ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 7, Issue 2, February 2020

Sentimental Analysis for Cryptocurrency using Deep Learning Techniques

P. Sugavaneshwari
Department of Computer Science and Engineering
Government College of Engineering
Dharmapuri, India
suga.eshwari@gmail.com

Dr. K. G. Saranya
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
kgs@psgtech.edu.in

***Abstract*— Cryptocurrency is decentralized digital money that is based on blockchain technology and secured by cryptography. More organizations and companies recognizing cryptocurrency has started to gain a larger following and is becoming more stable by the day. Cryptocurrencies and the general opinion about them can be obtained by performing sentiment analysis on the data extracted from tweets. Automatic analysis of customer opinion in the form of surveys, questionnaires, and social media forums allows businesses to hear their customers effectively. This helps them to understand their customers better through sentiment analysis and opinion mining. This paper is to extract public opinions about different cryptocurrencies, such as Bitcoin and Ethereum, from Twitter data using the Twitter API. The collected data is then pre-processed using a variety of pre- processing techniques, such as tokenization, stemming, lemmatization, etc. The processed data is subjected to Sentiment Analysis utilizing a variety of algorithms, including Naive Bayes, Support Vector Machines (SVM), and Recurrent Neural Network (RNN) variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Each technique is compared based on the results of parameters like accuracy, precision, recall, and f-score. Additionally, the benefits, drawbacks, and restrictions of each model are examined, and the best model for sentiment analysis is determined**

Keywords—Cryptocurreny, Naïve Bayes, SVM, RNN, LSTM, GRU

I. INTRODUCTION

The barter system, in which products and services are exchanged between two or more individuals, was in existence since the time of the cavemen. An example would be trading seven apples for seven oranges. Due to numerous obvious shortcomings, stated below, the barter system lost favor:

The requirements of the engaging parties need to coincide with each other, which is highly unlikely and thus the chances for both parties to come to an agreement are extremely low. This makes obtaining the resources necessary for survival, for any individual unnecessarily difficult as it has some aspect of luck to it.

There was no standard measure of value to make sure that both parties engaging in the transaction received articles of equal value. Both parties have to come to a conclusion about how much something is worth, which will be difficult as one of the parties will want it to be worth more while the other would want it to be not worth as much. Many articles can also not be split and thus making the negotiation process difficult.

The transaction can occur only when the goods themselves are physically present and either party can view and inspect the goods of the other in order to ensure that the value has been negotiated appropriately. The logistics and transportation of the goods are difficult, in case it is livestock, perishable goods, or other large goods.

After society realized that the barter system was not a very effective transactional system, a currency exchange has gone through multiple iterations. Paper notes and metal coins are used popularly in today's day and age. In our digital age, economic transactions regularly take place electronically, without the exchange of any physical currency. Digital cash in the form of bits and bytes will most likely continue to be the currency of the future.

Paper money, coins, credit cards, and digital wallets like Apple Pay, Amazon Pay, Paytm, PayPal, and so on are all examples of modern money. All of this currency is controlled by regulatory bodies, consisting of banks and governments, that limit how paper currency and credit cards work.

In future probably, Cryptocurrency can rule the world. Let's take two people, A and B, where one would like to pay the other a certain amount of money, by sending this money to the other person's bank account, online. There are several methods in which this transaction could be interrupted like the bank could have some technical issues and their services might be down, the other person's account could have been hacked due to a DDOS attack etc., the limit set for the receiver's account may have exceeded. In all of these scenarios, the central point of failure is the bank. This is the major reason why the future of currency is in cryptocurrencies. However, the following is how a comparable transaction between two users of the bitcoin app would happen. A notification asks the user if they are certain they are ready to send bitcoins. If so, the following happens: The system verifies the user's identification, determines whether they have the necessary balance to complete the transaction, and other things. Following that, the payment is transmitted, and the funds are deposited into the recipient's account. This everything occurs in a couple of minutes. Cryptocurrency thus removes all the major issues with the current banking system, there are no limits set on the transfer of currencies, and there is no central point of failure. As of 2018, there are more than 1,600 cryptocurrencies available and some of the popular cryptocurrencies are Bitcoin, Litecoin, and Ethereum.

