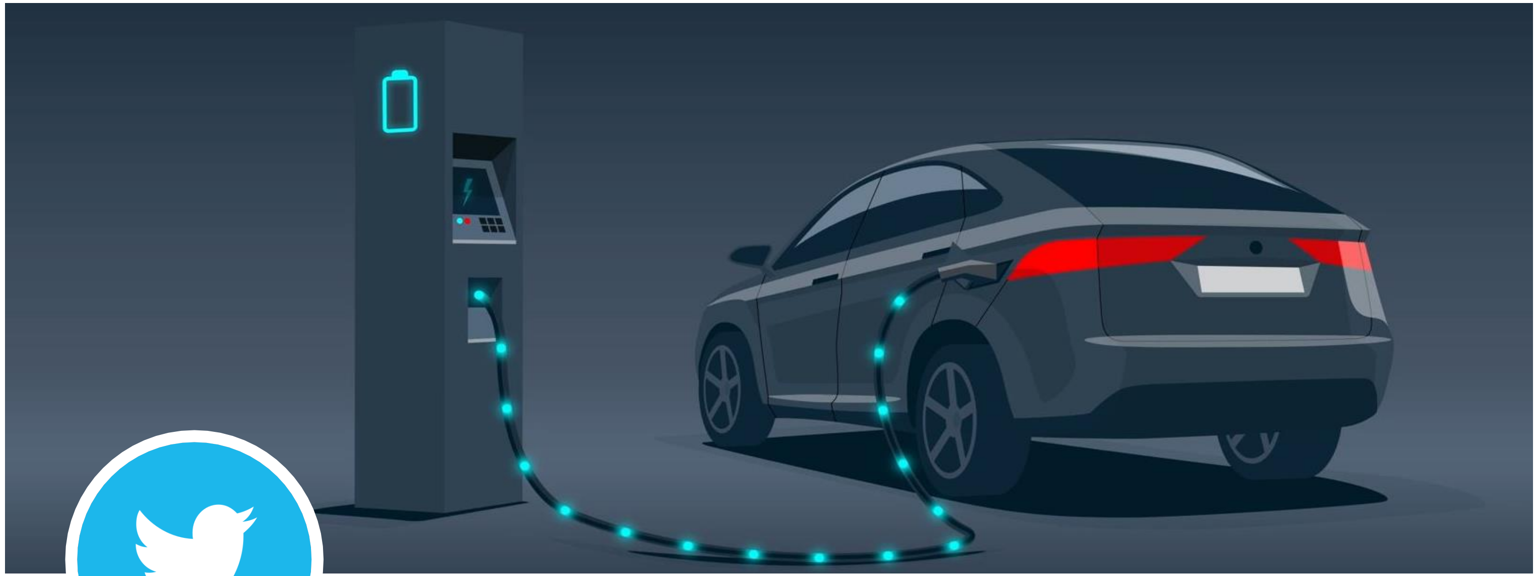
NILKAMAL SCHOOL OF MATHEMATICS,
APPLIED STATISTICS & ANALYTICS



Twitter Sentiment Analysis of Electric Vehicles

Group 8

Shraddha Kodavade	A037
Kaustubh Patil	A038
Leal Miranda	A039
Apoorva Singh	A040
Tanmay Talekar	A041

Mentored by:
Prof. Prashant Dhamale
Dr. Leena Kulkarni

Special Thanks:

BLUE ENERGY MOTORS

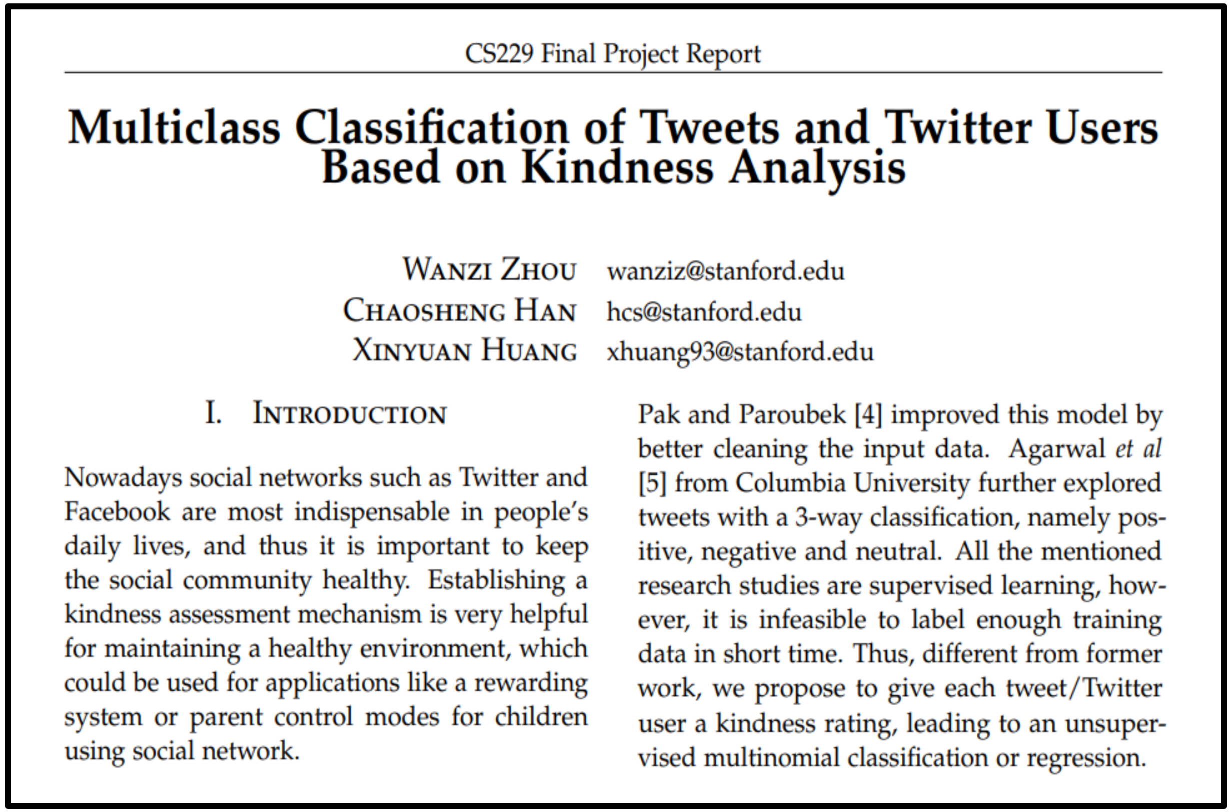
Subject:
Applied Multivariate Data Analysis & Financial Time Series Analysis

Objectives

- The project primarily focuses on analysing tweets to cluster the public sentiment about electric cars using unsupervised machine learning techniques.

- The sentiments thus obtained are then compared with the stock prices of an electric car company, checking for any correlation or association between the two.

Research Papers Referred



CS229 Final Project Report

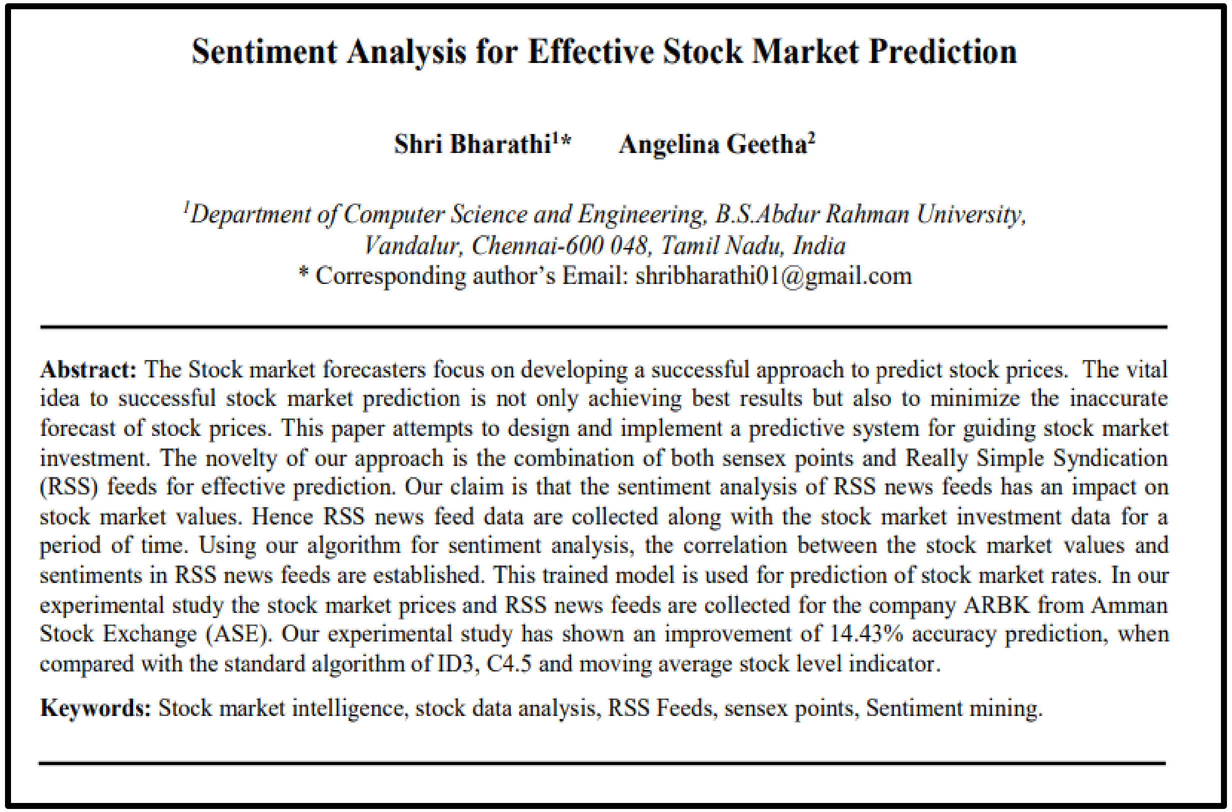
Multiclass Classification of Tweets and Twitter Users Based on Kindness Analysis

WANZI ZHOU wanziz@stanford.edu
CHAOSHENG HAN hcs@stanford.edu
XINYUAN HUANG xhuang93@stanford.edu

I. INTRODUCTION

Nowadays social networks such as Twitter and Facebook are most indispensable in people's daily lives, and thus it is important to keep the social community healthy. Establishing a kindness assessment mechanism is very helpful for maintaining a healthy environment, which could be used for applications like a rewarding system or parent control modes for children using social network.

Pak and Paroubek [4] improved this model by better cleaning the input data. Agarwal *et al* [5] from Columbia University further explored tweets with a 3-way classification, namely positive, negative and neutral. All the mentioned research studies are supervised learning, however, it is infeasible to label enough training data in short time. Thus, different from former work, we propose to give each tweet/Twitter user a kindness rating, leading to an unsupervised multinomial classification or regression.



