Bitcoin Tweet Sentiment Predictor

Lately, cryptocurrencies have become very popular on the internet. People increasingly find clever ways to move with the market and make quick money by trading one cryptocurrency with the other. However, Bitcoin, the original cryptocurrency, still stands on top and boasts a strong position.

However, Bitcoin followers aren't usually seen in that light and are blamed for being toxic and voicing negative opinions on the internet. This project will build a model that will predict the sentiment of a Bitcoin tweet.

First, we will import all the necessary libraries that'll be used in the context of this project to create the model

In [27]:

```
import numpy as np
import pandas as pd
import re
import nltk
import pickle
import joblib

import utils

# Natural Language Processing imports
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer

# Scikit learn imports
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.base import TransformerMixin, BaseEstimator
from sklearn.pipeline import Pipeline
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score

# Models
from sklearn.linear_model import LogisticRegression, Perceptron, SGDClassifier
from sklearn.naive_bayes import BernoulliNB, MultinomialNB, ComplementNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC, LinearSVC
from sklearn.ensemble import RandomForestClassifier
```

Load Dataset Into Memory

We will first load the dataset int memory and showcase the results. As you can see from the example below, there are quite a number of columns and unnecessary features that we need to remove. This will become a hinderance to the model and will bring unexpected results so we need to clean the data.

In [2]:

```
fullDataset = pd.read_csv("BTC_tweets_daily_example.csv", low_memory=False)
fullDataset
```

Out[2]:

	Unnamed: 0	Date	Tweet	Screen_name	Source	
0	0	Fri Mar 23 00:40:32 +0000 2018	RT @ALXTOKEN: Paul Krugman, Nobel Luddite. I h...	myresumerocket	[]	
1	1	Fri Mar 23 00:40:34 +0000 2018	@lopp @_Kevin_Pham @psycho_sage @naval But @Pr...	BitMocro	[u'Bitcoin']	href="ht
2	2	Fri Mar 23 00:40:35 +0000 2018	RT @tippereconomy: Another use case for #block...	hojachotopur	[u'blockchain', u'Tipper', u'TipperEconomy']	
3	3	Fri Mar 23 00:40:36 +0000 2018	free coins https://t.co/DiuoePJdap	denies_distro	[]	
4	4	Fri Mar 23 00:40:36 +0000 2018	RT @payvxofficial: WE are happy to announce th...	aditzgraha	[]	href="ht
...	...	...	...	...	...	
50882	50854	Fri Mar 23 08:55:16 +0000 2018	RT @fixy_app: Fixy Network brings popular cryp...	quoting_lives	[]	
50883	50855	Fri Mar 23 08:55:17 +0000 2018	RT @bethereumteam: After a successful launch o...	VariPewitt	[]	
50884	50856	Fri Mar 23 08:55:18 +0000 2018	RT @GymRewards: Buy #GYMRewards Tokens, Bonus ...	urbancoinerz	[u'GYMRewards', u'ICO', u'cryptocurrency', u'm...	
50885	50857	Fri Mar 23 08:55:19 +0000 2018	I added a video to a @YouTube playlist https:/...	MRDanishShahab	[]	
50886	50858	Fri Mar 23 08:55:19 +0000 2018	RT @Raybambs: Airdrop PhotoCoin Airdrop Round#...	Azriel020	[u'PhotoCoin']	href="ht

50887 rows × 16 columns

Cleaning The Dataset

We now have to clean the dataset to proceed with the project. Following things are at fault with the original dataset:

- There are a unnecessary amount of features
- There are some null values far into the dataset
- There are invalid values in the "Sentiment" column, which should only accept "positive", "neutral" and "negative"
- The sentiment column stores sentiment strings as stringified lists. We need to change that

The following changes will be made:

- All columns except "Tweet" and "Sentiment" will be dropped.
- Null values will be dropped
- Sentiment values will be converted from stringified lists to simple strings e.g. ['positive'] => positive
- Unknown values from "Sentiment" column will be deleted
- Cleaned dataset will be saved as cleaned_btc_tweets.csv file

In [3]:

```
# Dropping all unneeded columns except tweet and sentiment colums
btc_tweets = fullDataset.drop(fullDataset.columns[[0, 1, 3, 4, 5, 7, 8, 9, 10, 11,

btc_tweets = btc_tweets.dropna()
# btc_tweets = btc_tweets.head(10000)

indices = []

# This loop clears all sentiment values that aren't known
for index, row in btc_tweets.iterrows():

    sentiment = btc_tweets['Sentiment'][index].strip('][\'')

    if sentiment == "positive":
        btc_tweets.loc[index, 'Sentiment'] = sentiment
        continue
    if sentiment == "neutral":
        btc_tweets.loc[index, 'Sentiment'] = sentiment
        continue
    if sentiment == "negative":
        btc_tweets.loc[index, 'Sentiment'] = sentiment
        continue

    indices.append(index)

btc_tweets = btc_tweets.drop(labels=indices, axis=0)

btc_tweets.to_csv(r'cleaned_btc_tweets.csv', index = False)

print("After cleaning the sentiment column:")
pd.options.display.max_colwidth = 200
btc_tweets
```

```
After cleaning the sentiment column:
```

Out[3]:

	Tweet	Sentiment
0	RT @ALXTOKEN: Paul Krugman, Nobel Luddite. I had to tweak the nose of this Bitcoin enemy. He says such foolish things. Here's the link: htt…	neutral
1	@lopp @_Kevin_Pham @psycho_sage @naval But @ProfFaustus (dum b a ss) said you know nothing about #Bitcoin ... 😂😂😂 https://t.co/SBAMFQ2Yiy	neutral
2	RT @tippereconomy: Another use case for #blockchain and #Tipper. The #TipperEconomy can unseat Facebook and change everything! ICO Live No…	positive
3	free coins https://t.co/DiuoePJdap	positive
4	RT @payvxofficial: WE are happy to announce that PayVX Presale Phase 1 is now LIVE!\n\nSign up -->> https://t.co/dhprzsSxek\nCurrencies accept…	positive
...	...	...
50882	RT @fixy_app: Fixy Network brings popular cryptocurrencies and retailers as partners with benefits from blockchain. Partner Stores will acc…	positive
50883	RT @bethereumteam: After a successful launch of our Bounty campaign, we've managed to filter out the Bounty related questions to: https://t…	positive
50884	RT @GymRewards: Buy #GYMRewards Tokens, Bonus Time is ending! https://t.co/HDvhoZrz2J, #ICO #cryptocurrency #mobile #app #mining #exercisin…	neutral

	Tweet	Sentiment
50885	I added a video to a @YouTube playlist https://t.co/ntFJrNvSvZ How To Bitcoin Cloud Mining Free For Lifetime Urdu / Hindi	positive
50886	RT @Raybambs: Airdrop PhotoCoin Airdrop Round#2. 100 #PhotoCoin will be giving to everyone who complete the google form. Your account will…	positive

49751 rows × 2 columns

Checking Balance of Sentiments in Dataset

As we can see above, we have 50,886 total instances in the dataset even after cleaning. Now we check the balance of ratios between the instances with respect to sentiment so we have around equal of each sentiment. This step is crucial to designing a good data model. We have to check the ratio at which our tweets are divided by sentiment into the dataset.

