,  0.25254288,  0.22727452,
                 0.6592299 , -0.19492826, -0.3443741 , -0.16282928,  0.15724519,
                -0.26059058, -0.4698184 , -0.06933168, -0.59312505, -0.7647325 ],
               dtype=float32)
```

Results/Analysis

Results—Word2Vec Associations

Since the euclidean distances of the word vector representations carry meaning, we can examine words that are close to each other in the vector space to understand what concepts people associate with what other concepts. For instance, these are the words our word2vec model see closely associated with the word “free.” This is calculated using cosine similarity coefficient.

```
get_synonyms("employees")
```

```
employees
staff workers engineers remaining companies HB jobs actions considers businesses critical women Eu
ropean emails reasons members managers hardcore millions investors others half conditions changes
software foreign management shareholders plans teams quit firing cuts visas extra those weight ppl
thousands rules shares layoffs users children Teslas ones devs tens Neither owners
```

Notice mentions of words such as “remaining,” “hardcore,” “firing,” “cuts,” “engineers.” This seems to indicate most discourse involving the word employees are focussing on the recent mass layoffs of Twitter employees

Results—Examples

Musk donald andrew ye kanye daddy Fucking trump lmfao beast mr nah antisemitism nigga ur muskrat Musk fr realdonaldtrump Bro girl king Eat dumb Don bigot musks Reiner DT id Cruz Omg lmao biden elon rn omg clown twt riding pls tbh goat Someone tate Man hello fuck btw gates Ye

Destroy tank save burn bankrupt crash push ruin turn ship lead drive eliminate fail maintain kill solve tanking protect abandon direct steal desperately remain teach punish merge belongs offering anywhere blow manage exploit settle trashing use manipulate dying accept discuss cross failing avoid ignore sue publish stick challenge scare funding dump

Tanking trashing dying digging hellscape utility refreshing doomed ruins chan Basically mega picking reverse competition buzz shady admitting sabotage slowly struggling effectively realizing circus burned lit practice sinking mindset seeks falling tank dive drug covered failing arrested backfired unfair baiting priority crashing slap proves burning bid march dump bs improving drives

Employees staff workers engineers remaining companies HB jobs actions considers businesses critical women European emails reasons members managers hardcore millions investors others half conditions changes software foreign management shareholders plans teams quit firing cuts visas extra those weight ppl thousands rules shares layoffs users children Teslas ones devs tens Neither owners

Elon mr musk ye donald andrew Elon trump ur melon lmfao bro kanye TeslaEventHQ awesome mister Omg MrMusk FXMS ox messiah twt tate lmaoooooooooo haha goat Bro Man im nigga kinda Ur jack girl ya daddy trippings Tim gold tf christmas fuck amazon i Ok realdonaldtrump bills lol hay TwitterBlue west

Trump Trump kanye tRump tfg DT TFG DJT donald ye Ye tate andrew west realdonaldtrump biden Rump trumps Kanye beg Tate orange mr antisemitism Don AJ Andrew Orange traitor twt Cruz testing insurrectionist guns DeSantis Hitler Jr antisemite bigot presidency him West unbanned Donald lmfao knees dick Ron daddy gates begging

Libtard barrister trashfire rebuttal deplatforming poisonous quiver arena cocaine informative forums maintenance lube illuminating albeit Driving sanctions neuro smooth MAJORITY tease Guarantee Words angle gathered rot TERFs swarming trained absent sleazy powder illegally circumstances reasonably sharp addiction awake damages functions triggers Artificial commits hoe awesomeness sweaty stacks disappoint selecting manufacture Igor

Results—Word2Vec Associations: Musk vs Elon

Musk donald andrew ye kanye daddy Fucking trump lmfao beast mr nah antisemitism nigga ur muskrat Musk fr realdonaldtrump Bro girl king Eat dumb Don bigot musks Reiner DT id Cruz Omg lmao biden elon rn omg clown twt riding pls tbh goat Someone tate Man hello fuck btw gates Ye

Elon mr musk ye donald andrew Elon trump ur melon lmfao bro kanye TeslaEventHQ awesome mister Omg MrMusk FXMS ox messiah twt tate lmaoooooooo haha goat Bro Man im nigga kinda Ur jack girl ya daddy trippings Tim gold tf christmas fuck amazon i Ok realdonaldtrump bills lol hay TwitterBlue west

It seems people who refer to the billionaire as “Elon” hold vastly different opinions from people who refer to him as “Musk”

Here positive words such as “awesome” or “goat” (slang for ‘greatest of all time’) are highlighted in green. Whereas negative words such as “clown” or “dumb” are highlighted in red.