Bitcoin - The original cryptocurrency and currently the most traded, Bitcoin was established an individual or group of people whose precise identity remainsunknown.

Ethereum - It is the second most popular cryptocurrency after Bitcoin. Developed in 2015, Ethereum is a blockchain platform with its own cryptocurrency, called Ether (ETH) or Ethereum.

Ripple - Ripple was founded in 2012. It is a distributed ledger system. Ripple can be used to track various transactions, not just cryptocurrency. The company that developed it, worked with sev

eral banks and other financial institutions.

Non-Bitcoin cryptocurrencies are collectively known as “altcoins” to distinguish them from the original. Cryptocurrencies are strings of data that are coded to represent a unit of currency. The buying and selling of cryptocurrencies is monitored by a network called blockchains.

Cryptocurrencies act as both the currency and the accounting system too, by the use of encryption. Cryptocurrency is a virtual currency that is supposed to be used as a medium of exchange. It is similar to a physical currency except for the fact that it has a physical embodiment, and uses encryption to work. As cryptocurrencies work in a decentralized manner, (i.e.) there are no banks or central authorities, there are certain restrictions set to be able to add new units of the currency to the system. With the drastic rise in popularity of social media, like Twitter, Instagram, and Reddit, more than half of the world’s population, around 59%, are active users of social media. This makes it possible that the users of social media can affect the market, both the traditional stock market and virtual cryptocurrencies, by publishing their opinions on the servers. Thus, opinion mining can be extremely useful for anyone trying to find the overall opinion about a certain topic.

Opinion mining is also known as Sentiment Analysis. Sentiment Analysis is the Natural Language Processing (NLP) technique that is used to determine whether a certain piece of digital text’s emotional tone is positive, negative, or neutral. Many companies use the results from sentiment analysis to improve the service and the products that they can provide, and improve their reputations.

Sentiment analysis provides objective insights, as companies can avoid any personal bias when compared to marketers. It allows the company to manufacture better products and services as they can find genuine feedback about their services and products. Sentiment analysis on social media can allow companies to obtain feedback from a lot of people, (i.e.) at a large scale. It allows companies to obtain results in a real-time. The three main approaches that are used in sentiment analysis are as follows:

Rule-based approach - The rule-based methodology uses preset lexicons to identify, categorize, and score certain terms. Lexicons are collections of words used to convey the intention, feeling, and tone of the author. Marketers rate the emotional impact of various terms by assigning sentiment ratings to positive and negative lexicons. The software searches for terms from the lexicon to identify whether a statement is good, negative, or neutral before calculating the sentiment score. To establish the overall emotional bearing, the final score is contrasted with the sentiment bounds. It is simple to put up a rule-based sentiment analysis system, but it is challenging to scale. For instance, you'll need to continue adding new terms to the lexicons as you find new ways to express purpose in text input. Additionally, this method might not be reliable when analyzing statements that have been impacted by several civilizations.

Machine learning - This method teaches computer software to recognize emotional sentiment from text using machine learning (ML) techniques and sentiment classification algorithms, including neural networks and deep learning. To create a sentiment analysis model that can accurately predict the sentiment in new data, it must first be built and frequently trained on known data. Because it properly analyses a large variety of text information, ML sentiment analysis is helpful. The emotional tone of the communications may be correctly predicted by ML sentiment analysis provided that the machine is trained with enough instances. A trained ML model, however, is unique to a single industry. As a result, sentiment analysis software that has been educated using marketing data cannot be utilized to monitor social media without being retrained.

Hybrid approach - Hybrid sentiment analysis works by combining both ML and rule-based systems. It uses features from both methods to optimize speed and accuracy when deriving contextual intent in text. However, it takes time and technical effort to bring the two different systems together.