This is important because if one of the sentiments is less in number with respect to the others, the model may get trained on inaccurate data and, hence, provide inaccurate results

The following loop will count all sentiment instances in the cleaned dataset:

In [4]:

```
positiveCount = 0
neutralCount = 0
negativeCount = 0

for index, row in btc_tweets.iterrows():
    sentiment = btc_tweets['Sentiment'][index].strip('][\'')

    if sentiment == "positive":
        positiveCount += 1

    if sentiment == "neutral":
        neutralCount += 1

    if sentiment == "negative":
        negativeCount += 1

print("Positive Tweets:", positiveCount)
print("Neutral Tweets:", neutralCount)
print("Negative Tweets:", negativeCount)
```

```
Positive Tweets: 22656
Neutral Tweets: 21150
Negative Tweets: 5945
```

Data is Unbalanced

As we can see, negative sentiment tweets are almost 4 times lower than positive and neutral sentiments. This will affect the prediction results of the model if we train it on unbalanced data.

For this reason, we will cut down positive and neutral tweets until they match exactly the negative tweets' number and then proceed to the next step. The following loop does exactly that:

In [5]:

```
newDF = pd.DataFrame(columns=["Tweet", "Sentiment"])

positiveCount = 0
neutralCount = 0

for index, row in btc_tweets.iterrows():
    sentiment = btc_tweets['Sentiment'][index]

    if sentiment == "positive":
        if positiveCount == negativeCount:
            continue
        positiveCount += 1
        newDF.loc[len(newDF)]=[row["Tweet"], row["Sentiment"]]

    if sentiment == "neutral":
        if neutralCount == negativeCount:
            continue
        neutralCount += 1
        newDF.loc[len(newDF)]=[row["Tweet"], row["Sentiment"]]

    if sentiment == "negative":
        newDF.loc[len(newDF)]=[row["Tweet"], row["Sentiment"]]

newDF
```

Out[5]:

	Tweet	Sentiment
0	RT @ALXTOKEN: Paul Krugman, Nobel Luddite. I had to tweak the nose of this Bitcoin enemy. He says such foolish things. Here's the link: htt…	neutral
1	@lopp @_Kevin_Pham @psycho_sage @naval But @ProfFaustus (dum b a ss) said you know nothing about #Bitcoin ... 😂😂😂 https://t.co/SBAMFQ2Yiy	neutral
2	RT @tippereconomy: Another use case for #blockchain and #Tipper. The #TipperEconomy can unseat Facebook and change everything! ICO Live No…	positive
3	free coins https://t.co/DiuoePJdap	positive
4	RT @payvxofficial: WE are happy to announce that PayVX Presale Phase 1 is now LIVE!\n\nSign up -->> https://t.co/dhprzsSxek\nCurrencies accept…	positive
...	...	...
17830	RT @PumaPay: Why Did Credit Cards Fail to Adopt to the Modern Needs? https://t.co/u1qB3gxA3T #pumapay #creditcards #banking #finance #block…	negative
17831	Bitcoin Will Be World's 'Single Currency' Says Twitter CEO https://t.co/f4hsEbLgkk https://t.co/P3fuHSLwkX	negative
17832	RT @CloudMiningX: Use the code: HF18BDAY30 at purchase to get a 30% discount for all contracts. The offer is limited. \n\n10 Ghs = 0.84$\n1000…	negative
17833	Twitter CEO Says Bitcoin Will Be World’s ‘Single Currency’ Within A Decade https://t.co/2obg7hKwm5	negative
17834	RT @UTEMISUTS: Decentralizing businesses reputation enables Latin American companies that have never met each other to conduct internationa…	negative

17835 rows × 2 columns

Check Balance Again

Balancing the tweets has reduced our dataset to 17,834 instances. We will check balance again now after balancing the sentiments with each other:

In [6]:

```
btc_tweets = newDF

positiveCount = 0
neutralCount = 0
negativeCount = 0

for index, row in btc_tweets.iterrows():
    sentiment = btc_tweets['Sentiment'][index].strip('][\'')

    if sentiment == "positive":
        positiveCount += 1

    if sentiment == "neutral":
        neutralCount += 1

    if sentiment == "negative":
        negativeCount += 1

print("Positive Tweets:", positiveCount)
print("Neutral Tweets:", neutralCount)
print("Negative Tweets:", negativeCount)
```

```
Positive Tweets: 5945
Neutral Tweets: 5945
Negative Tweets: 5945
```

Data is Balanced

The data has successfully been balanced and now we have 5,945 instances of each sentiment in the dataset. Now we can proceed to split between training and testing data.

The following code block will do this:

- First, separate dataframes with both all columns will be made with 80% training data and 20% testing data. This data will be saved as .csv files.
- Then, the original data will be split again. Now we need to separate with 80 - 20% difference as well separate the columns. This is crucial in the code to follow after this step

In [12]:

```
# Creating csv's for presentation purpose
train_tosave, test_tosave = train_test_split(btc_tweets, test_size=0.2)

train_tosave.to_csv(r'train_set.csv', index = False)
test_tosave.to_csv(r'test_set.csv', index = False)

# 80% training set, 20% testing set
tweets = btc_tweets["Tweet"]
sentiments = btc_tweets["Sentiment"]

train_data, test_data, train_sentiment, test_sentiment = train_test_split(tweets, s

rand_indexs = np.random.randint(1,len(train_data),50).tolist()

print("Number of training instances:", len(train_data.index))
print("Number of testing instances:", len(test_data.index))
```

```
Number of training instances: 14268
Number of testing instances: 3567
```

Training Data

In [13]:

```
print("Training data:")
train_data.head(60)
```

```
Training data:
```