Sentiment Analysis for Effective Stock Market Prediction

Shri Bharathi[1]* Angelina Geetha[2]

[1]*Department of Computer Science and Engineering, B.S.Abdur Rahman University, Vandalur, Chennai-600 048, Tamil Nadu, India*
* Corresponding author's Email: shribharathi01@gmail.com

Abstract: The Stock market forecasters focus on developing a successful approach to predict stock prices. The vital idea to successful stock market prediction is not only achieving best results but also to minimize the inaccurate forecast of stock prices. This paper attempts to design and implement a predictive system for guiding stock market investment. The novelty of our approach is the combination of both sensex points and Really Simple Syndication (RSS) feeds for effective prediction. Our claim is that the sentiment analysis of RSS news feeds has an impact on stock market values. Hence RSS news feed data are collected along with the stock market investment data for a period of time. Using our algorithm for sentiment analysis, the correlation between the stock market values and sentiments in RSS news feeds are established. This trained model is used for prediction of stock market rates. In our experimental study the stock market prices and RSS news feeds are collected for the company ARBK from Amman Stock Exchange (ASE). Our experimental study has shown an improvement of 14.43% accuracy prediction, when compared with the standard algorithm of ID3, C4.5 and moving average stock level indicator.

Keywords: Stock market intelligence, stock data analysis, RSS Feeds, sensex points, Sentiment mining.

Multiclass Classification of Tweets based on Kindness Analysis

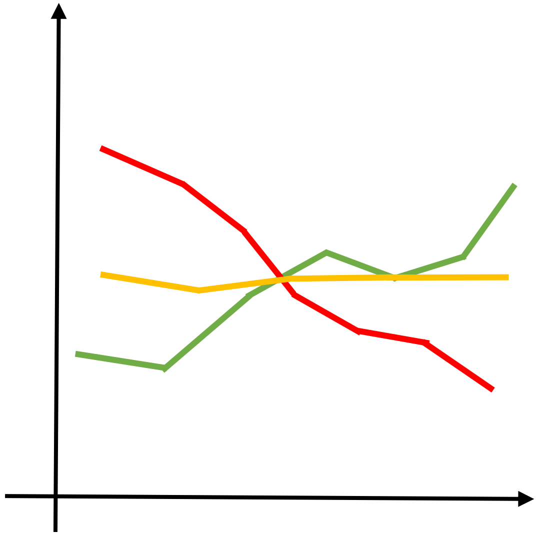
Published in 2016
Authors: Wanzi Zhou
Chaosheng Han
Xinyuan Huang

Sentiment Analysis for Effective Stock Market Prediction

Published in 2017
Authors: Shri Bharathi
Angelina Geetha

The Premise


Apoorva
@apoorva
Electric cars will change the way we move and how we make a living!



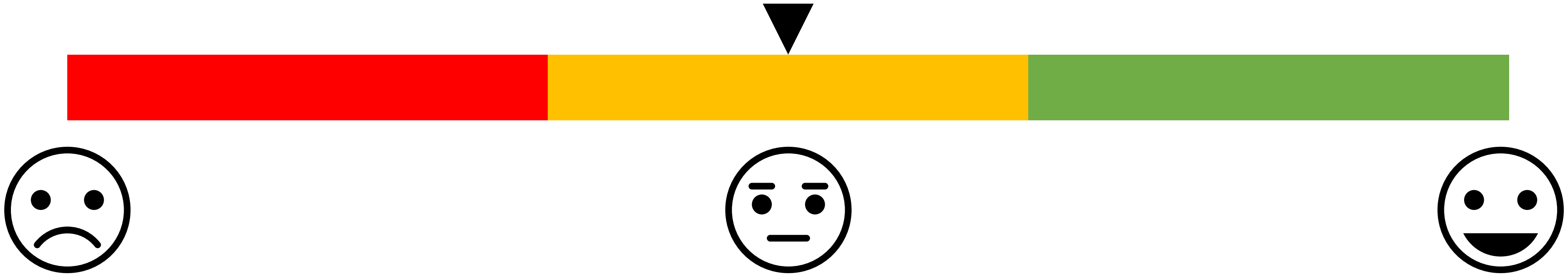


Table Of Contents

1 **Data Profile and Cleaning**

Collected Raw tweets using Twitter API and cleaned the tweets by using NLTK package.

2 **Word2Vec**

Used the Word2Vec algorithm on the cleaned tweets to convert tweets to vectors

3 **K means**

We run unsupervised learning algorithm using K-Means with K=3 for positive, negative and neutral.

4

Exploratory Data Analysis

Performed EDA on the dataset

5

Tesla Stock Prices and Superimposition

Collected Tesla Stock Prices for 2018-01-01 to 2020-12-31
and the correlation between stocks and the sentiments obtained using K-Means

6

Moving Average and Binning Results

Use MA method and binning method to perform analysis of association
on the sentiments with the movement of Tesla stocks

7

Conclusion

Concluded the project, which was backed by industry expert- Blue Energy Motors

The Sentiments

Sentiment Analysis attempts to divide the language units into three categories:

- **Positive**
- **Negative**
- **Neutral**

Apoorva
@apoorva

Electric cars will change the way we move and how we make a living! 💵

Shraddha
@shraddha

As if building electric cars and shooting rockets to Mars weren't enough work, Elon Musk has a new project !!! 🚀 🌕
Check out this link: https://t.co/sIBtpTU2S6

Tanmay
@tanmay

As we change our batteries for a new way of driving, here are the questions we should be asking https://t.co/rC7

Kaustubh
@kaustubh

Electric cars are coming. It is a question of time, not if. As a global electric utility, we aim to get infrastructure ready to accelerate the #MobilityRevolution

Leal
@leal

Tesla crashes into fire truck while reportedly on autopilot https://t.co/Mf1kzQoqch

Data Profile and Description

Date	1-1-2018 to 31-12-2020
Number of Tweets	903962
Likes > 10	43983

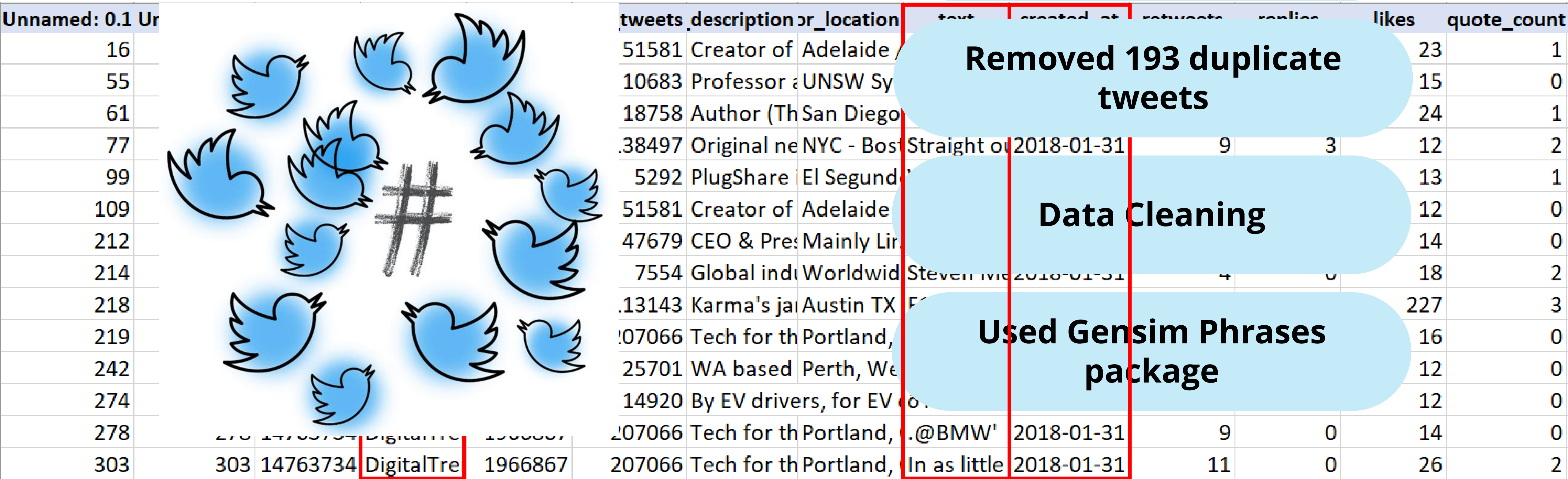
Collected tweets with 'electric cars' keyword