Sentiment analysis is a sort of textual contextual mining in which subjective information is discovered and extracted from source material to assist organizations in understanding the social sentiment of their brand, product, or service while monitoring online conversations. However, most social media analysis is limited to simple sentiment analysis and count-based metrics.

Sentiment analysis, the most common text categorization tool, analyses an incoming message and determines whether the underlying sentiment is positive, negative, or neutral. Sentiment analysis is considered a classification task since it categorizes the orientation of a text as positive or negative. Machine learning, in addition to lexicon-based and linguistic approaches, is a widely used approach to sentiment categorization. According to some, these approaches do not perform as well in sentiment classification as they do in subject categorization because the nature of an opinionated text necessitates more comprehension of the text, whereas the presence of specific keywords may be the key to an appropriate classification. In sentiment classification, machine learning classifiers such as naive Bayes, maximum entropy, and support vector machine (SVM) are used to achieve accuracies ranging from 75% to 83%, compared to 90% or better accuracy in topic-based categorization. Sentiment analysis can be performed to obtain different types of results as shown below:

Fine-grained scoring - The term "fine-grained sentiment analysis" refers to the division of text intent into several emotional gradations. According to this technique, user sentiment is often rated on a scale of 0 to 100, with each equal segment denoting very positive, positive, neutral, negative, and very negative opinions. A 5-star rating system is used by e-commerce shops as a fine-grained grading tool to evaluate the purchasing experience.

Aspect-based - Aspect-based analysis concentrates on specific features of a good or service. For instance, laptop makers ask users about their experiences with the touchpad, keyboard, and graphics. They link client intent with phrases pertaining to hardware using sentiment analysis methods.

Intent-based - When performing market research, intent-based analysis helps to identify customer sentiment. Marketers utilize opinion mining to comprehend where a particular consumer group stands in the buying cycle. When they hear things like discounts, bargains, and reviews in watched conversations, they conduct targeted advertising on clients who are interested in purchasing.

Emotional detection - Analyzing a person's psychological state at the time they were producing the text is a component of emotional detection. Sentiment analysis is a more intricate field than emotional detection since it does more than just categorize data. With this method, sentiment analysis models try to decipher different emotions—like happiness, rage, sadness, and regret—through the words that are spoken

II. LITERATURE SURVEY

S. No	Title	Authors	Description	Future Scope
1	Social Media Sentiment Analysis for Cryptocurrency Market Prediction (Submitted on 19 Apr 2022 in arXiv)	Ali Raheman, Anton Kolonin, Igors Fridkins, Ikram Ansari, Mukul Vishwas	The feasibility of numerous models for social media sentiment analysis is applied to the forecast of the financial markets, using the bitcoin sector as a paradigm.	Exploring the predictive power of the connection to improve the reliability of the price prediction and business applications for decentralized finance relying on such predictions.
2	Mining netizen's opinions on cryptocurrency: sentiment analysis of Twitter data (Submitted on 5 November 2021 in Emerald Publishing Limited)	M. Kabir Hassan, Fahmi Ali Hudaefi, Rezzy Eko Caraka	To investigate netizens' perspectives on cryptocurrency using emotion theory and lexicon sentiments analysis via machine learning.	Gathering more objective information from different media channels to better comprehend public opinion.
3	Twitter sentiment data analysis of user behaviour on crypto currency (Submitted in 2021 at IGI Global)	Hashitha Ranasinghe, Malka N Halgamuge	This study examines consumer preferences for Bitcoin and Ethereum, the two most widely used cryptocurrencies worldwide, using Twitter sentiment analysis.	It would have been a more viable solution to use tweets collected over a longer period of time, as well as multiple cryptocurrency options.
4	Cryptocurrency – Sentiment Analysis in social media (Submitted in 2019 at Sciendo)	Tudor Mircea Dulau, Mircea Dulau	This paper suggests investigating, identifying, and developing a method for separating the sentiment linked to the cryptocurrency phenomenon from the posts on	The findings could be incorporated into a as an additional parameter for the machine learning system identifying potential growths in the

			some well-known social media sites.	main Currency exchange and investment.