Out[13]:

```
4538      Name: Raiden Network Token\nSymbol: RDN\n24 hour change: -
4.69%\nPrice: 1.58653\nRank: 129\nTotal Supply: 100000000.0\nVo… htt
ps://t.co/KHrEl2usc2 (https://t.co/KHrEl2usc2)
3436                    RT @cryptomsn: Home of Bitcoin Crypto Currenc
y - https://t.co/O0gPKD0Ie8 (https://t.co/O0gPKD0Ie8) \n#BTC #Crypto
CurrencyNews #Eth #Ltc https://t.co/76hvwd6DlH (https://t.co/76hvwd6
DlH)
8986         RT @OnWindowly: Lightning Network Problems — wow!\n#Bit
coinCash is #Bitcoin\n\n@el33th4xor  @ Satoshi Vision Conference in
 Tokyo, Japan. https…
5119             RT @DrDenaGrayson: @ericgeller Agree w/@ericgeller
👉likely signals #indictments of state-backed hackers. I believe tha
t these hackers will…
17443        RT @CloudMiningX: Use the code: HF18BDAY30 at purchase
 to get a 30% discount for all contracts. The offer is limited. \n\n
10 Ghs = 0.84$\n1000…
12025     Name: CRYPTO20\nSymbol: C20\n24 hour change: -7.7%\nPrice:
1.29942\nRank: 168\nTotal Supply: 40656082.0\nVolume: 2603890.… http
s://t.co/h3G19w0ywl (https://t.co/h3G19w0ywl)
14952                                        Current Bitcoin Rate in U
SD : 8,414.4538 Check other Currencies: https://t.co/KqQpwIzXrs (htt
ps://t.co/KqQpwIzXrs) #BitsRate #BTC #Bitcoins
5168     RT @UppercoinC: #Giveaway $100 in $ETH!\n\n-Like\n-Retweet
\n-Follow\n-Comment down below with your #Ethereum (preferably) ERC2
0 address.\n\nLike a…
14985      Name: Metaverse ETP\nSymbol: ETP\n24 hour change: -11.43%
\nPrice: 0.895199\nRank: 217\nTotal Supply: 57379980.0\nVolume:… htt
ps://t.co/oHYFNZYod2 (https://t.co/oHYFNZYod2)
10767             RT @hackinjeebs: @LouiseBagshawe @LoolooMagdalena @
patribotics I’m hoping for a bitcoin announcement. Something that re
ferences that block…
10570           Britain Introduces Crypto Task Force To Foster Finte
ch Innovation #cryptocurrency #crypto #bitcoin #altcoin #inve... htt
ps://t.co/kEaQVJe6PF (https://t.co/kEaQVJe6PF)
377                                            ICE Agency Charges P
ayza and Two Canadian Citizens With Bitcoin Money Laundering #ico #c
ryptocurrency #token
4018
EBay ‘Seriously Considering’ Adding #Bitcoin Payments https://t.co/7
SgygfezrI (https://t.co/7SgygfezrI)
4755                                                    @CryptoT
rendy I conceive you can be a unique bitcoin expert please visit htt
ps://t.co/6rortIStnD (https://t.co/6rortIStnD)
4789     RT @MsxNetwork: Last day for #AION token holder!\n\nCheck A
NN thread for more detail \nhttps://t.co/8Mbvhqbvem\n\nDo not miss i
t!\n\n#microstack #…
243           RT @OfficialMusards: Our Telegram community is reachin
g 28k users...\nWe want to thank you all!\nNew features and news on
 our website, coming…
5517          RT @PeerMountain: Check out again our video to underst
and how Peer Mountain works :)\n\n#blockchain #cryptocurrency #bitco
in #ethereum #Techn…
```

```
8409            RT @bethereumteam: Have you seen any of the Animated
Motion Pictures that are nominated for the Oscars? Would you bet tha
t you can guess th…
1332                                    Tom Lee: “The Altcoin Bear Ra
lly is Almost at an End,” but still Stick to Bitcoin\nhttps://t.co/s
jyoiczwwI\n$BTC #BTC
4442            @zabala_jeric @buzzshownetwork Bitcoin Gold (Bitcoin
-Gold) do  gaining popular also apt more expensive, marely Your… htt
ps://t.co/0Ayk8TtEfk (https://t.co/0Ayk8TtEfk)
3248      Name: COSS\nSymbol: COSS\n24 hour change: -7.4%\nPrice: 0.
254438\nRank: 335\nTotal Supply: 104000000.0\nVolume: 1023840.0… htt
ps://t.co/a2EvYrM2ly (https://t.co/a2EvYrM2ly)
5310            RT @bethereumteam: We already have an iOS prototype
 of the betting process through Bethereum on the AppStore! Are you r
eady to challenge yo…
15900                                                                Bit
coin Loses $9k Support After Binance Confusion Shakes Confidence htt
ps://t.co/gwV5C80LUc (https://t.co/gwV5C80LUc)
4588      Name: Asch\nSymbol: XAS\n24 hour change: -12.34%\nPrice:
 0.720606\nRank: 141\nTotal Supply: 114855331.0\nVolume: 4978880.… h
ttps://t.co/piGLGJ9X92 (https://t.co/piGLGJ9X92)
7733       RT @PhotoCoin_io: 2,000,000 PHT TOKEN #airdrop \n1. Follo
w \n2. Like\n3. Retweet ,tag 5 Friends with #PHT\n4. Comment your ET
H address \nTotal s…
5545            #bitcoin (Coincheck: NEM Foundation Stops Tracing St
olen Coins, Hackers’ Account At Zero) has been published on Bit… htt
ps://t.co/ExNb1Pe5To (https://t.co/ExNb1Pe5To)
9912            RT @adamludwin: 1/Satoshi said Bitcoin was for “comm
erce on the internet” (the first four words of the Bitcoin whitepape
r). Turns out she w…
3435           RT @bethereumteam: Do you remember the last time you
 had fun while betting?\nWell, with #Bethereum you will! We're bring
ing the holy trinity…
4159           RT @MinerGate: Why Proof-of-capacity could be the fut
ure of #cryptocurrency? The answer you will find in our new blog pos
t:\n➡🖊️https://t.c…
8887                                    #bitcoin Growing mistrust thre
atens Facebook after data mining scandal https://t.co/63ARp5eBQO (ht
tps://t.co/63ARp5eBQO) https://t.co/UwuYRYF2V8 (https://t.co/UwuYRYF
2V8)
15916
@aliraja How can you 'predict' when bitcoin goes down?
11614         RT @bethereumteam: We're revealing our surprise tomorr
ow! \nAre you ready to celebrate with us? 😮\n#surprise #presents #c
rypto #bitcoin #ethe…
5227          RT @bethereumteam: Checkout an interview with our tea
m!\nWe feel that interviews get us even closer to our community. 🤝
\nhttps://t.co/AUpyskY…
3738         RT @izx_io: #TOKEN2049 is finished💫\nWe want to thank
 everybody who supported us in Hong Kong!🙌\n\n#izx #izetex #blockch
ain #VR #AR #vrparks #…
13704           RT @XVGDolphin: @WinwithRick We are a crypto communi
ty and we stick together, one small step taken by thousands will mak
e Verge known to Mi…
15531            RT @PeerMountain: Security requires safeguards that
make it difficult or impossible for criminals to gain access to your
information. Read…
205             #power #SelectionSunday #Cryptopia #CryptoNews #cryp
tocurency #ICO #Bitcoin #ethereum #blockchaintechnology #Market… htt
ps://t.co/8tn2Ouektj (https://t.co/8tn2Ouektj)
1675        Bitcoin (-0.15): $8,716.30\nEthereum (0.1): $540.09\nRip
```

```
ple (-0.62): $0.66\nBitcoin Cash (0.54): $1,017.03\nLitecoin (-… https://t.co/s2R2uEbCXR (https://t.co/s2R2uEbCXR)
3590                                                  I added a video to a @YouTube playlist https://t.co/jRPJqOQsL7 (https://t.co/jRPJqOQsL7) I kissed a Bitcoin and I liked it! 💋
5164                                          Earning $100K Mining Bitcoin Ethereum ZCash! — Mining BTC ETH ZEC — CryptoCurrency Mining https://t.co/5G4WDWa4MA (https://t.co/5G4WDWa4MA)
4145           RT @bethereumteam: Create custom group bets and invite your friends, choose the buy in amounts, select what sport to bet on and enjoy the g…
13721          RT @InvResDynamics: SPX down another 20 points from the close.  10yr yield below 2.80.  Gold is flying.  Bitcoin is tanking.  Markets are r…
2064                    Our goal is to be the best cryptocurrency trading platform in the world for traders https://t.co/EOjqDZfjNo… (https://t.co/EOjqDZfjNo…) https://t.co/zRICEPeZs0 (https://t.co/zRICEPeZs0)
3165       Name: BlockMason Credit Protocol\nSymbol: BCPT\n24 hour change: -9.14%\nPrice: 0.493775\nRank: 220\nTotal Supply: 116158… https://t.co/6NZzYxPAyj (https://t.co/6NZzYxPAyj)
11996          RT @CoinbayExchange: Giveaway still on!!🎉🎊🎆🎉 Our #Airdrop is worth $1,000,000 and started on 18th March. https://t.co/muULREbL0Z (https://t.co/muULREbL0Z) RETWEET an…
1704                                                  Researchers Discover Child Pornography Hidden in Bitcoin’s Blockchain  #btc https://t.co/dQtgUu1Qrd (https://t.co/dQtgUu1Qrd)
13688          @CoinRaffles I'm thinking of putting my current #bitcoin giveaways over to your system because it automates the who… https://t.co/Gir5oAzBcL (https://t.co/Gir5oAzBcL)
15690           RT @muirfieldip: "TAOs... offer enhanced safety for investors and accountability for the issuing firm." - Tom Zaccagnino (@tomzaccagnino),…
7472      RT @truegameSRL: 🚀 Truegame team is fully ready to conquer Tallinn! ✔\n\nTomorrow we'll be attending Blockchain &amp;amp; Bitcoin Conference Tallinn…
8628           RT @Marvel_euphoria: Enough is enough. Chinese govt tell companies who are falsefuly claiming Blockchain affiliation just for the sake of b…
15694           RT @PeerMountain: Security requires safeguards that make it difficult or impossible for criminals to gain access to your information. Read…
8037          RT @WealthE_Coin: Have you ever had the question “What is Bitcoin?”...check out this video #Bitcoin #Blockchain #WealthMigrate #WealthE \nht…
904                                     Ben is a chatbot that lets you learn about and buy Bitcoin https://t.co/qNp7qNKSfd (https://t.co/qNp7qNKSfd) #CryptocurrencyNews #bitcoin #Bots
16511                                     WSJ: #SEC To Examine Up To 100 Crypto-Related Hedge Funds: In the US, the… https://t.co/feQqjx7ffI (https://t.co/feQqjx7ffI) #bitcoin #crypto
16664          I followed this one “crypto guru” as a joke and then every day some other crypto account follows me. This one is my… https://t.co/SCNQvlcn11 (https://t.co/SCNQvlcn11)
17467     RT @btccloud: Official Bitcoin Cloud #Airdrop 1\n\nhttps://t.co/wyumnhGR49\n\nLimited to 10,000 Members\nTotal Supply # 20M\n4M Coins Will be Air…
10830          RT @bethereumteam: Have you seen any of the Animated Motion Pictures that are nominated for the Oscars? Would you bet that you can guess th…
12590          RT @Marvel_euphoria: You know what else makes sense,
```

```
is making Millions in 💰💰. A spokesperson has announced on Thursday
that, the U.S. Mars…
7747                          💯 Earn FREE Bitcoin https://t.co/aFMj2a1
5LU (https://t.co/aFMj2a15LU) 💰 #BTCPeek #Btc #BitcoinB #BtcSales #
BitcoinPh #BitcoinPoker… https://t.co/yxsM08r35V (https://t.co/yxsM0
8r35V)
10240                                      RT @thehackfund: Investors b
ullish on bitcoin now that the 'Tokyo Whale' has stopped selling htt
ps://t.co/3ZCULmPqmM (https://t.co/3ZCULmPqmM)
Name: Tweet, dtype: object
```