Select 3 columns

Removed 193 duplicate tweets

Data Cleaning

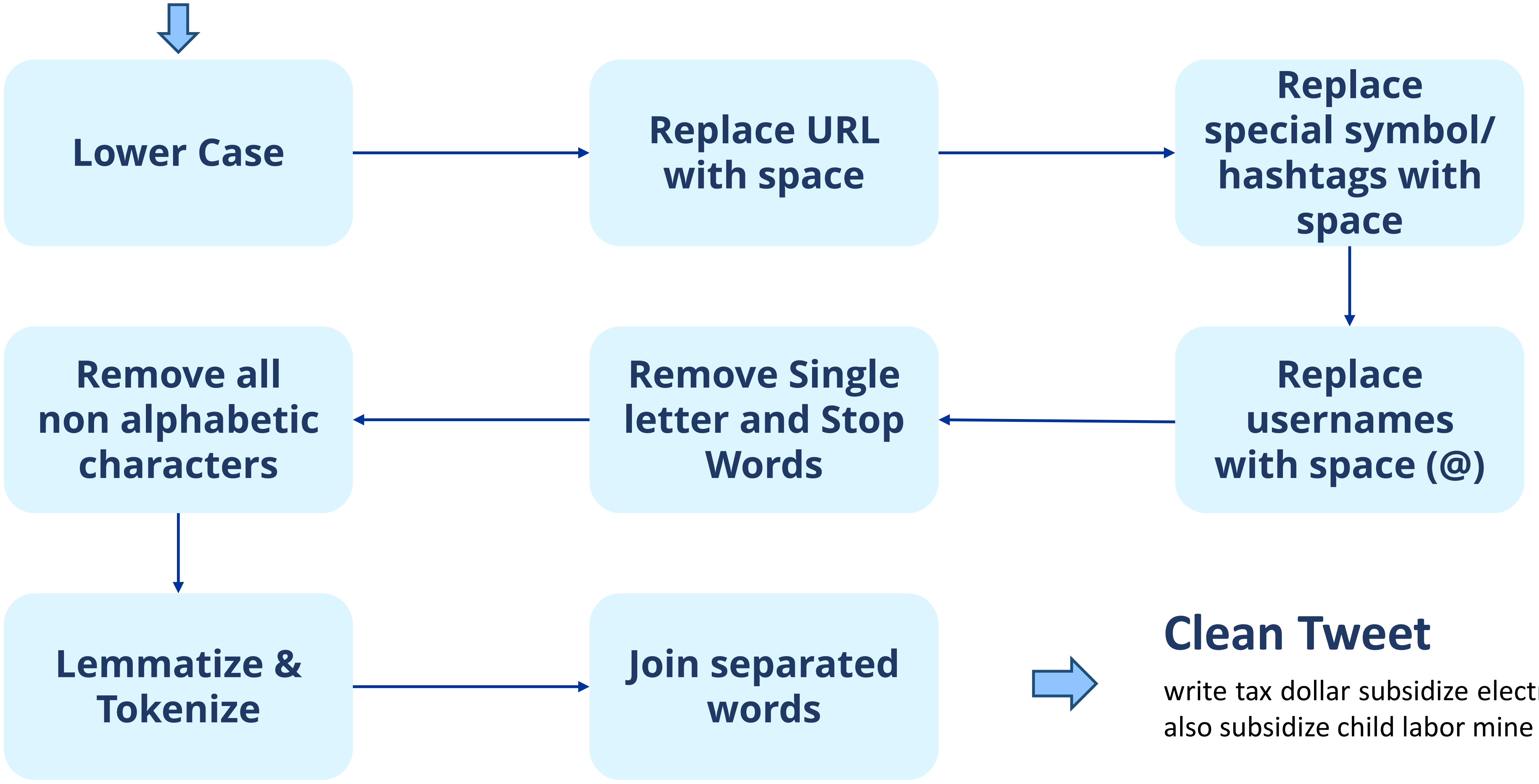
Used Gensim Phrases package



Bethany Shondark
@bethanyshondark

Raw Tweet

I wrote for @Ricohet about how our tax dollars subsidizing electronic cars is also subsidizing child labour in mines in Africa: https://t.co/12XHRuhlfO



Clean Tweet

write tax dollar subsidize electronic car also subsidize child labor mine africa

Word2Vec

Bag of Words

Word → **Numbers**

Example : **1. I love cars** **2. I love Tesla**

	I	love	cars	Tesla
D1	1	1	1	0
D2	1	1	0	1

One Hot Encoding Method

Drawback :

1. **Sparse Matrix**
2. **Similarity between words (cars and Tesla) is not captured**

TF / IDF Feature

Text vectorizer that transforms the text into a usable vector.

- **Term Frequency (TF) :** The number of occurrences of a specific term.
- TF_{ij}: *No of repeated words in a sentence / No of words in a sentence.*
- **Inverse Document Frequency (IDF) :** To reduce the weight of a term if the term's occurrences are scattered throughout all the sentences.
- IDF_i: *log(No of sentences/No of sentences containing the word)*

$$idf_i = \log\left(\frac{n}{df_i}\right)$$

- $\mathbf{idf_i}$: IDF score for term i
- $\mathbf{df_i}$: Number of sentences containing term i
- $\boldsymbol{n}$: Total number of sentences.

$$w_{i,j} = tf_{i,j} \times idf_i$$

- $\mathbf{W_{ij}}$: TF-IDF score for term i in sentence j
- $\mathbf{tf_{ij}}$: Term Frequency for term i in sentence j
- $\mathbf{idf_i}$: IDF score for term i

Tweets	
1	I like tesla
2	I love electric vehicles
3	I love electric vehicles but It is expensive

Step 1 : TF

	i	like	tesla	love	electric	vehicles	but	it	is	expensive
1	1/3	1/3	1/3	0	0	0	0	0	0	0
2	1/4	0	0	1/4	1/4	1/4	0	0	0	0
3	1/8	0	0	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Step 2 : IDF

Term	i	like	tesla	love	electric	vehicles	but	it	is	expensive
IDF	log(3/3)=0	log(3/1)	log(3/1)	log(3/2)	log(3/2)	log(3/2)	log(3/1)	log(3/1)	log(3/1)	log(3/1)

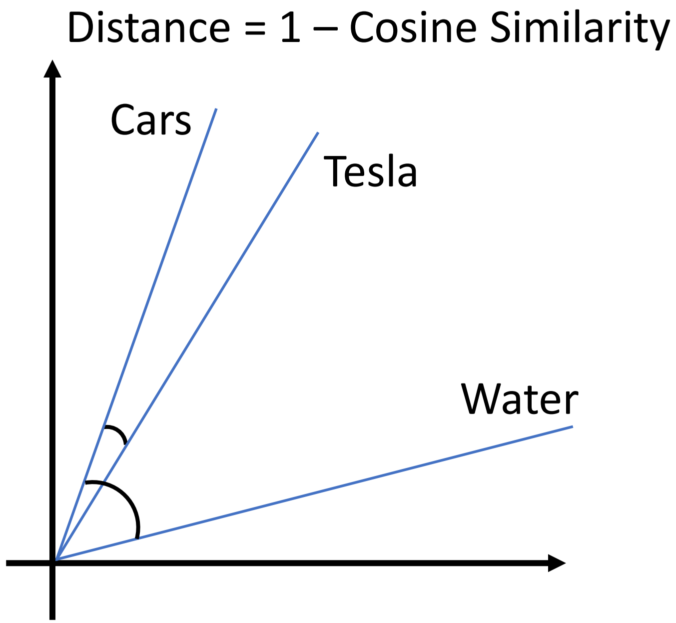
Step 3 : TF×IDF

	i	like	tesla	love	electric	vehicles	but	it	is	expensive
1	0	0.477121	0.477121	0	0	0	0	0	0	0
2	0	0	0	0.176091	0.176091	0.176091	0	0	0	0
3	0	0	0	0.176091	0.176091	0.176091	0.477121	0.477121	0.477121	0.477121

O/P

Word Embeddings

Words / Features	is_vehicle	needs_fuel	can_flow
Cars	0.9	0.95	0.01
Tesla	0.8	0.89	0.02
Water	0.01	0.02	0.93



How does it work? - By Using Neural Network

I love cars

Window=1

I	[1,0,0]
love	[0,1,0]
cars	[0,0,1]