Notice that generally tweets referencing the billionaire by “Elon” have almost no negative words at all, while tweets referencing “Musk” are use mostly negative words.

Given the cultural significance of referencing someone by their first name being an indicator of familiarity/closeness, this seems to indicate that people who view Elon Musk positively have some level of parasocial relationship with the billionaire.

Meanwhile people who reference the billionaire as Musk are a more mixed bag, as the cultural significance of referencing someone by their first name is both a sign of respect (IE: “Mr. Musk is a hard working CEO”) as well as a sign of distance/disapproval (“Musk just fired half of Twitter LMAO”). Although generally, it seems people who tend towards a negative perception of Musk associate him with either alleged racism (“antisemitism,” “bigot”), poor management decisions (“clown”, “dumb”), or his cult-like following (“riding” (as in so called ‘d*ck riding’).

Results—Word2Vec Associations: Billionaire

Billionaire narcissist fascist Nazi loser buddies genius moron total racist businessman egomaniac artist despicable unlike creates asshole creating represents sociopath grifter fool troll dipshit horrible famous idiot sociopathic human woman clown complex powerful evil con class bigot male supposedly capitalist boss manchild lying self stable criminal supremacy chess society rare antisemitic

Interestingly, it seems billionaire is overwhelmingly used more as an insult than as an honorific.

Notably, there seem to be 4 main “types” of negative opinions.

- Accusations of mental disorders(narcissist, complex, sociopathic, egomaniac).
- Accusations of racism (Bigot, racist, antisemitic, Nazi)
- Accusations of “conning” (con, artist, criminal)
- Mudslingers involving intelligence(fool, idiot, clown)

Results—Word2Vec Associations: Genius

Genius businessman master brilliant moron stable taste chess villain billionaire fool asshole egomaniac role clever narcissist absolute marketing sociopath incredible petty artist Nazi entrepreneur grifter entrepreneurs visionary loser parent manchild plain disturbing coward stan innovative fantastic Clearly D marketer twat complex capitalist creating dipshit rare besides liar trolling dickhead greedy scientist

It seems Tweets mentioning both Elon Musk and “Genius” fall into one of 2 categories, either Elon Musk supporters praising his intelligence, or critics calling into question otherwise.

Notable here is the inclusion of the word “chess,” this is likely in reference to the idea that Elon Musk has some sort of hidden ulterior plan and the current pandemonium at Twitter is part of this plan of some sort. This idea is often satirized by critics as a sarcastic exclamation that Elon Musk is playing “4D chess.”

Results—Word2Vec Basis Vector

Since the vector space encodes semantic meaning of words and euclidean operations such as vector addition or subtraction can be used to transform one word into another, it is not unreasonable to ask “what do the basis vectors of this vector space represent.” IE: is there a number of fundamental underlying concepts words are built off of.

To answer this I looked at the 100 basis vectors that make up our vector space and listed the known word vector closest to each of the basis vectors by cosine similarity. These were some notable results.

Of note: most of these cosine similarity values are rather small—roughly in the 0.2 range, so no strong conclusions can be made. However, there were some interesting patterns that emerged, so I have elected to include these results. Rather than drawing concrete conclusions from this, it is advisable instead to look at them as conjectures and areas of future investigation rather than concrete conclusions.

Next slide shows results

Results—Word2Vec Basis Vector

('Bitcoin', 0.19671478867530823-----------Perhaps representative of Elon Musk's presence in the cryptocurrency space
('democracy', 0.24716810882091522)----Maybe in relation to discourse surrounding free speech, censorship, and democracy
('I', 0.16360676288604736)------------------Indicative of how most tweets are first-person proclamations of personal opinions
('Twitter', 0.20414158701896667)----------Seems like a natural choice
('news', 0.20933282375335693)------------Shows Twitter's usage as a source of news platform
('CryptoDrafter', 0.2787479758262634)---Again indicative of Elon Musk's presence in the crypto space
('Filth', 0.1613483875989914)---------------Criticism of Elon Musk(?)