Testing Data

In [14]:

```
print("Testing data:")
test_data.head(60)
```

```
Testing data:
```

Out[14]:

```
12501             #Blockchain simplified: @CBinsights / #crypto #finte
ch #bitcoin, #ICO https://t.co/WuEnTKIq6r (https://t.co/WuEnTKIq6r) /
 @BourseetTrading… https://t.co/tuU6Yjy2Vh (https://t.co/tuU6Yjy2Vh)
607      Name: Tidex Token\nSymbol: TDX\n24 hour change: -38.18%\nPric
e: 0.312157\nRank: 647\nTotal Supply: 10000000.0\nVolume: 29… https://
t.co/UvqZKkplhU (https://t.co/UvqZKkplhU)
13116                 [USD] 23/03/2018 03:00:01 Bitcoin: $8451.49 Ethe
reum: $516.85  #bitcoin #ethereum #altcoin #coin #blockchain… https://
t.co/dZ4Wv0WcS5 (https://t.co/dZ4Wv0WcS5)
10555      RT @CherylPreheim: City of Atlanta’s computers being held h
ostage by hacker demanding $51,000 ransom in bitcoin. FBI &amp;amp; Homela
nd Security in…
15317                 RT @ErikVoorhees: CNBC: Jack Dorsey expects bitc
oin to become the world's 'single currency' in about 10 years https://
t.co/ERONOX5cH1 (https://t.co/ERONOX5cH1)
13995    Name: AdEx\nSymbol: ADX\n24 hour change: -8.72%\nPrice: 0.754
066\nRank: 155\nTotal Supply: 100000000.0\nVolume: 6612940.0… https://
t.co/ghfLGDKUgr (https://t.co/ghfLGDKUgr)
2243                                  Why Blockchain Will Survive Ev
en If Bitcoin Doesn't\nhttps://t.co/agHTUR7Ur5 #Blockchain #Bitcoin #b
tc #crypto #p2p
13629                 Bitcoin falls after report that one of the bigge
st crypto exchanges is facing regulatory trouble https://t.co/QlZtzVXh
NM (https://t.co/QlZtzVXhNM) via @markets
11201            RT @metalpaysme: A study showed that 30% of millennia
l's would rather invest $1,000 in Bitcoin than $1,000 in government bo
nds or stocks.…
16171       RT @CloudMiningX: Use the code: HF18BDAY30 at purchase to
 get a 30% discount for all contracts. The offer is limited. \n\n10 Gh
s = 0.84$\n1000…
13651                         RT @Bitcoin: Jack Dorsey expects bitcoin
to become the world's 'single currency' in about 10 years\n\nhttps://
t.co/V0wy32Fy38
13573        RT @bethereumteam: It's been 72 hours since the official
 launch of our Bounty campaign.\nIs there a stronger crypto community
 out there?\n#cr…
366            RT @SteveRichFXCorp: #Breaking #SteveRichFXCorp #News A
lert : https://t.co/UUGvxAIMYh (https://t.co/UUGvxAIMYh) is now live!
 Latest News and Updates on Cryptocurrenc…
17346          RT @WorldCoinIndex: #Bitcoin All Set to Replace the U.S
#Dollar as World’s Single Currency, Says Twitter co-founder Jack Dorse
y https://t.c… (https://t.c…)
5966           RT @ErikVoorhees: Many have questioned how Bitcoin work
s, and stay away from it due to this uncertainty. Meanwhile, not 1 in
 100 of them kn…
15509    Name: SunContract\nSymbol: SNC\n24 hour change: -4.86%\nPric
e: 0.18774\nRank: 277\nTotal Supply: 122707503.0\nVolume: 465… http
s://t.co/mZkXYp9Qw2 (https://t.co/mZkXYp9Qw2)
1157           RT @NickSzabo4: In particular, every full node watches
 every other full node (including watching the miners).  Robots in gre
en eyeshades ti…
9961          RT @egamexofficial: our community in the telegram  http
```

```
s://t.co/2S0DGb5fko (https://t.co/2S0DGb5fko) \n#egamex #egamexcoin #s
wap #bitcoin #litecoin #yobit @CoinExchan…
9296     RT @LitePalOfficial: #Bitcoin &amp;amp; #Litecoin\n\nFor far too
 long have we assigned a value to the worthless paper we're told has w
orth, now we en…
17259    Name: COSS\nSymbol: COSS\n24 hour change: -10.91%\nPrice: 0.2
43701\nRank: 336\nTotal Supply: 104000000.0\nVolume: 1056420… https://
t.co/yHPYFm6nrQ (https://t.co/yHPYFm6nrQ)
7870                          Great, easy to use platform,best bounty
 company \nC clear interface and a good reward.\n@WealthE_Coin http
s://t.co/6vK6tGmryR (https://t.co/6vK6tGmryR)
465           RT @SmartTaylorApp: Buy pressure: China and the USA are
forbidden to buy TAY until after the token sale. What do you think it
 will hapen th…
5586     RT @AivarasTop: Binance Just Open Registrations. Hurry! 🚀\n
\n➡ https://t.co/22NwNrwmyM (https://t.co/22NwNrwmyM)        \n\nSign
 up here to receive 20$ free coin 💰\n\n$BT…
3166      RT @RandolphMlny: #Bitcoin #cryptocurrency  #Airdrop\nNew A
irdrop #Tron 📢\n\nHuobi Exchange Airdrop 80 TRON (#TRX) to grow thei
r user base!\nEa…
7455              [TR] 23/03/2018 01:59:02 Bitcoin: ₺34416 Ethereum:
 ₺2137  #bitcoin #ethereum #altcoin #coin #blockchain #crypto… http
s://t.co/0lgiTl5ulH (https://t.co/0lgiTl5ulH)
7553                                    Current Bitcoin Rate in USD
 : 8,608.7650 Check other Currencies: https://t.co/KqQpwIzXrs (http
s://t.co/KqQpwIzXrs) #BitsRate #BTC #Bitcoins
7958          RT @RC_Mining: Forecast algorithm has detected that #Bt
cZ is a fantastic #investment📈🚀https://t.co/6Ood8V3zyw @BitcoinZTea
m  @RealTimeCrypt…
16147                                                            M
onthly Web Traffic for Major #bitcoin Exchanges Falls by Half https://
t.co/ffX48OxiKE (https://t.co/ffX48OxiKE)
10047         RT @maxkeiser: This also applies to Myron Scholes new
 ‘stable crypto coin’. || Centralized State Digital Tokens ‘Can’t Comp
ete With Bitcoin…
3731        RT @bethereumteam: We're revealing our surprise tomorrow!
\nAre you ready to celebrate with us? 😮\n#surprise #presents #crypto