III. DATASET

The dataset required for this project has been procured from Twitter using the Twitter API to extract tweets about various cryptocurrencies like,

a) Bitcoin,

b) Ethereum,

c) Cardano,

d) Dogecoin etc

The tweets are first extracted from Twitter data using the Twitter API. The collected data is then pre-processed using a variety of pre-processing techniques, such as tokenization, stemming, lemmatization, etc. This gives us clean, processed data on which sentiment analysis can be performed.

The information obtained by utilizing the Twitter API has the following features after being pre-processed,

a) uncleaned_tweet – The raw tweet that is obtained from Twitter API.

b) name – The name of the user who has published that particular tweet.

c) created_at – The date and time of publication of the tweet.

d) favourites count – The count of users who have favourited that particular tweet.

e) screen_name – The on-screen name chosen to be displayed to the public by the user.

f) Source – This feature shows us how the tweet was published, (i.e.) using an Android device, or the Web App, etc.

g) retweet_count – The count of users who have retweeted that particular tweet.

h) clean_tweet – The data obtained after performing pre-processing.

i) polarity – The value that represents the polarity of that particular tweet.

j) subjectivity – The value that represents how subjective that particular tweet is.

k) Sentiment – This represents the overall sentiment of a particular tweet.

The data is stored as a comma-separated value file after all the pre-processing is done.

In order to include more inclusive natural language in our dataset data from google news is included as well. The data from google news is collected using google news API. The data after pre-processing has the following features:

Fig. 1. Dataset from Twitter

Title - The news title is obtained from google news API.

Desc - The clear description of the news is displayed in the desc feature

polarity – The value that represents the polarity of that particular text.

Sentiment – This represents the overall sentiment of a particular text.

Fig.2. Dataset from Google News

The data is stored as a comma-separated value file after all the pre-processing is done. Both the datasets are merged with the following common features:

uncleaned_text - The raw text before pre-processing is displayed here.

cleaned_text - The cleaned text after pre-processing is displayed here.

polarity - The value that represents the polarity of that particular text.

Sentiment - This represents the overall sentiment of a particular text.

	uncleaned tweet	name	created_at	favorite_cour	screen_name	source	retweet_count	clean tweet	polarity	Subjectivity	sentiment
0	RT @MetaBeatOffl: QUIZ TIME on today's # HINT FOR TONIGHT'S #AMA with @V(	ImDahDude	2022-09-10 13:33:05	80	DahDude22	Twitter Web App	5	quiz time todays morningbeat metabeat	0	0	neutral
1	RT @GGslayerstudios: Massive #Giveav Reward Pool - Tokens Worth $5,000 To Win Follow @Digifinex & @GGslayerOf	SanjayVaghri	2022-09-10 13:33:04	3547	sanjay_vaghri	Twitter Web App	992	massive giveaway reward pool tokens w	0.55	0.25	positive
2	Current #Cardano price Alert: #ADA $0.52 USD #Bitcoin 2.443773242280849e-05 BTC #Ethereum 0.0003015191928563145 E... h	Cardano Tracker	2022-09-10 13:33:03	2	CardanoTracker	Twitter Web App	0	current cardano price ale052 usd bitcoir	0	0.4	neutral
3	RT @evai_io: What's in it for me? Rewarding Evai community members that t	godfrey™	2022-09-10 13:33:01	184359	godfrey_G_	AiSpaceBot	3	whats rewarding evai community memb	0	0	neutral
4	RT @CLOCKMENNFT: Join us in revolution 15x Knightlist...	rul	2022-09-10 13:33:01	6600	bromenangbro	Twitter for Android	801	join revolutionizing web3 partnerships di	0.6	0.9	positive
5	RT @ZakaWaqar: #Bitcoin changed my life It can change yours too. #WaqarZaka #Crypto #Cryptocurrency #NF	Haziq Khan	2022-09-10 13:33:00	152	HaziqKh6290996	Twitter for Android	61	bitcoin changed life change waqarzaka	0	0	neutral
6	RT @JoaoDMendes: I just published a New "Luna CLassic: Under Coordinated Whale A #luncburn #LUNAC #LUNACLASSIC #lunct	Shibnobiway	2022-09-10 13:33:00	8875	wayofshibnobi	Twitter for Android	28	published article luna classic coordinate	0.1666666667	0.1666666667	positive
7	RT @evai_io: Ratings Review 1 Minute Rou Changes in Evai Rating for $POLY forecast Watch...	godfrey™	2022-09-10 13:32:59	184359	godfrey_G_	AiSpaceBot	6	ratings review 1 minute roundup ep 34 c	0	0	neutral
	RT @adscoinofficial: Push the switch and r										