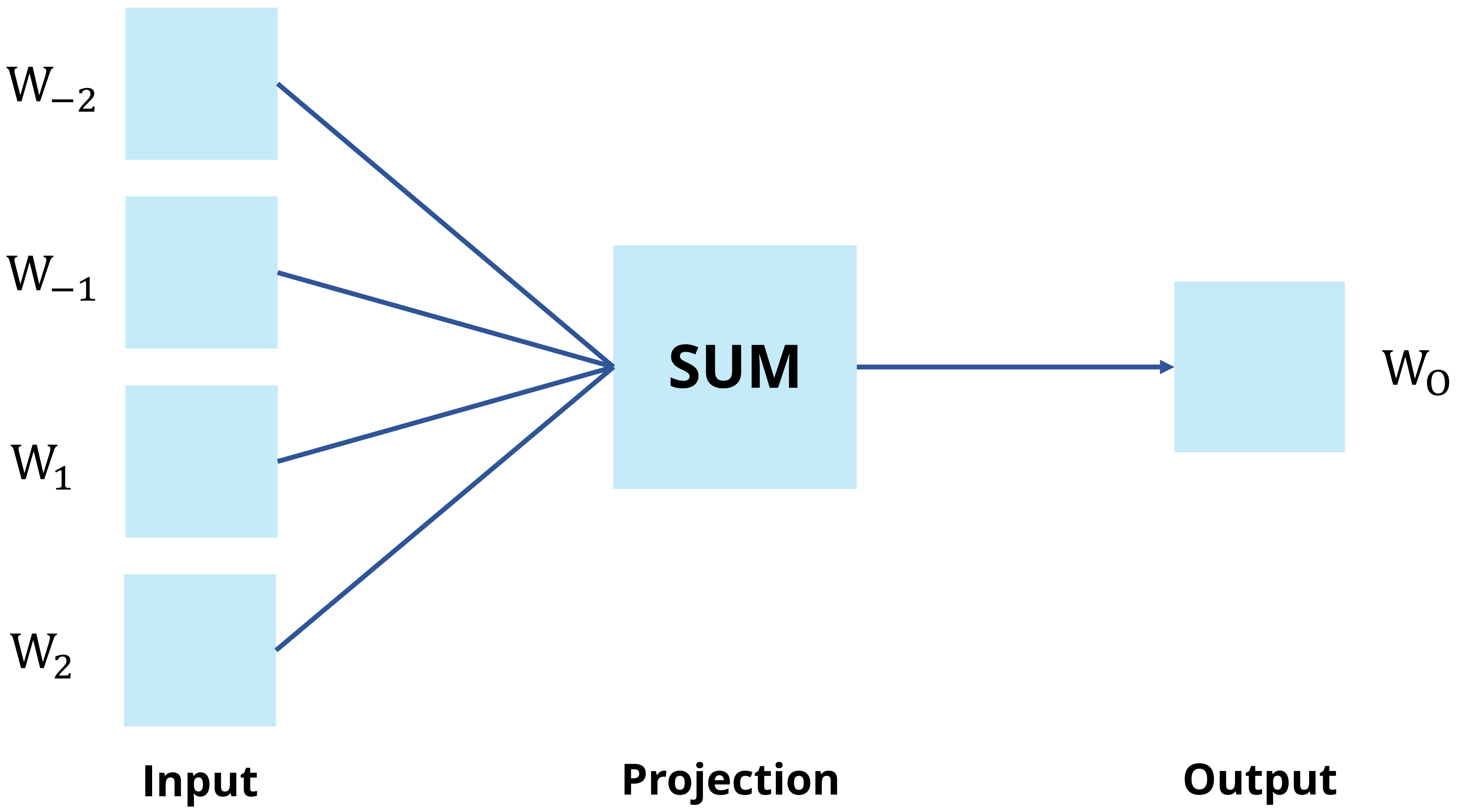
Input

Output

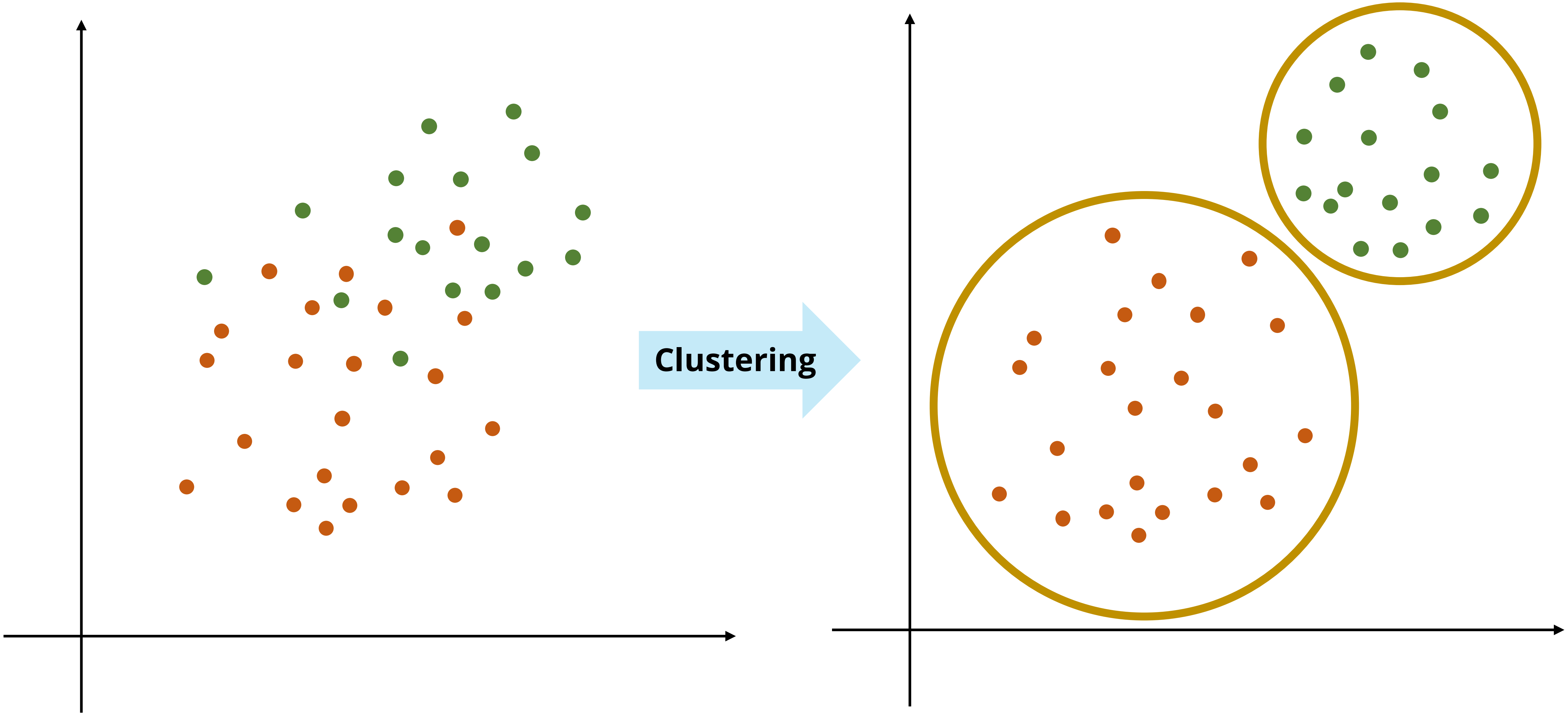
Window=2

I	[1,0,0]
love	[0,1,0]
cars	[0,0,1]

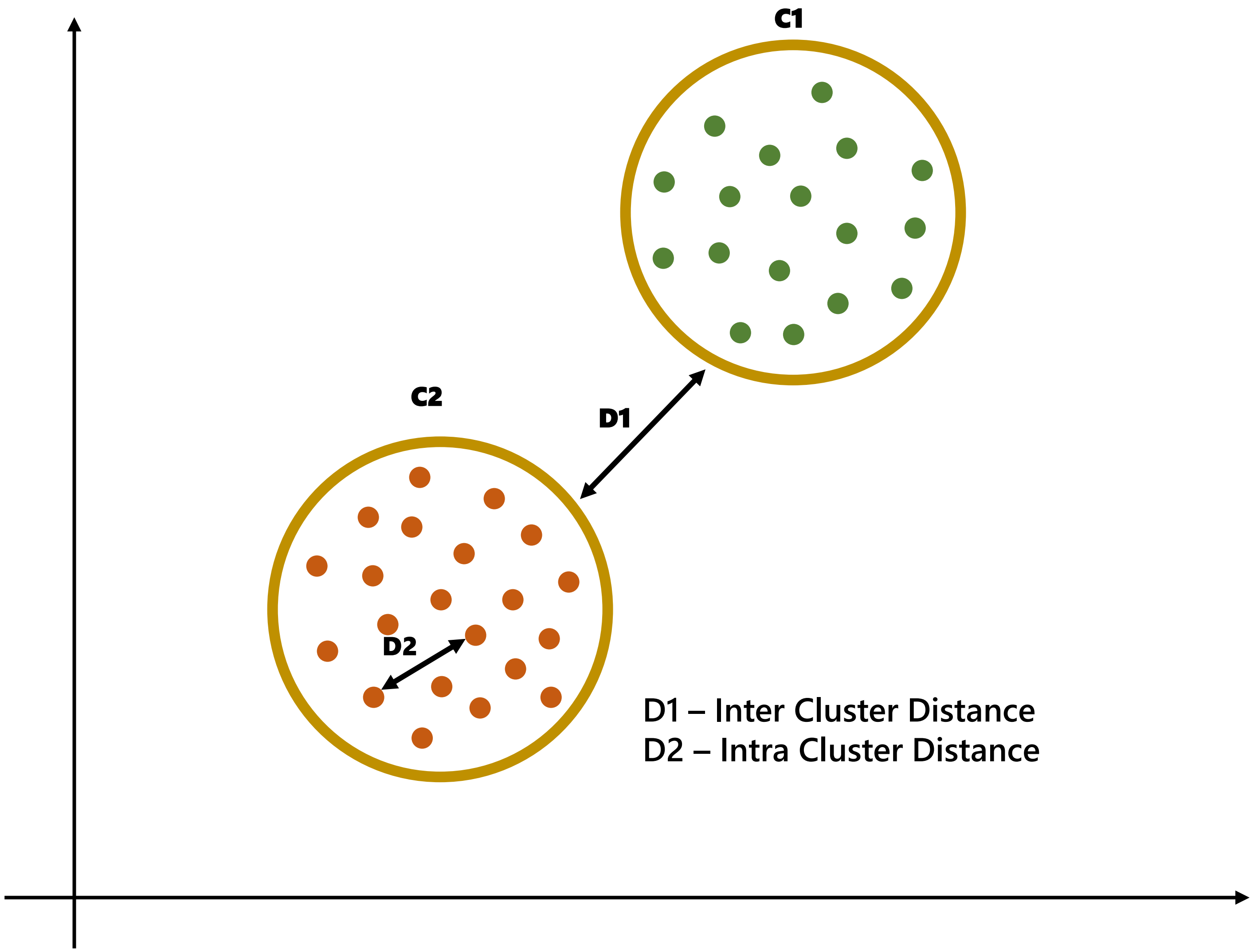
Continuous Bag of Words (CBOW)



K Means



K Means



The Formula

$$\underbrace{J}_{\text{Objective Function}} = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance Function}}$$

Where,

$k = no.of\ clusters$

$n = no.of\ cases$

$x_i^{(j)} = case\ i\ in\ j^{th}\ cluster$

$c_j = centroid\ for\ cluster\ j$

K Means on Word2Vec

Train Word2Vec Model

Extract Word Vectors

Cluster the Word Vectors

Interpret the Clusters

Use the Clusters to understand the Sentiments

K Means Result

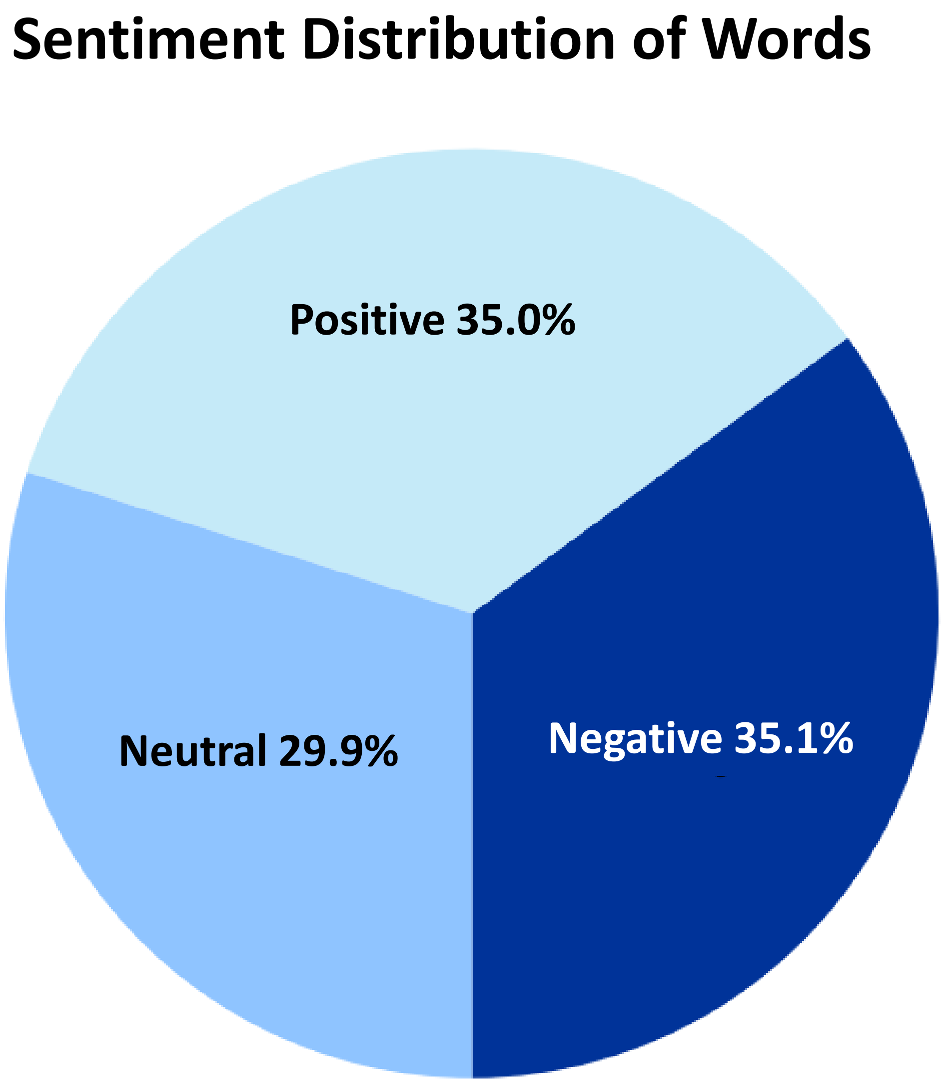
Cluster 1	
moron	0.163284
decline	0.110696
climate_change	0.137932
crisis	0.112269
expensive	0.169612