#bitcoin #ethe…
5405                RT @alexposadzki: TMX enters bitcoin market with
 new cryptocurrency platform @willis_andrew /via @globeandmail http
s://t.co/KzDpZR4E80 (https://t.co/KzDpZR4E80)
15338       RT @TubiPlatform: "bitcoin is the future currency. It's w
hat we will all be using."\n\nTim Draper who previously predicted that
BTC would hit…
9392                                    #spentenoughmoneytoday Ho
w much money can bitcoin miners make? https://t.co/c0auOFILBj (http
s://t.co/c0auOFILBj) https://t.co/p8uMdZukVf (https://t.co/p8uMdZukVf)
8789                 i earn $2000 weekly click here to subscribe and
 find out how http://airdrop_er20\n#crypto (http://airdrop_er20\n#cryp
to) #altcoin #dogecoin… https://t.co/YvFTy3JqPa (https://t.co/YvFTy3Jq
Pa)
3414                 Here Are 7 Crypto Comparison Sites Chasing Coinm
arketcap’s Crown - https://t.co/sHByHF4Vz0 (https://t.co/sHByHF4Vz0) -
Generate Bitcoin. Take your free Bitcoin
14675         RT @FreeZone_one: Privacy-focused cryptocurrency zcash
 is gearing up for its first hard fork. #cryptocurrency #investment #i
nvesting #crypt…
5618        RT @AnselLindner: There's been more SW txs than #bcash tx
s.\n\nAfter the 2 year blocksize conflict, where we were told by FUDst
ers that 'we m…
6503                RT @Applancer_pro: Why Blockchain Will Survive Ev
```

```
en If Bitcoin Doesn't\nhttps://t.co/chtRI5PCyB #Blockchain #Bitcoin #b
tc #crypto #p2p
7888          #Bitcoin value in the next 2 days should be somewhere a
bout 💰 $8,735.52.\n📈 Gain: $112.26\n📈 Gain percentage: 1.30%… http
s://t.co/63hHxCjkki (https://t.co/63hHxCjkki)
15297                                                          Bitcoi
n News: With the New Casa Bitcoin Cold Storage Wallet Hack... https://
t.co/yU8CwD8gWh (https://t.co/yU8CwD8gWh)
13790                                RT @BTCTN: Crypto Collectibles Are
Worthless Without a Website https://t.co/XpfAkSSF7j (https://t.co/XpfA
kSSF7j) #Bitcoin https://t.co/3H0aT2v0lV (https://t.co/3H0aT2v0lV)
10942                                  The Future Of Bitcoin, From A Fi
nance Perspective https://t.co/ALEvkVWnFF (https://t.co/ALEvkVWnFF) #b
itcoin #ethereum #btc #crypto #blockchain
15582
Bitcoin ‘could become illegal’ https://t.co/K0XMIYxKlo (https://t.co/K
0XMIYxKlo) via @newscomauHQ
4249         RT @bethereumteam: Our transparent and easy-to-use bounty
solution is almost ready for a live launch!\n\nJoin our community on T
elegram: http…
13181    Name: Asch\nSymbol: XAS\n24 hour change: -27.15%\nPrice: 0.56
9067\nRank: 160\nTotal Supply: 114855331.0\nVolume: 3411400.… https://
t.co/5rNf3axNRi (https://t.co/5rNf3axNRi)
1910       @PinoyGameStore Bitcoin Gold (BTG) and become popular &amp;
right extra expensive, marely Your still donTt hold it? You… https://
t.co/LSLGQ9W0KE (https://t.co/LSLGQ9W0KE)
187                                CRYPTOLOANS ICO FIRST BLOCKCHAIN PLAT
FORM FOR SECURE LENDING\nTRADING AND EXCHANGE CRYPTOCURRENCY… https://
t.co/hQ2hQqW5jw (https://t.co/hQ2hQqW5jw)
4761                                                          RT @Gamblic
a: We need our own #crypto Groundhog Day for things like that https://
t.co/K84cRqjQdZ (https://t.co/K84cRqjQdZ)
744         RT @BuzToken: Airdrop 5000 people can get $10,00,000 worth
of buzz\nCrowdsale price is $0.12\n✔ Like and follow \n✔ Retweet. Tag
 5 your frie…
8651          RT @bethereumteam: After a successful launch of our Bou
nty campaign, we've managed to filter out the Bounty related questions
to: https://t… (https://t…)
7952          (The Australian dollar tanked) has been published on Fr
ee Forex Signals - Forex/ Bitcoin Signals Service - Manage A… https://
t.co/DkKmtlNg2C (https://t.co/DkKmtlNg2C)
4564        RT @bethereumteam: The Bether #token.\nSimple, #safe, #tra
nsparent and socially engaging!\nLearn more: https://t.co/C5UxE6TPGJ\n
#crypto (https://t.co/C5UxE6TPGJ\n#crypto) #block…
10771         RT @Blockchainlife: There is now $1.000.000 of US debt
 for every #Bitcoin that will ever be mined. At $21 trillion, the nati
onal debt is gr…
3249          RT @GymRewards: https://t.co/Bm9sIxiiwU (https://t.co/B
m9sIxiiwU)  Checkout our #bitcointalk #ANN https://t.co/J5xnJJr7Sa (ht
tps://t.co/J5xnJJr7Sa) … #Gymrewards #tokenssale #ethereum #bitcoi…
3871           RT @DomusCoins: Transfers are regulated automatically
 by Smart Contracts which are electronic rules, published on the block
chain, that are…
15874                                                          Bitcoi
n Loses $9k Support After Binance Confusion Shakes Confidence https://
t.co/crXsJKWe2K (https://t.co/crXsJKWe2K)
14767         RT @MAVRO_COIN: The ICO will be closed before you know
 it! Get your tokens now! #Mavro #cryptocurrency #blockchain #crypto #
bitcoin #bitcoi…
14662                                    #EOS Price is 0.00079152 (+0.
00000808) #BTC / 6.68708 (+0.11217) #USD. Market rank is 7. #eos #bitc
```

```
oin #blockchain
7222                                @Just_Hash_Me @YoustockProject @YouS
tockAura This article explains a little about what I'm doing\nhttps://
t.co/lpvM7Uc8ky
3236
How to Kill Bitcoin? https://t.co/7pqUFTyBUH (https://t.co/7pqUFTyBUH)
https://t.co/yMaUEHqfLL (https://t.co/yMaUEHqfLL)
Name: Tweet, dtype: object
```