Fig. 3. Merged dataset

IV. LEARNING MODELS

A. NAIVE BAYES MODEL

The Naive Bayes classification technique is a basic and powerful classification problem in machine learning. The employment of Bayes' theorem with a strong independence assumption between the characteristics is the cornerstone of naive Bayes classification. When utilized for textual data analysis, such as Natural Language Processing, the Naive Bayes classification yields good results.

Simple Bayes models and independent Bayes models are other names for naive Bayes models. All of these terms refer to the classifier's decision rule, which is based on Bayes' theorem. The Bayes theorem is used in practice via the Naive Bayes classifier. The power of Bayes' theorem is used with this classifier.

$P(A|B) = P(B|A) . P(A)/P(B)$

$P(A|B)$ is Posterior probability: Probability of A on the observed event B.

$P(B|A)$ is Likelihood probability: The probability of the evidence given that the probability of a hypothesis is true.

$P(A)$ is Prior Probability: The probability of the hypothesis before observing the evidence.

$P(B)$ is Marginal Probability: Probability of Evidence.

The Bayes theorem is used by the Naive Bayes Classifier to forecast membership probabilities for each class, such as the likelihood that a specific record or data point belongs to that class. The most likely class is the one with the greatest chance. This is also known as the Maximum A Posteriori (MAP).

B. SVM

Support Vector Machine, or SVM, is a popular Supervised Learning approach for classification and regression problems. However, it is largely used in Machine Learning to solve classification problems. The goal of the SVM method is to discover the best line or decision boundary for categorising n-dimensional space so that simply add new data points in the appropriate category in the future. The best suitable boundary is a hyperplane. SVM chooses the extreme points/vectors that will help create the hyperplane. The technology is known as the Support Vector Machine, and these extreme cases are referred to assupport vectors.

Support vector machines can handle classification and regression both on linear and non-linear data. SVM is used to solve both linear and non-linear problems and also works well for different practical problems. SVM is more productive in high-dimensional spaces. It is effective in instances where the totalnumber of dimensions is larger than the total number of specimens. Support Vectors are the data points or vectors that are closest to the hyperplane and have an effect on its location. Because these vectors support the hyperplane, they are referred to as Support vectors.

The SVM kernel is a function that turns a non-separable problem into a separable problem by converting a low-dimensional input space to a higher- dimensional space. It works quite well in nonlinear separation scenarios. Simply said, the kernel does some very complex data transformations before deciding how to partition the data based on the labels or outputs supplied.

C. Decision Trees

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

The given dataset's features are used to execute the test or make the decisions.

It is a graphical representation for obtaining all feasible answers to a decision or problem based on predetermined conditions.

It is called a decision tree because similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

Decision Tree Terminologies:

Root Node: The root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

Branch/Sub Tree: A tree formed by splitting the tree.

Pruning: Pruning is the process of removing unwanted branches from the tree.

Parent/Child node: The root node of the tree is called the parent node, and other nodes are called the child nodes.