('thanks', 0.21898919343948364)----------Possibly representative of the underlying type of positive sentiment towards Musk
('CryptoMo', 0.257974773645401))---------Again indicative of Elon Musk's presence in the crypto space
('accounts', 0.15835785865783691)-------Related to discourse surrounding free speech, censorship, and democracy
('created', 0.19386467337608337)----------Related to discourse surrounding free speech, censorship, and democracy
('Reinstatement', 0.23132707178592682)-Related to discourse surrounding free speech, censorship, and democracy

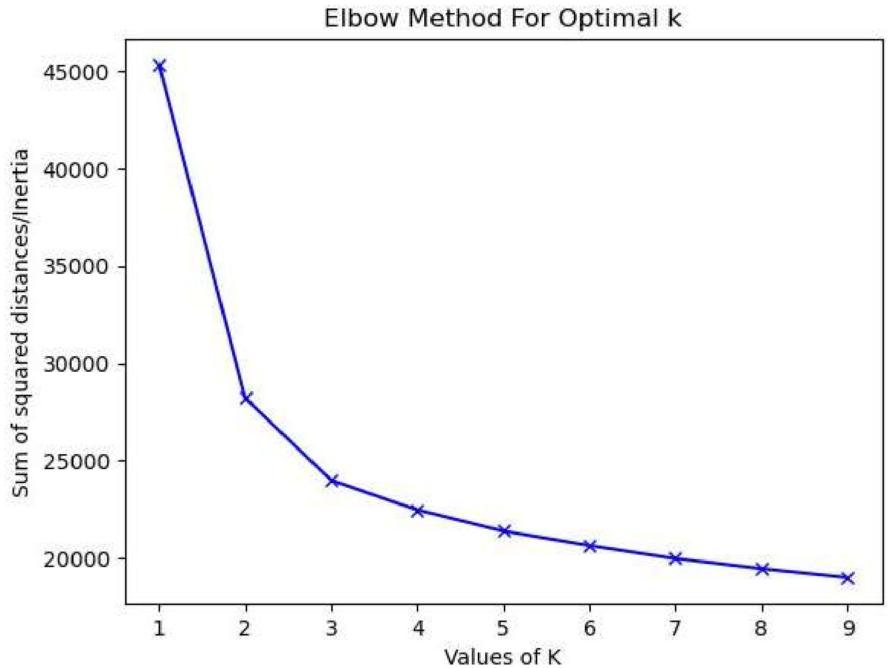
K-Means Clustering

Another idea is to conduct K-Means clustering on the vector representations of the words as a means of seeing what clusters of concepts are commonly associated with what in our Tweet space. For instance, maybe things will neatly separate into “Democrat vs. Republican” or “Pro-Musk vs. Anti-Musk” etc.

The sklearn k-means implementation is used for this project.

K-Means—picking optimal number of clusters

The elbow curve method was used to determine the optimal number of clusters, which resulted in k = 3.



K-Means—Results

```
for i in range(NUM_CLUSTERS):
    print(extract_words(model_w2v.wv.similar_by_vector(model.cluster_centers_[i], topn=50, restrict_vocab=None)))
    print('\n')
```

```
['DianeCoffeecrzy', 'bornacoric', 'eilmach', 'DEzperat', 'thehindu', 'Alyssa', 'salometerantodo', 'Nikhils', 'dorianmusk',
'KAJhwMlzzvoJ', 'NTR', 'zayalucky', 'YDF', 'xetcx', 'CofRiveriaNFT', 'RickPetree', 'elonkwon', 'rocihietoreibon', 'maricopac
ounty', 'airdropstosi', 'uuu', 'brankajovic', 'Toland', 'vickyyyever', 'mlkhattar', 'SGPCAmritsar', 'SMM', 'Navy', 'DueFact
s', 'illustratorby', 'HilltopLeader', 'tsukeyakibacoin', 'OpEd', 'TOIChandigarh', 'Andyold', 'iepunjab', 'Juliecooly', 'BDS
M', 'GoldUser', 'ylecun', 'belonmusk', 'juaniimaciel', 'rnkellie', 'ernestleenot', 'Yellow', 'GianChandBjp', 'Heh', 'BKmel
o', 'LynnGri', 'southasia']


['Shit', 'massively', 'dipshit', 'wasting', 'Basically', 'suggests', 'covered', 'unfair', 'attached', 'Cuban', 'buttons', 't
heirs', 'Anyway', 'FSD', 'comeback', 'burned', 'shenanigans', 'clone', 'Drumpf', 'reverse', 'split', 'dandy', 'Knowing', 'in
stantly', 'updates', 'circle', 'hat', 'popularity', 'round', 'Thus', 'cheap', 'note', 'football', 'meaning', 'struggling',
'Dumb', 'hammer', 'dems', 'outright', 'Guy', 'shadowban', 'corrected', 'fell', 'pool', 'exploit', 'choosing', 'reasonable',
'serve', 'dummy', 'lib']


['celebrates', 'classs', 'Adidas', 'MAJORITY', 'kayDawg', 'mylifeisabiglie', 'renamed', 'poker', 'slope', 'shithead', 'AriMe
lber', 'Opinions', 'Gotcha', 'Houston', 'Games', 'Patriotic', 'bureaucracy', 'hyper', 'researching', 'abc', 'bites', 'Jrs',
'sour', 'deer', 'twin', 'disparaging', 'reader', 'Facetoface', 'upto', 'execution', 'costume', 'Gus', 'spectacularly', 'Fren