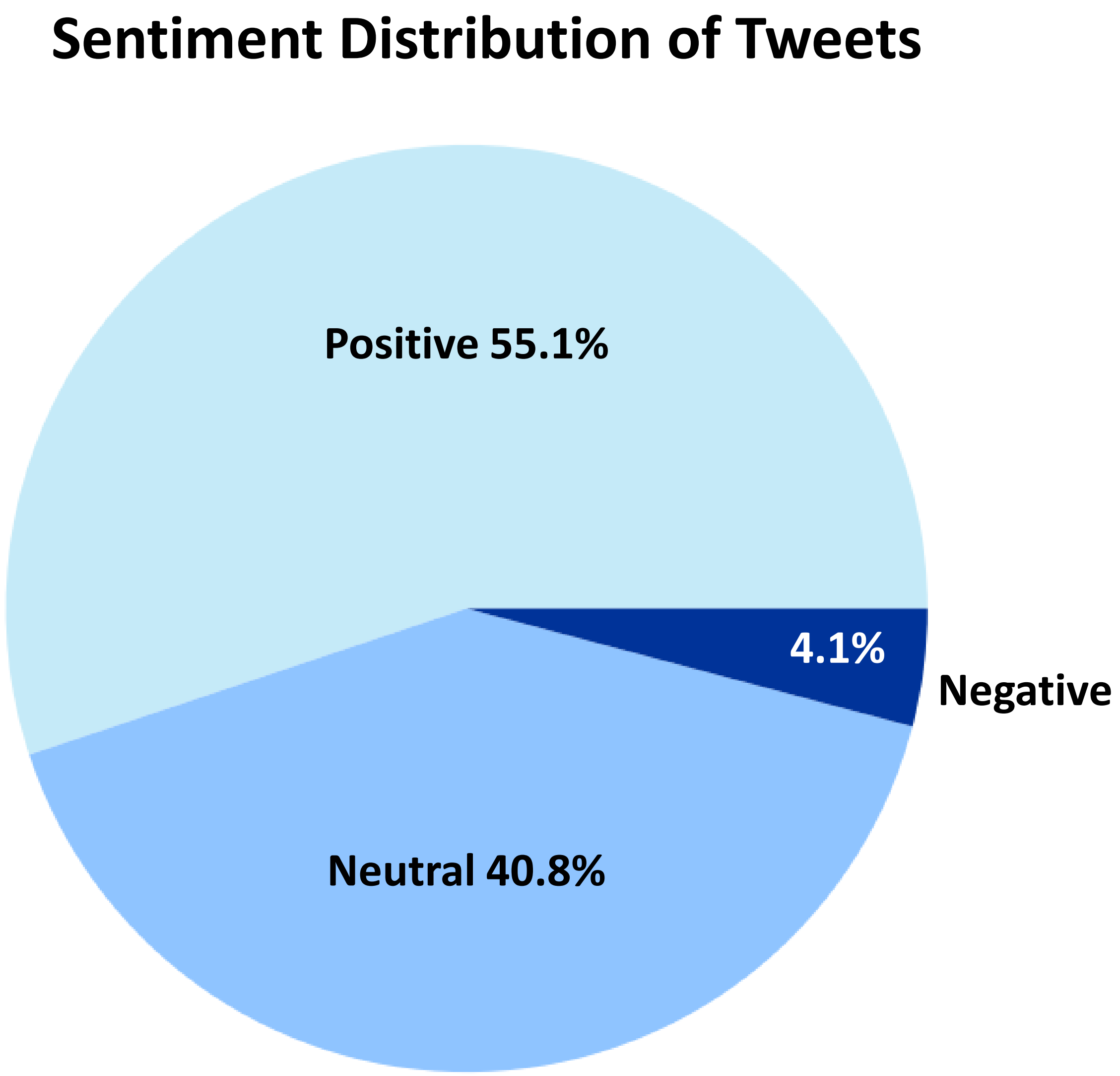
Cluster 2	
innovate	0.135221
impressive	0.149991
faster	0.147905
efficient	0.140876
price_drop	0.126916

Cluster 3	
car	0.157001
package	0.149991
electric	0.147905
motorist	0.144958
automotive	0.126916

Recoding	
-1	Negative
0	Neutral
1	Positive

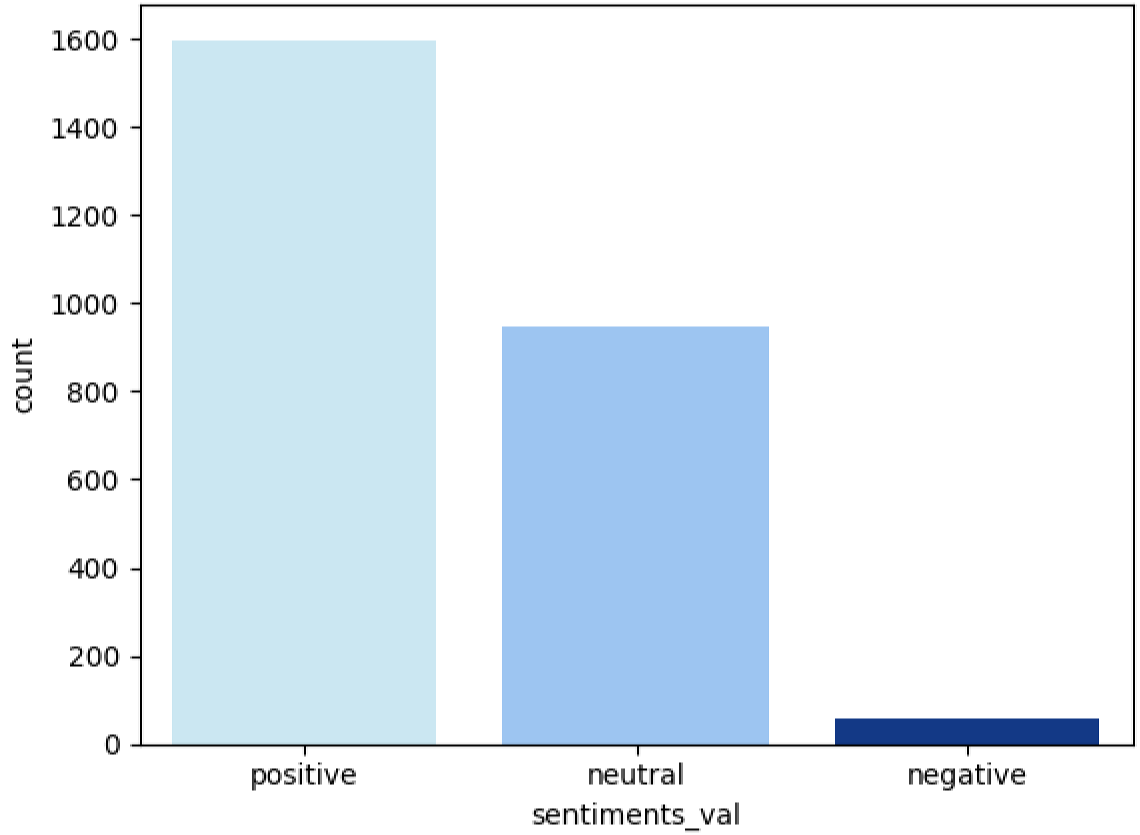
The Clusters



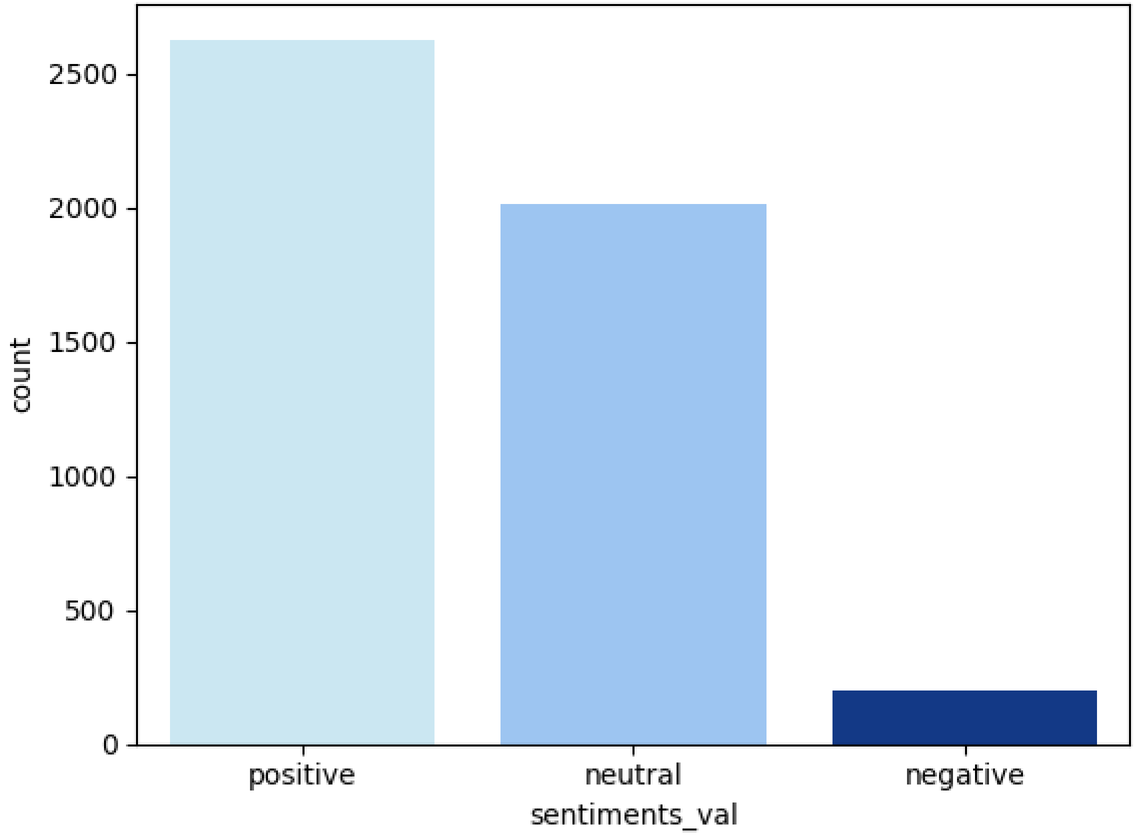


EDA and Visualizations

Sentiments for other brands:
(Ford, BMW, Audi, Tesla, Hyundai)

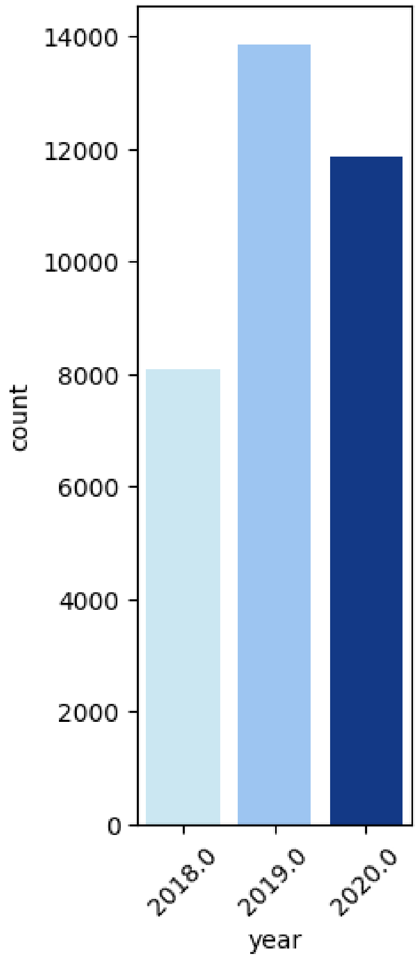


Sentiments for hashtags:
(costs, batteries, climate, fuel, price, tax, afford, money)