Value of Emoticons as Sentiment

Emoticons are extremely important for sentiment analysis as they are clear exressors of emotions. The following code matches all emojis in our training dataset that match with a specific regular expression that is designed to generate all emojis.

This step is only for presentation purposes. These values will not actually be used later on.

In [16]:

```
# Checking which emoticons are used in data set
tweets_text = train_data.str.cat()

emos = set(re.findall(r" ([xX:;][-']?.) ", tweets_text))
emos_count = []
for emo in emos:
    emos_count.append((tweets_text.count(emo), emo))
print("Emoticons used in dataset:")
sorted(emos_count,reverse=True)
```

```
Emoticons used in dataset:
```

Out[16]:

```
[(14363, ': '),
 (115, ':…'),
 (39, 'XM'),
 (36, ':)'),
 (7, ':('),
 (2, ';)'),
 (2, ':D')]
```

Emoticons in Our Dataset by Sentiment

In [17]:

```
# Checking frequency of happy and sad emoji encounters

HAPPY_EMO = r" ([xX;:]-?[dD)]|:-?[\)]|[;:][pP]) "
SAD_EMO = r" (:'?[/|\(]) "

print("Emoticons specifying happy and sad expressions:\n")

print("Happy emoticons used:", set(re.findall(HAPPY_EMO, tweets_text)))
print("Sad emoticons used:", set(re.findall(SAD_EMO, tweets_text)))
```

```
Emoticons specifying happy and sad expressions:

Happy emoticons used: {';)', ':D', ':)'}
Sad emoticons used: {':('}
```

Most Used Words

The following function will check for most used words in our dataset so we have a clear visual of what we're working with. The nltk library will first download a set of words and then check the dataset for most used words before printing them for presentation and clearing purposes.