ch', 'Vlad', 'magats', 'diarrhea', 'lemme', 'dp', 'joins', 'technological', 'Third', 'Winter', 'neuro', 'Artificial', 'Ps',
'contedo', 'discern', 'Acct', 'circumstances']
```

K-Means—Results Interpretation

Based on this clustering of words, it does seem like there is some level of political divide.

Words highlighted in blue are generally closer aligned to democrats—IE they are either used to refer to democrats or often used by democrats. For instance: left-leaning politicians such as AOC generally is a critique of Elon Musk, so left-leaning Twitter users are more likely to view Elon Musk negatively and use words such as “dummy” or “dipsh*t”.

On the other hand words highlighted in red are generally closer aligned to republicans. IE: “MAJORITY” for example is often used in discourse surrounding ‘rigged elections’ in recent memory it is used in discourse surrounding the recent Twitter poll to reinstate former president Donald J. Trump’s account. The all-capitalization spelling of the word is also more commonly seen among right-wing enthusiasts. “Magats” on the other hand is a derogatory slang referring to Trump supporters in reference to the ‘MAGA’ slogan (‘Make America Great Again’). “Patriotic” is generally also a common value held by American republicans.

'Shit', 'massively', 'dipshit', 'wasting', 'Basically', 'suggests', 'covered', 'unfair', 'attached', 'Cuban', 'buttons', 'theirs', 'Anyway', 'FSD', 'comeback', 'burned', 'shenanigans', 'clone', 'Drumpf', 'reverse', 'split', 'dandy', 'Knowing', 'instantly', 'updates', 'circle', 'hat', 'popularity', 'round', 'Thus', 'cheap', 'note', 'football', 'meaning', 'struggling', 'Dumb', 'hammer', 'dems', 'outright', 'Guy', 'shadowban', 'corrected', 'fell', 'pool', 'exploit', 'choosing', 'reasonable', 'serve', 'dummy', 'lib'

'celebrates', 'classs', 'Adidas', 'MAJORITY', 'kayDawg', 'mylifeisabiglie', 'renamed', 'poker', 'slope', 'shithead', 'AriMelber', 'Opinions', 'Gotcha', 'Houston', 'Games', 'Patriotic', 'bureaucracy', 'hyper', 'researching', 'abc', 'bites', 'Jrs', 'sour', 'deer', 'twin', 'disparaging', 'reader', 'Facetoface', 'upto', 'execution', 'costume', 'Gus', 'spectacularly', 'French', 'Vlad', 'magats', 'diarrhea', 'lemme', 'dp', 'joins', 'technological', 'Third', 'Winter', 'neuro', 'Artificial', 'Ps', 'contedo', 'discern', 'Acct', 'circumstances'

Ending Notes

Based on this analysis, I believe using Word2Vec is a valuable method of semi-supervised sentiment analysis where labeled data may not be available or feasible. Further, it’s versatility also lends itself to further future investigations as seen with the example of using K-Means.

Github/Sources:

Github:

https://github.com/macmacmacmac/DS504_Final_Tw

Reference Paper Word2Vec:

https://arxiv.org/abs/1301.3781

Twitter API:

https://developer.twitter.com/en/docs/twitter-api

Word2Vec Implementation:

https://radimrehurek.com/gensim/

K-Means Implementation:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html