D. Long Short-Term Memory

Long Short-Term Memory is a type of recurrent neural network (RNN). In RNN, the output from the previous step is used as input in the next phase. Hochreiter and Schmidhuber created the LSTM. It addressed the problem of RNN long-term dependency, in which the RNN cannot predict the word stored in long- term memory but can make more accurate predictions based on the current input. RNN does not function efficiently as the gap length rises. By default, LSTM may save information for a long time. It is used to analyze, forecast, and categorize time-series data.

Structure of LSTM

The LSTM has a chain structure that includes four neural networks and several memory units known as cells. The cells store information, while the gates perform memory manipulation. There are three gates, the Forget gate, the Input gate, and the Output Gate. Each of the gates and their purposes is given below. The following figure shows the architecture of the LSTM Model and how the different gates in it interact with one another.

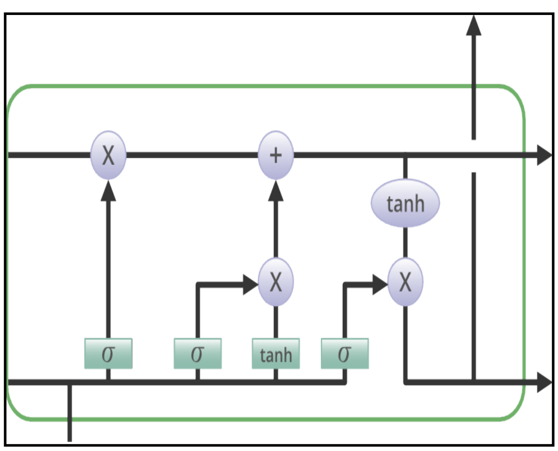

Fig. 4. Long Short-Term Memory Model

Forget Gate: The forget gate removes information from the cell that is no longer useful. Before bias is applied, two inputs, x t (at-the-time input) and h t-1 (previous cell output), are sent into the gate and multiplied using weight matrices. The outcome is sent into an activation function, which generates a binary output. If the output for a particular cell state is 0, the information is lost; if it is 1, the information is preserved for future use.

Input Gate: The input gate is in charge of contributing critical information to the cell state. First, the sigmoid function is used to regulate the information, and the values to be remembered are filtered using the h t-1 and x t inputs, similar to the forget gate. The tanh function is then used to build a vector with values ranging from -1 to +1 that includes all of the potential h t-1 and x t values. Finally, the vector and controlled values are multiplied to obtain useful data.

Output gate: The output gate is responsible for extracting useful information from the current cell state and displaying it as output. To begin, the tanh function is used to the cell to generate a vector. The information is then filtered by the values to be remembered via the h t-1 and x t inputs and regulated using the sigmoid function. Finally, the vector and controlled values are multiplied and output and input to the next cell are delivered.

E. Gated Recurrent Unit (GRU)

GRU, or Gated recurrent unit, is an advanced RNN, or recurrent neural network. GRUs are similar to Long Short-Term Memory (LSTM). GRU, like LSTM, uses gates to regulate the flow of information. They are a more recent technology than LSTM. This is why they outperform LSTM and have a simpler architecture. Another noteworthy aspect of GRU is that, unlike LSTM, it does not have a distinct cell state (Ct). It has just one state: hidden (Ht). Because of their simplified architecture, GRUs are easier to train.

The different gates of a GRU are described below:-

Update Gate: It defines how much previous information must be transmitted into the future. In an LSTM recurrent unit, it is comparable to the Output Gate.

Reset Gate: It decides how much prior information to forget. In an LSTM recurrent unit, it is comparable to the combination of the Input Gate and the Forget Gate.

Current Memory Gate: The current memory gate is frequently forgotten in discussions about Gated Recurrent Unit Networks. It is a component of the Reset Gate in the same way that the Input Modulation Gate is a component of the Input Gate, and it is used to bring nonlinearity into the input as well as to make the input Zero- mean. Another rationale for including it in the Reset gate is to lessen the impact of prior knowledge on current information being sent in the future.