This step is only for presentation purposes. These values will not actually be used later on.

In [19]:

```
nltk.download('punkt')
def most_used_words(text):
    tokens = word_tokenize(text)
    frequency_dist = nltk.FreqDist(tokens)
    print("There is %d different words in training dataset" % len(set(tokens)))
    return sorted(frequency_dist,key=frequency_dist.__getitem__, reverse=True)

most_used_words(train_data.str.cat())[:100]
```

```
[nltk_data] Downloading package punkt to /home/zozu/nltk_data...
[nltk_data]   Package punkt is already up-to-date!

There is 30513 different words in training dataset
```

Out[19]:

```
[':',
 '#',
 'https',
 '@',
 'Bitcoin',
 '.',
 'the',
 'to',
 ',',
 '!',
 '$',
 'a',
 'is',
 'bitcoin',
 'and',
 'of',
 'in',
 'for',
 'you',
 '?',
 'on',
 'Airdrop',
 '(',
 ')',
 '’',
 '%',
 'with',
 'that',
 'I',
 '-',
 'cryptocurrency',
 'bethereumteam',
 'our',
 'blockchain',
 'crypto',
 'we',
 "'s",
 'Price',
 'The',
 'your',
 'will',
 'Supply',
 'Total',
```

```
'be',
'1',
'it',
'24',
'are',
'change',
'hour',
's',
';',
'out',
'BTC',
'Symbol',
'at',
'Rank',
"'re",
'We',
'...',
'have',
'by',
'&',
'Volume',
'this',
'Blockchain',
'what',
'about',
"'",
'--',
'*',
'Ethereum',
'Will',
'can',
'`',
'New',
'Twitter',
'has',
'from',
'📢',
'ICO',
'Satoshi',
'amp',
'make',
'“',
'all',
'”',
'article',
'Crypto',
'how',
'or',
'as',
'…',
'ethereum',
'not',
'now',
'A',
"'",
'ETH',
'money']
```

Feature Extraction, defining Vectorizer and Pipeline

This is the most important step. For natural language processing, we cannot feed raw text data to models. We have to convert them into a machine understandable format. Here is where our vectorizer will come in.

For the purpose of this project, we are using Bag of Words and TF-IDF feature extraction method. The way it works is the it creates a table with each tweet in our dataset as a row, and each column being each word encountered in the dataset atleast once. The tweets per row will then have numerical values with respect to columns to demonstrate the number of said words encountered in the tweet.

This simple diagram eases the concept:

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

We will use the TfidfVectorizer() from scikit-learn library and make a vectorizer of our own. In the same directory as this .ipynb file, there should be a utils.py file alongside. In said file, a lemmetizer function and an text preprocessing class have been defined which we are using in the code block below.

The reason for putting them separately was so that the pickled pipeline works with them later when they're deployed to a Heroku server.

In [20]:

```
vectorizer = TfidfVectorizer(tokenizer=utils.lemmatize_tokenize, ngram_range=(1,2))

pipeline = Pipeline([
    ('text_pre_processing', utils.TextPreProc(use_mention=True)),
    ('vectorizer', vectorizer),
])

training_data = pipeline.fit_transform(train_data)

joblib.dump(pipeline, 'pipeline.pkl')

print("Processed data ready to be passed to the model:")
print(training_data)
```

```
Processed data ready to be passed to the model:
  (0, 11436)    0.20075486694643815
  (0, 55414)    0.10031328191932584
  (0, 6107)     0.19068189100977978
  (0, 11857)    0.19068189100977978
  (0, 6553)     0.2452675750855544
  (0, 12050)    0.2452675750855544
  (0, 49429)    0.09740334370414654
  (0, 1171)     0.10024074751618044
  (0, 8589)     0.2452675750855544
  (0, 12675)    0.2452675750855544
  (0, 25698)    0.10013234145865559
  (0, 37269)    0.10024074751618044
  (0, 50499)    0.1815464986234954
  (0, 13866)    0.1815464986234954
  (0, 59557)    0.1542536565856081
  (0, 45080)    0.16530099330427037
  (0, 50342)    0.1815464986234954
  (0, 13861)    0.1815464986234954
  (0, 44817)    0.10020455957761988
  (0, 5257)     0.20075486694643815
  (0, 55408)    0.09408198338361679
  (0, 6106)     0.19068189100977978
  (0, 6552)     0.2452675750855544
  (0, 49422)    0.08723363948657727
  (0, 1110)     0.08896602587879256
  :     :
  (14267, 27298)        0.18985952615516027
  (14267, 65890)        0.18985952615516027
  (14267, 342)  0.18985952615516027
  (14267, 21835)        0.17868324112355058
  (14267, 20258)        0.18985952615516027
  (14267, 59856)        0.18985952615516027
  (14267, 55290)        0.18985952615516027
  (14267, 27300)        0.18985952615516027
  (14267, 65624)        0.18788309633948375
  (14267, 30059)        0.18985952615516027
  (14267, 65889)        0.18985952615516027
  (14267, 20255)        0.18841140320473038
  (14267, 27297)        0.37898394501169186
  (14267, 55284)        0.18268505456046782
  (14267, 1254) 0.1603501729747892
  (14267, 16025)        0.15536279755951415
  (14267, 59850)        0.18617628963833194
  (14267, 30141)        0.1670138377116816
```

```
  (14267, 30095)        0.14957756213493043
  (14267, 14249)        0.11706683588032979
  (14267, 64316)        0.11641852573705856
  (14267, 63407)        0.16557828529244598
  (14267, 63373)        0.12002016518689004
  (14267, 0)    0.09289969551699381
  (14267, 21834)        0.056568206781440866
```

Model Training

The above data is the data that was transformed by passing through our defined pipeline according to our lemmetizer and text preprocessing rules class. This will work fine when passed into a model.

We will now retain the performance and accuracy of 9 machine learning algorithms on our training data.