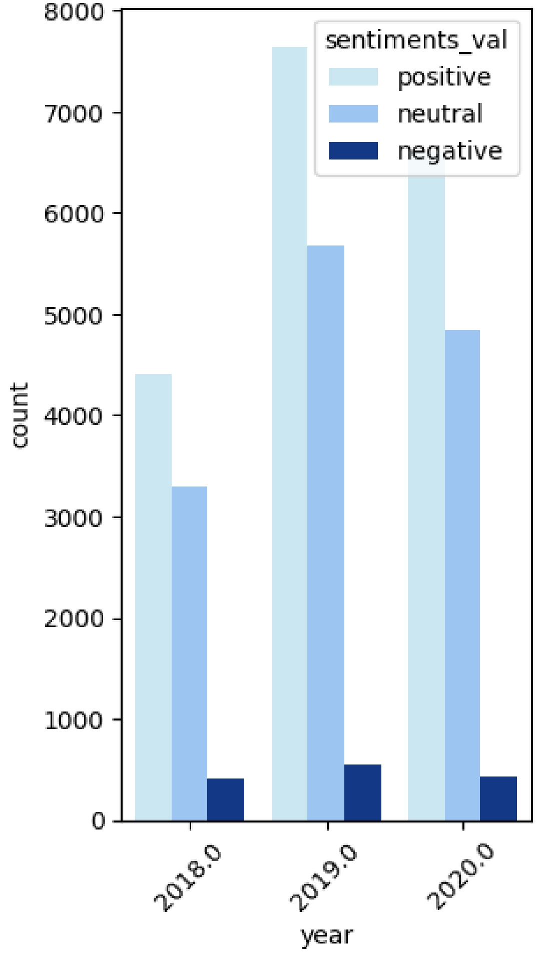


EDA and Visualizations

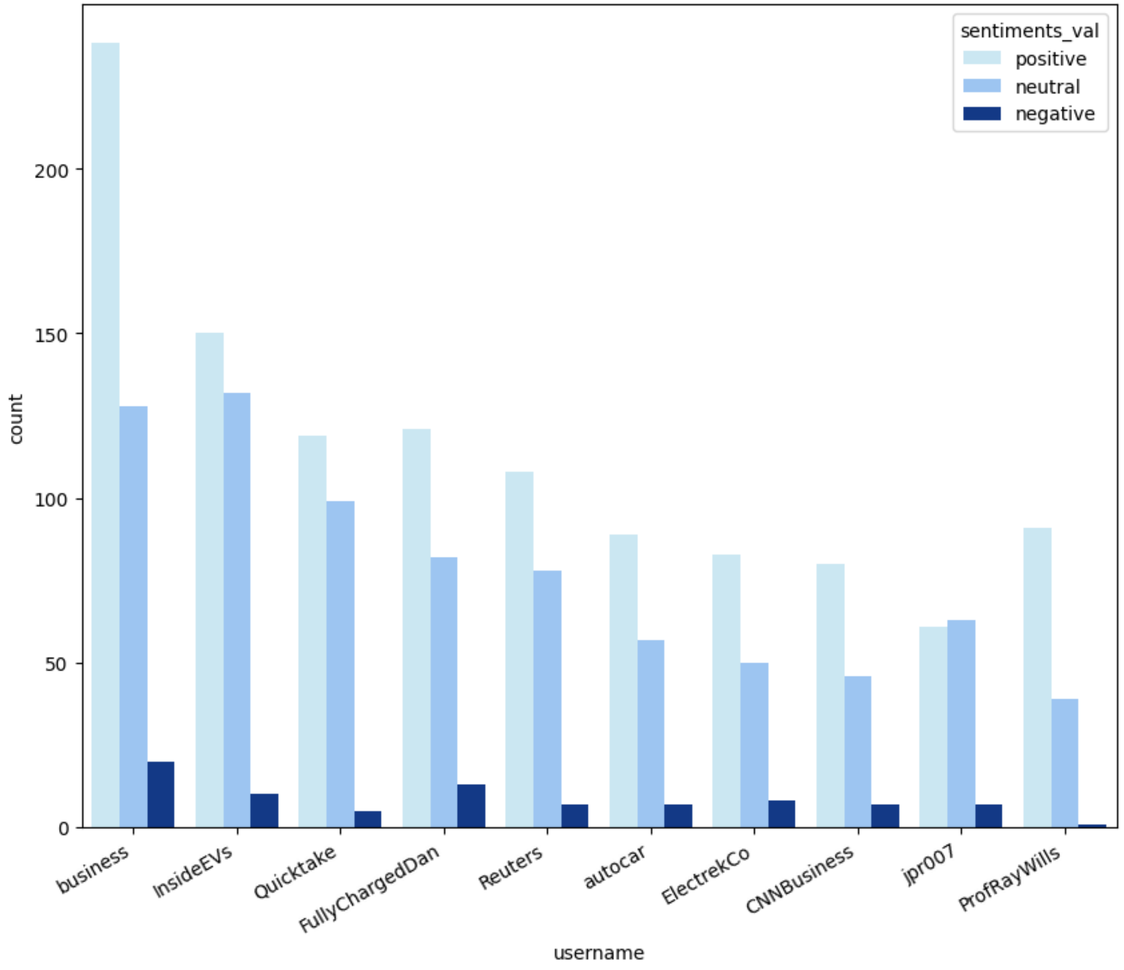
Tweets per Year



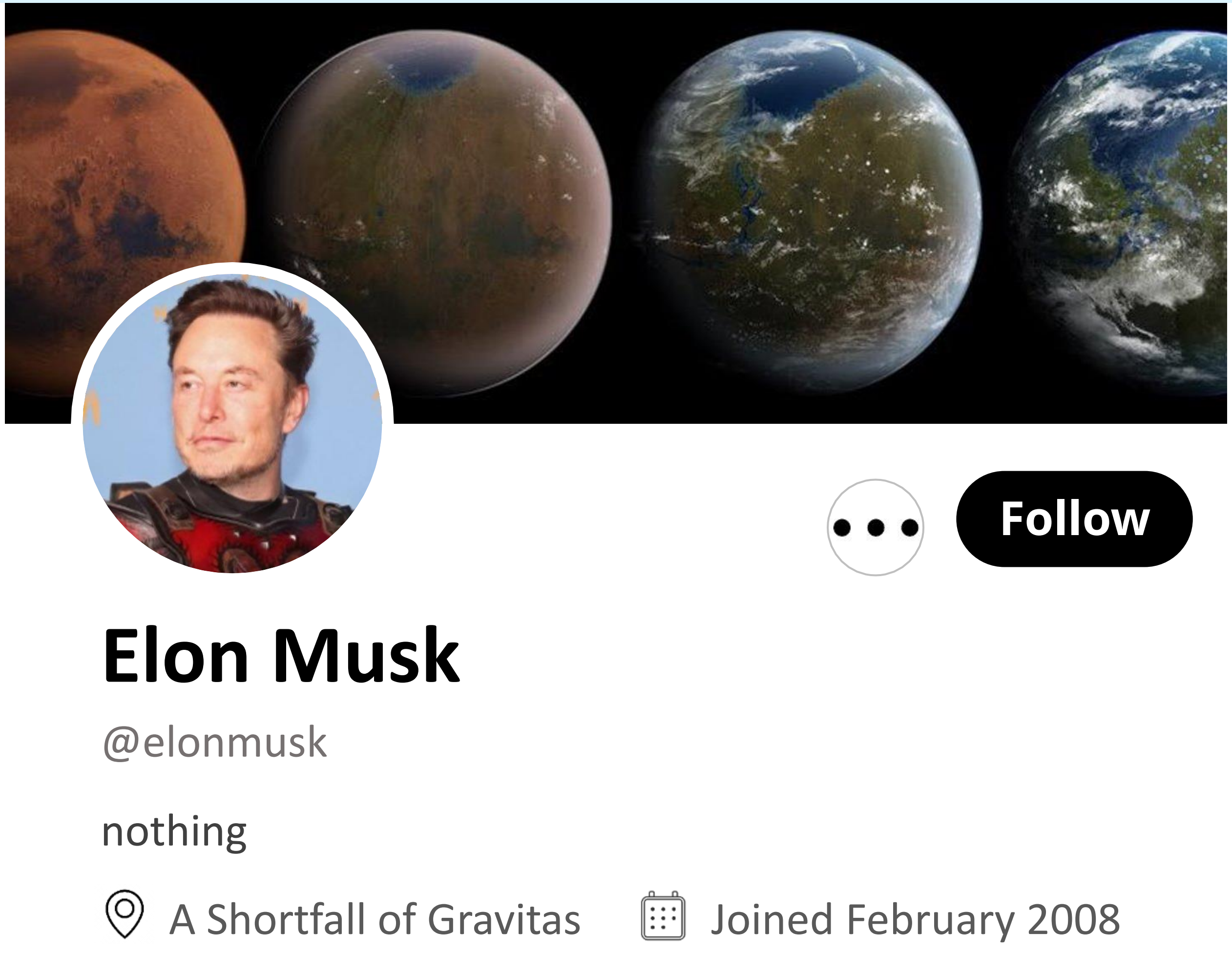
Tweets Sentiments per Year

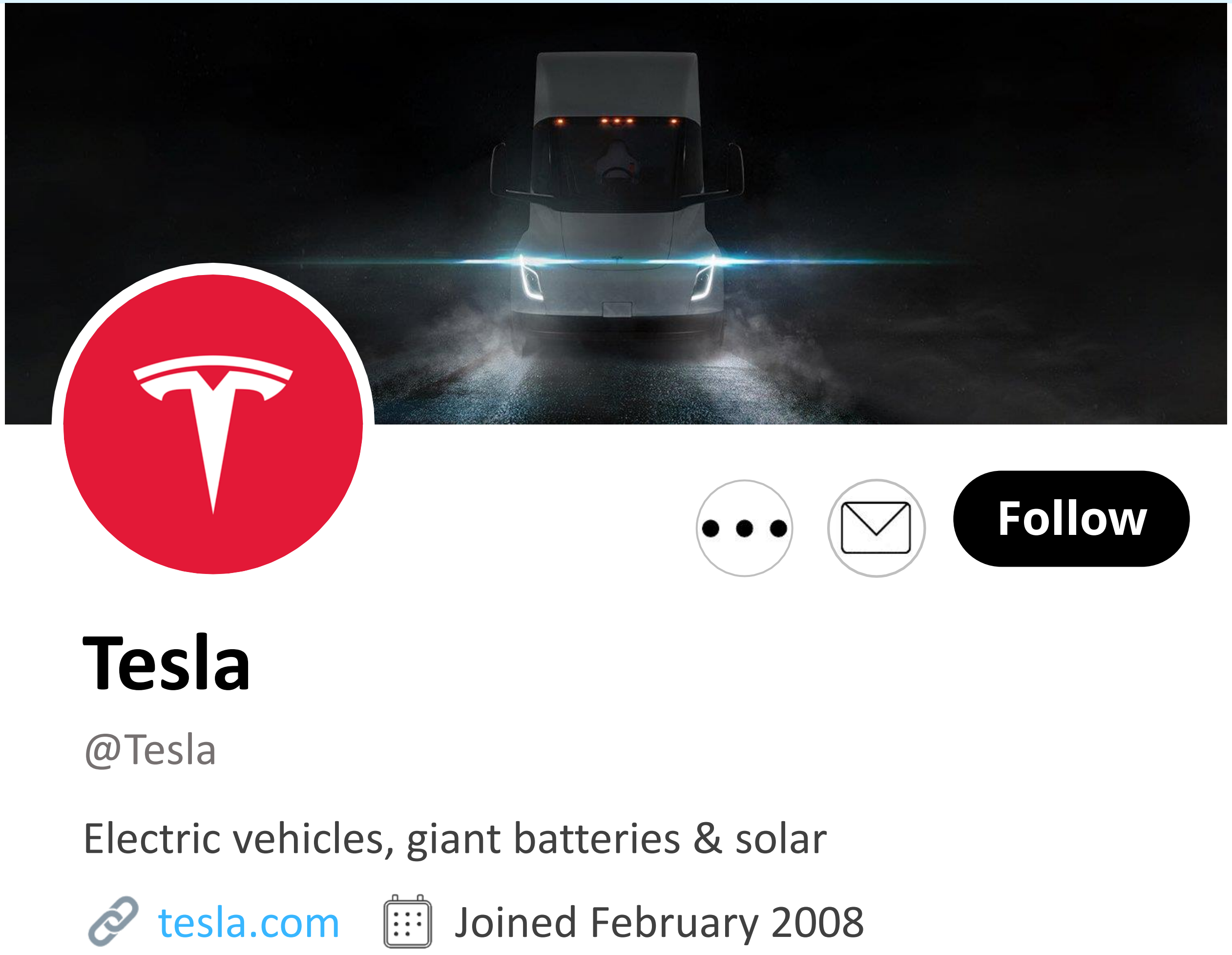


Top 10 Highest Tweeting Usernames



Tesla Stock Association with Sentiments





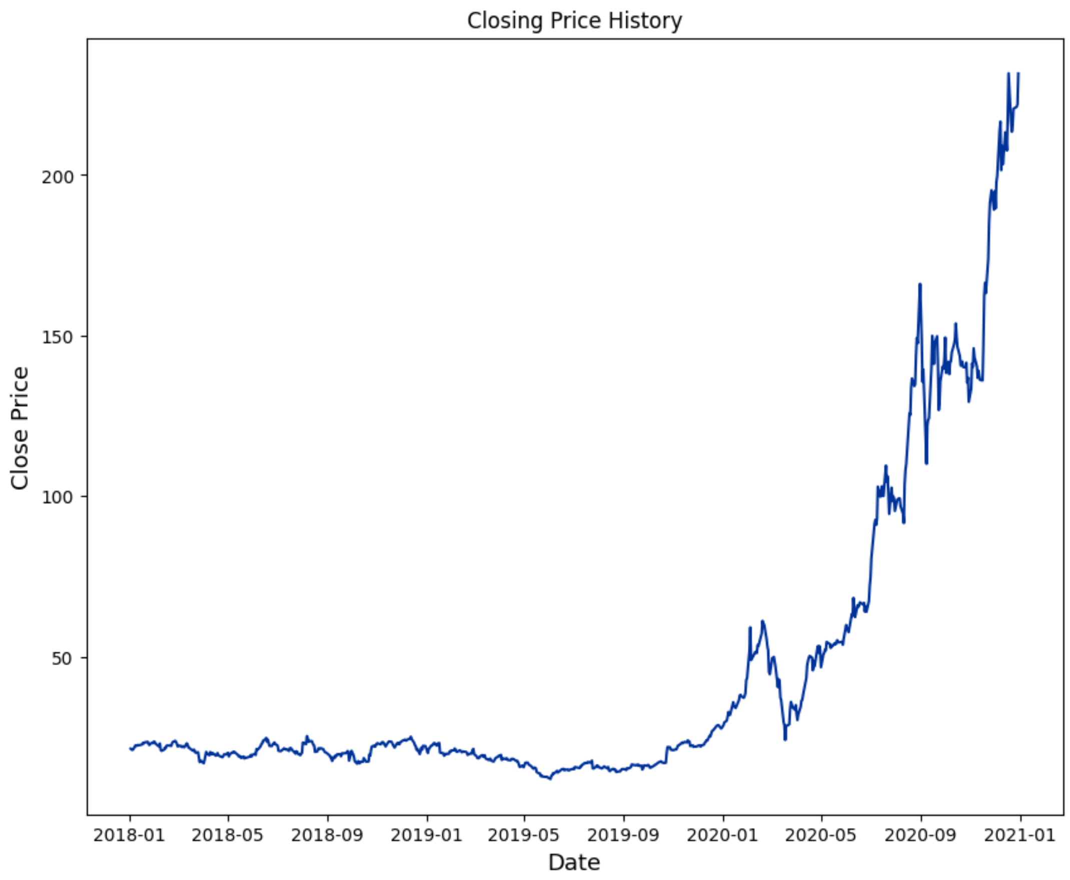
Tesla Dataset

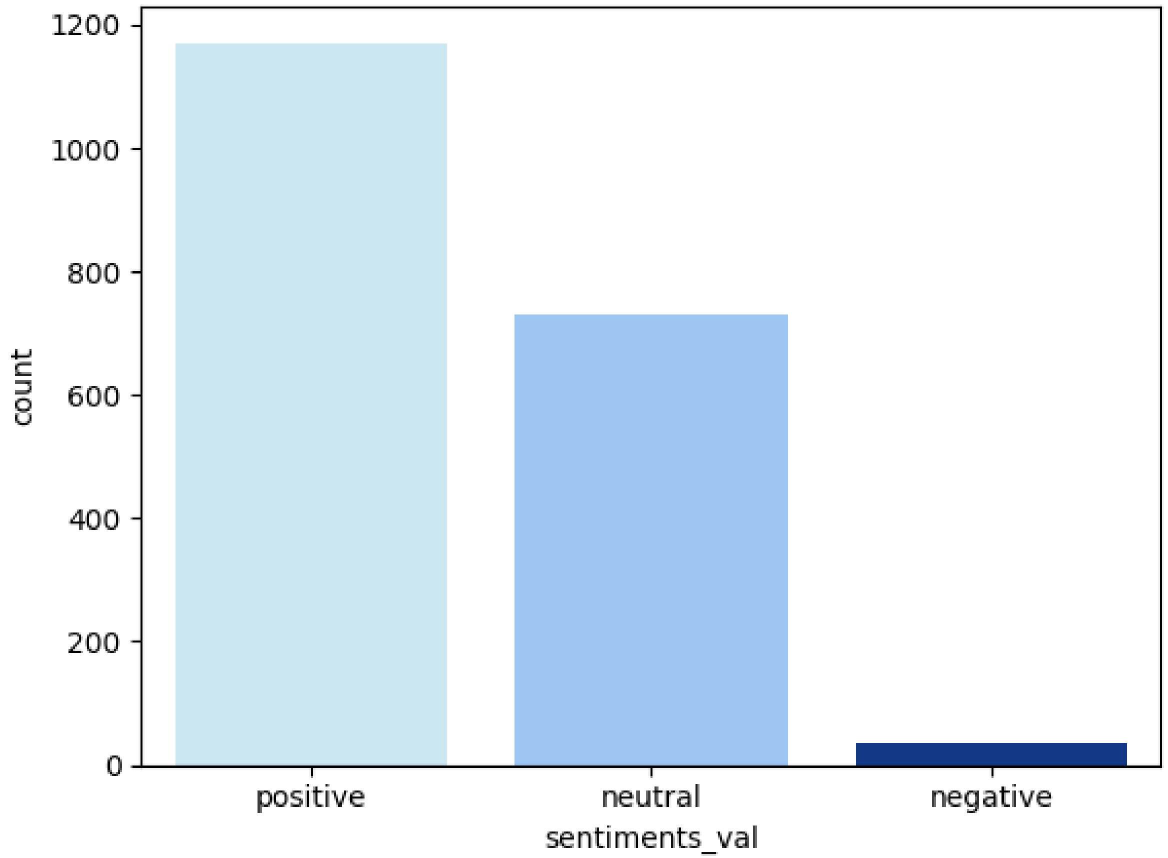
JANUARY 2018						
Sun	Mon	Tue	Wed	Thurs	Fri	Sat
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

DECEMBER 2020						
Sun	Mon	Tue	Wed	Thurs	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Sr. No.	Date	Open	High	Low	Close	Adj Close	Volume	Symbol	Month	Year
0	02-01-2018	20.799999	21.474001	20.733334	21.368668	21.368668	65283000	TSLA	1	2018
1	03-01-2018	21.4	21.683332	21.036667	21.15	21.15	67822500	TSLA	1	2018
2	04-01-2018	20.858	21.236668	20.378668	20.974667	20.974667	149194500	TSLA	1	2018
3	05-01-2018	21.108	21.149332	20.799999	21.105333	21.105333	68868000	TSLA	1	2018
4	08-01-2018	21.066668	22.468	21.033333	22.427334	22.427334	147891000	TSLA	1	2018
5	09-01-2018	22.344	22.586666	21.826668	22.246	22.246	107199000	TSLA	1	2018
6	10-01-2018	22.146667	22.466667	22	22.32	22.32	64648500	TSLA	1	2018
7	11-01-2018	22.349333	22.987333	22.217333	22.530001	22.530001	99682500	TSLA	1	2018
8	12-01-2018	22.575333	22.694	22.244667	22.414667	22.414667	72376500	TSLA	1	2018
9	16-01-2018	22.502666	23	22.32	22.670668	22.670668	97114500	TSLA	1	2018
10	17-01-2018	22.698	23.266666	22.65	23.143999	23.143999	106552500	TSLA	1	2018