F. Transformers

A Transformer is a deep learning model that uses the attention mechanism to evaluate the significance of each component of the incoming input separately. It is extensively used in natural language processing (NLP) and computer vision (CV). A transformer is composed of an encoder and a decoder, both of which are composed of modules that may communicate with each other several times. Because this cannot utilized directly, first wrap the inputs and outputs in n-dimension space. So, whatever inputs are, it must be encrypted. A minor but critical feature of this paradigm is the location and coding of separate words. It is needed to provide a relative placement to each word or component of a sequence since a sequence is dependent on the order of the parts and don't have a recurrent neural network that can remember how the sequence is fed into the model. These places are added to the embedded. representation of each word. Transformers is an architecture for converting one sequence into an antidote while aiding the other two parts, encoders and decoders, but it differs from the previously described sequence in that it does not act in the same manner as GRUs do. As a consequence, no recurring neural networks are used. Until recently, recurrent neural networks were one of the most effective methods for capturing the tiniest dependency on a sequence.

Transformers do not always handle data in the same sequence. In contrast, the attention mechanism provides context for any point in the input stream. Rather, it establishes the context that provides meaning to each word in the statement. This feature enables more parallelization than RNNs.

Transformers are now the most advanced type of model for dealing with sequences. The most prominent use of these models is in text processing jobs, particularly machine translation. In fact, transformers and their conceptual offspring have gone to the top of practically every NLP benchmark scorecard, from question answering to grammatical correction. Transformer designs are being developed at a breakneck pace, for better or worse, compared with Convolutional Neural Networks (CNN) during the 2012 ImageNet competition.

The transformer can be divided into three parts:

An Encoder that turns an input sequence into state representation vectors.

An Attention approach that allows our Transformer model to focus on the important sequential input stream components. This is used throughout the encoder and decoder to help them contextualize the incoming data.

A Decoder that decodes the state representation vector to generate the target output sequence

V. SYSTEM DESIGN

The Figure 1.1 represents the flow of data and the architecture of our application. Initially, the data sources of Twitter and Google News are used, from which data is extract ed using the Twitter API and NewsAPI. Then data is processed to obtain clean text which is then stored and can be used in different models to perform sentiment analysis. Here multiple models are used to find their performance metrics, and then find the best model among all by comparing these metrics using a pre-trained Transforms model.

Initially, data regarding various cryptocurrencies like bitcoin, Ethereum, and dogecoin from a source like Twitter and google news are extracted. In order to accomplish this, Twitter API and News API are used for our respective tasks. The data collected is further pre-processed and their sentiment is found using a lexicon model. Then the data from both sources are merged and represented as a single data frame which is then stored as a comma-separated value file.

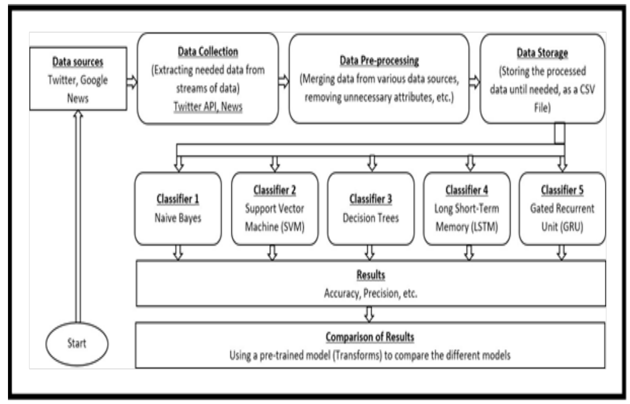

Fig 5. System Flow Overview

The data is then passed through various machine learning and deep learning algorithms. The respective metrics and generate the results are calculated. The generated results are then further compared with the results obtained from a pre-trained transformers model (here hugging face Roberta model and a distilbert model is used).