In [28]:

```
perceptron = Perceptron()
bnb = BernoulliNB()
mnb = MultinomialNB()
cnb = ComplementNB()
tree = DecisionTreeClassifier()
lsvc = LinearSVC()
sgdc = SGDClassifier()
randFor = RandomForestClassifier()
lr = LogisticRegression(max_iter=1000)

models = {
    "Random Forest Classifier": randFor,
    "Perceptron": perceptron,
    "Bernoulli Naive Bayes": bnb,
    "Multinomial Naive Bayes": mnb,
    "Complement Naive Bayes": cnb,
    "Decision Tree Classifier": tree,
    "Linear Support Vector Classification": lsvc,
    "Stochastic Gradient Descent": sgdc,
    "Logistic Regression": lr,
}


for model in models.keys():
    scores = cross_val_score(models[model], training_data, train_sentiment)
    print("\n===", model, "===")
    print("scores = ", scores)
    print("mean = ", scores.mean())
    print("variance = ", scores.var())
    models[model].fit(training_data, train_sentiment)
    acc_score = accuracy_score(models[model].predict(training_data), train_sentimen
    print("score on the learning data (accuracy) = ", acc_score)
    print("")
```

```
=== Random Forest Classifier ===
scores =  [0.90749825 0.9159075  0.91065172 0.91342447 0.90606379]
mean =  0.9107091442367186
variance =  1.3257659739066788e-05
score on the learning data (accuracy) =  1.0


=== Perceptron ===
scores =  [0.92676945 0.93622985 0.93587947 0.93901157 0.93340343]
mean =  0.9342587536791698
variance =  1.7184494186189898e-05
score on the learning data (accuracy) =  1.0


=== Bernoulli Naive Bayes ===
scores =  [0.88367204 0.90049054 0.89313245 0.90536278 0.88818787]
mean =  0.8941691345934437
variance =  6.245940032019872e-05
score on the learning data (accuracy) =  0.9718951499859826


=== Multinomial Naive Bayes ===
scores =  [0.88156973 0.88822705 0.88367204 0.89730109 0.88958991]
```

```
mean =  0.8880719615271155
variance =  2.9828665773171636e-05
score on the learning data (accuracy) =  0.9670591533501542


=== Complement Naive Bayes ===
scores =  [0.88717589 0.89698669 0.8917309  0.90851735 0.89309499]
mean =  0.8955011641442109
variance =  5.218850994221847e-05
score on the learning data (accuracy) =  0.9748388001121391


=== Decision Tree Classifier ===
scores =  [0.88997898 0.89978977 0.89138052 0.88608482 0.87767263]
mean =  0.888981342498129
variance =  5.197006157597027e-05
score on the learning data (accuracy) =  1.0


=== Linear Support Vector Classification ===
scores =  [0.9278206  0.93482831 0.9351787  0.94356817 0.92534175]
mean =  0.9333475059509029
variance =  4.0929391057163947e-05
score on the learning data (accuracy) =  0.9999299130922343


=== Stochastic Gradient Descent ===
scores =  [0.92186405 0.9302733  0.92992292 0.9404136  0.92393971]
mean =  0.9292827157191523
variance =  4.177439385731084e-05
score on the learning data (accuracy) =  0.9954443509952341


=== Logistic Regression ===
scores =  [0.90504555 0.91520673 0.90889979 0.92393971 0.91132142]
mean =  0.9128826391821049
variance =  4.1476146110314495e-05
score on the learning data (accuracy) =  0.9889262685730306
```

Testing Accuracy on Test Data

We will now test the accuracy of each model on test data and check which one gives the highest result. As is visible from below, Linear Support Vector Classification gives the highest rating with respect to accuracy for testing data.

In [29]:

```
# We now test each model on trainset
for model in models.keys():
    test_model = models[model]
    test_model.fit(training_data, train_sentiment)

    testing_data = pipeline.transform(test_data)
    print("Accuracy on test data for " + model + ":", test_model.score(testing_data
```

```
Accuracy on test data for Random Forest Classifier: 0.913372582001682
Accuracy on test data for Perceptron: 0.9425287356321839
Accuracy on test data for Bernoulli Naive Bayes: 0.8982338099243061
Accuracy on test data for Multinomial Naive Bayes: 0.8901037286234932
Accuracy on test data for Complement Naive Bayes: 0.895991028875806
Accuracy on test data for Decision Tree Classifier: 0.8873002523128679
Accuracy on test data for Linear Support Vector Classification: 0.9428
090832632464
Accuracy on test data for Stochastic Gradient Descent: 0.9408466498458
088
Accuracy on test data for Logistic Regression: 0.927950658816933
```

Comparing Test Sentiments to Predictions

The time has come to compare test dataset sentiments with our predictions from the model. In the display as follows, both the original sentiment and predicted sentiments are compared side by side. As you can see, the result is quite impressive.

In [30]:

```
# We choose Linear Support Vector Classification due to highest accuracy
test_model = lsvc
test_learning = pipeline.transform(test_data)

tempDF = pd.DataFrame(test_sentiment)
tempDF["Predicted Sentiment"] = test_model.predict(test_learning)

tempDF.head(60)
```

Out[30]:

	Sentiment	Predicted Sentiment
12501	neutral	neutral
607	negative	negative
13116	neutral	neutral
10555	neutral	neutral
15317	negative	negative
13995	negative	negative
2243	neutral	neutral
13629	negative	negative
11201	neutral	neutral
16171	negative	negative
13651	negative	negative
13573	neutral	neutral
366	positive	positive
17346	negative	negative
5966	positive	positive
15509	negative	negative
1157	positive	positive
9961	neutral	neutral
9296	negative	negative
17259	negative	negative
7870	positive	positive
465	neutral	neutral
5586	positive	positive
3166	negative	negative
7455	neutral	neutral
7553	negative	negative
7958	positive	positive
16147	negative	negative
10047	positive	positive

	Sentiment	Predicted Sentiment
3731	positive	positive
5405	positive	positive
15338	negative	negative
9392	positive	positive
8789	neutral	neutral
3414	positive	positive
14675	negative	negative
5618	positive	positive
6503	neutral	neutral
7888	neutral	neutral
15297	negative	positive
13790	negative	negative
10942	neutral	neutral
15582	negative	negative
4249	positive	positive
13181	negative	negative
1910	positive	positive
187	positive	positive
4761	positive	positive
744	positive	positive
8651	positive	positive
7952	positive	positive
4564	positive	positive
10771	neutral	neutral
3249	neutral	neutral
3871	positive	positive
15874	negative	negative
14767	negative	negative
14662	negative	negative
7222	negative	negative
3236	neutral	neutral

Finalize the Model

In [31]:

```
model = test_model
```

Test Custom Input

We will now test our own input to test the model. As you can see, both the predictions are on point.

In [32]:

```
tweet = pd.Series([input(),])
tweet = pipeline.transform(tweet)

sentiment_predicted = model.predict(tweet)[0]

print("This tweet is", sentiment_predicted)
```

```
I hate bitcoin
This tweet is negative
```

In [33]:

```
tweet = pd.Series([input(),])
tweet = pipeline.transform(tweet)

sentiment_predicted = model.predict(tweet)[0]

print("This tweet is", sentiment_predicted)
```

```
I love bitcoin
This tweet is positive
```

Save Model

We now save the model as a pickle file so we can use it to deploy on a server.

In [34]:

```
pickle.dump(model, open("model.pkl", 'wb'))
print("Model Saved")
```

```
Model Saved
```