EDA

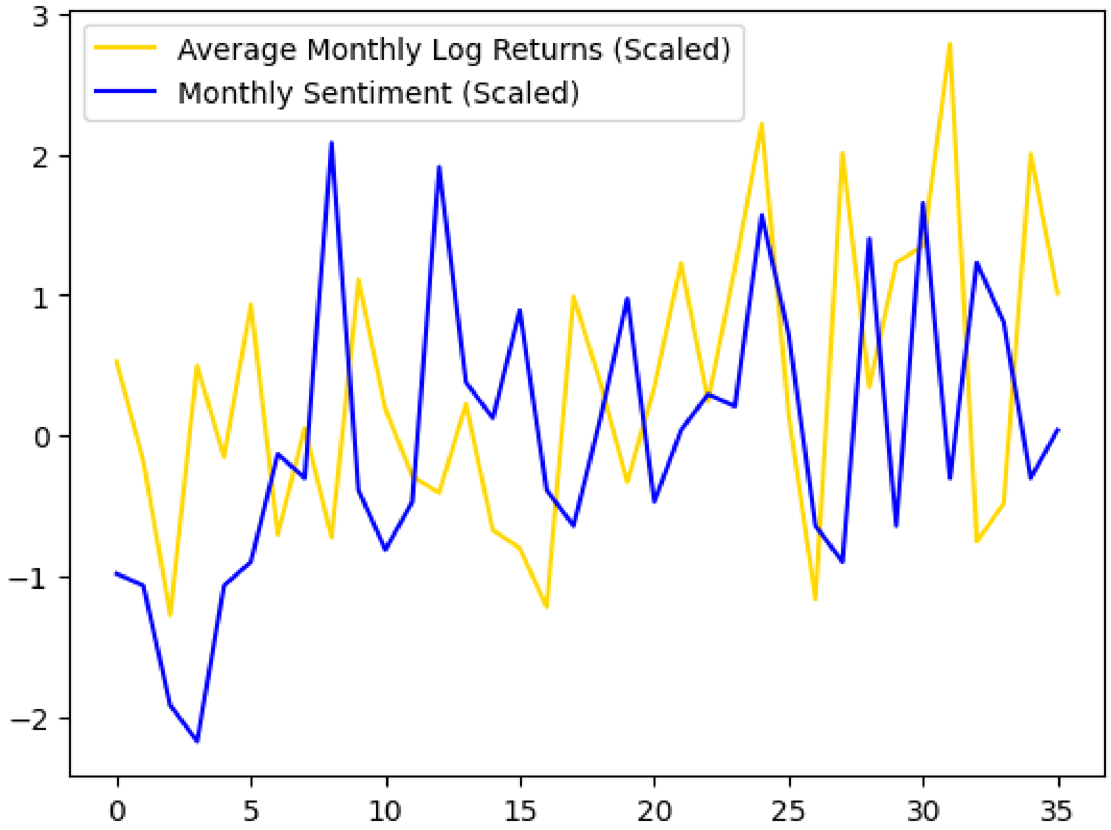




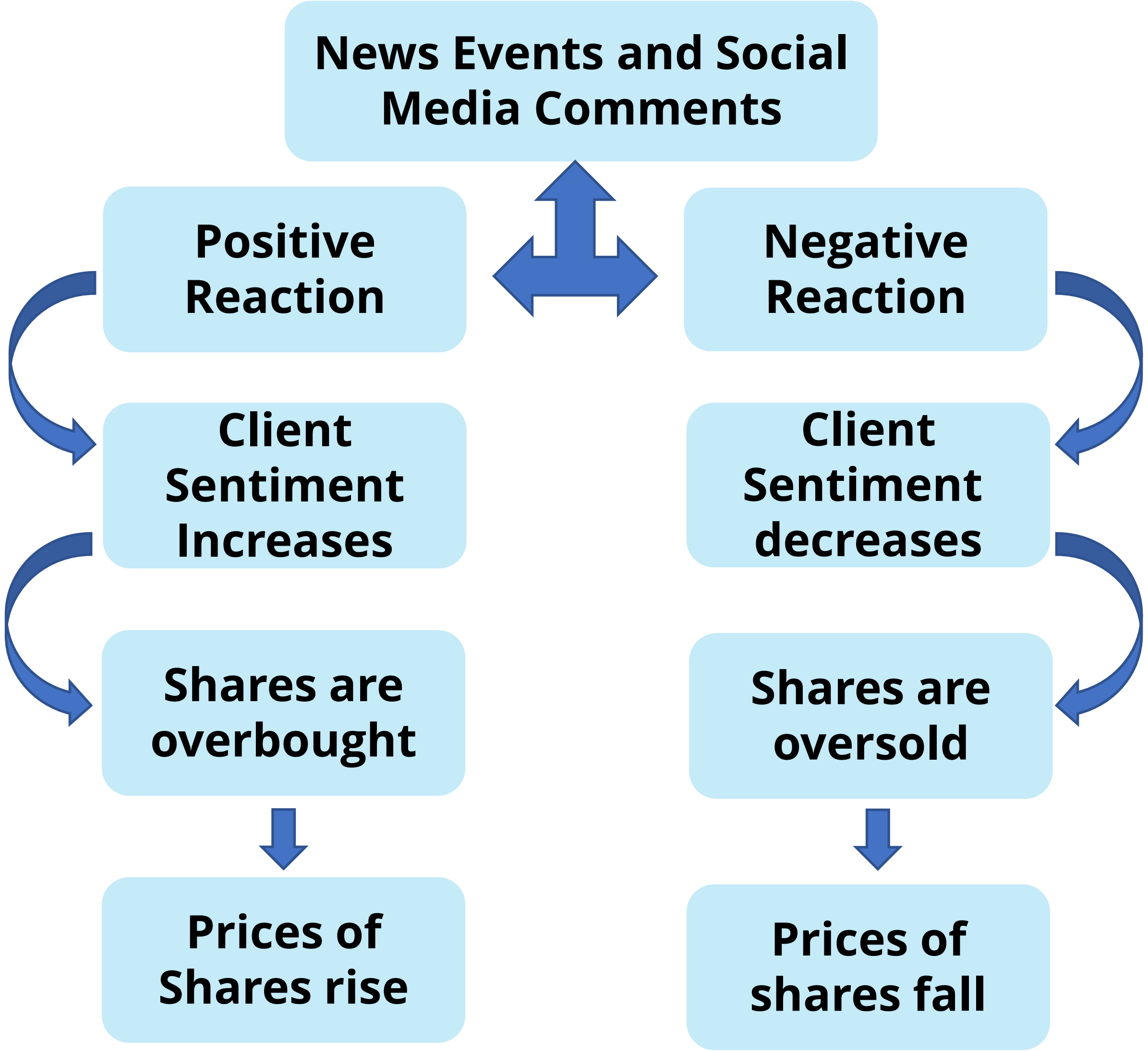
Closing prices of Tesla **Sentiments about Tesla**

Mapping the Sentiment with the Stocks

(Monthly Sentiments + Monthly Log Returns)



Correlation = - 0.04098799



Moving Average Method as
Stock Level Indicators

- Moving Average is a Technical Analysis tool in which the actual index data is compared with its average taken over a period of time.

- We have employed Simple Moving Average (SMA), the periods for moving averages are 5 days, 10 days, and 15 days.

- The main advantages of Moving Average Stock Level Indicator is that it offers a smooth line and also helps to cut down the amount of noise on price chart compared with other level of indicators.

Formula:

$$F_t = \frac{A_t + A_{t-1} + A_{t-2} \dots + A_{t-(n-1)}}{n}$$

$$n : Number\ of\ periods\ to\ be\ averaged$$

$$F_t : n^{th}\ order\ MA\ at\ time\ t$$

$$A_{t-n} : Actual\ occurence\ in\ the\ past\ period\ for\ up\ to\ 'n'\ periods$$

Proposed Predictive System (Moving Average)	Sensex-Moving Average Result
5 day MA > 10 day MA > 15 day MA	Positive
5 day MA < 10 day MA < 15 day MA	Negative
5 day MA < 10 day MA > 15 day MA	Neutral
5 day MA > 10 day MA < 15 day MA	Neutral

The Mapping Relation for Moving Average

Sentiment Analysis Result	Sensex-Moving Average Result	Sentiment + MA
Positive	Positive	Positive
Negative	Negative	Negative
Negative	Positive	Neutral
Positive	Negative	Neutral
Neutral	Positive / Negative / Neutral	Neutral
Positive / Negative / Neutral	Neutral	Neutral

Chi-Square Results

H_0: The sentiments and the direction of stock movements from MA are independent
H_1: There is dependence between sentiments and the direction of stock movements
LOS $= \alpha = 10\% = 0.1$

Sentiment ⟶ MA Result ↓	Negative	Neutral	Positive
Negative	5	20	185
Neutral	9	19	169
Positive	7	18	309

Alpha	0.1
P-value	0.0996

SInce p-value $\leq \alpha$, we reject H_0
Hence, there is dependence between sentiments and the direction of stock movements from MA

Binning Method as Stock Level Indicators

Proposed Predictive System (Binning)	Stock Direction
Closing Value > [Open Value + ADV×Open Value]	Positive
Closing Value < [Open Value – ADV×Open Value]	Negative
Otherwise	Neutral

ADV = Average Daily Variation

Comparisons

Binning

MA Labels

Comparison of direction of stock price movements

Binning

Sentiment + MA

Comparison of direction of stock price movements by adding effect of sentiment on Moving Averages

Conclusion

- Unlike the conventional stock market prediction systems, our novel approach combines the sentiments of common people through the tweets and NYSE data to analyze the behavior of Tesla stock.
- Sentiments and log returns have very low negative correlation of **–0.04.**
- The Moving Average method, when compared to the true values (obtained by binning), gives a percentage match of **37.43%**.
- By taking the sentiments into consideration and re-labelling the stock trends, the percentage match for the same increases to **51.67%**.

Limitations

- This project compares tweets on electric cars with key words like '*Tesla' and 'Teslarati*'. But there could be a pool of tweets outside of the electric car keyword. This might not even be a comparison.

- K means, due to being an unsupervised algorithm, does not give us well defined clusters for our case.

- We considered Tesla stocks, which is listed in the NYSE market. But the twitter data collected comprised of English tweets from all over the world.

References

1. **Multiclass Classification of Tweets based on Kindness Analysis**
   http://cs229.stanford.edu/proj2016/report/HanHuangZhou-MulticlassClassificationOfTweetsBasedOnKindnessAnalysis-report.pdf

2. **Sentiment Analysis for Effective Stock Market Prediction**
   https://www.google.com/searchq=sentiment+analysis+for+effective+stock+market+prediction&domains=inass.org&sitesearch=inass.org

3. **How to Label Unlabelled Tweets – Unsupervised Learning**
   https://medium.com/geekculture/how-to-label-unlabeled-tweets-fb701b97ebf

4. **What is Word2Vec ?**
   https://www.youtube.com/watch?v=lEzzgLh_SFA

5. **Github Link**
   https://github.com/Leal-Miranda/Twitter-Sentiment-Analysis-of-Electric-Vehicles

Thank You