VI. INFERENCE OF LEARNING MODEL

After running our machine learning algorithms (in our case Support vector machine, Naive Bayes classifier, and Decision Tree Classifier) the efficiency of each of our models using various classification metrics like precision score, recall score, f1 score, and accuracy score are measured.

ML Algorithm	Precision	Recall	F-Score	Accuracy
Support Vector Machine	0.9130	0.9106	0.9036	0.9106
Naive Bayes Classifier	0.8363	0.8419	0.8290	0.8419
Decision Tree Classifier	0.8363	0.8419	0.8290	0.9412

Table. 1. Performance metrics for ML models

Precision - Precision is defined as the proportion of correctly predicted positive observations to all predicted positive observations. The question that this metric answer is, how many of the classes labelled as such are actually correct? High precision is associated with a low false positive rate. Precision is a measure of how frequently correct positive. The formula for it is:

$$\frac{\text{True Positives}}{\text{True Positives + False Positives}} = \frac{\text{N. of Correctly Predicted Positive Instances}}{\text{N. of Total Positive Predictions you Made}}$$

Fig. 6. Precision Formula

Recall - Recall is a measure of how many positive cases the classifier predicted correctly out of all the positive cases in the data. It is also known as Sensitivity at times.

$$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}} = \frac{\text{N. of Correct Predictions}}{\text{N. of all Predictions}} = \frac{\text{N. of Correct Predictions}}{\text{Size of Dataset}}$$

Fig. 7. Recall Score Formula

F1-Score - The F1-Score statistic combines accuracy and recall. It is also known as the harmonic mean of the two. Harmonic mean is just another method for calculating a "mean" of numbers, and it is sometimes promoted as being more suited for ratios (such as accuracy and memory) than the usual arithmetic mean. The formula for it is:

$$2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

Fig. 8. F1-Score Formula

Accuracy Score - Accuracy Score, which describes the number of right predictions overall predictions, is frequently used as the basis metric for model evaluation.

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}} = \frac{\text{N. of Correctly Predicted Positive Instances}}{\text{N. of Total Positive Instances in the Dataset}}$$

Fig. 9. Accuracy Formula

Performance of Deep Learning Models

Both the deep learning algorithms are allowed to run for a period of 10 epochs and their accuracy and loss are measured for each epoch. After fitting the model, the performance of the model is tested using our testing data. Both LSTM and GRU use categorical cross-entropy loss as their loss function. An Adam optimizer is used an optimization technique for faster convergence of our model.

DL Algorithm	**Loss**	**Accuracy**	**Runtime (for 10 epochs)**
Long Short-Term Memory	0.4352	0.9032	222 seconds
Gated Recurrent Unit	0.5375	0.8730	129 seconds

Table. 2. Performance metrics for DL models

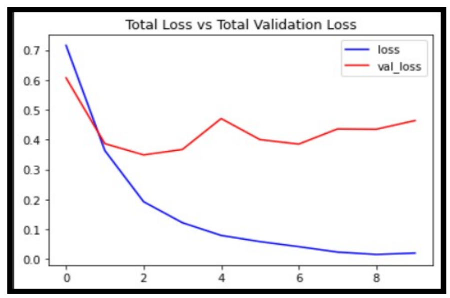

Fig. 10. Performance graph (Loss vs Validation Loss) for LSTM Model

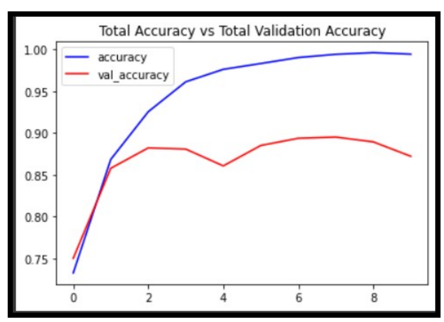

Fig. 11. Performance graph (Accuracy vs Validation Accuracy) for LSTMModel

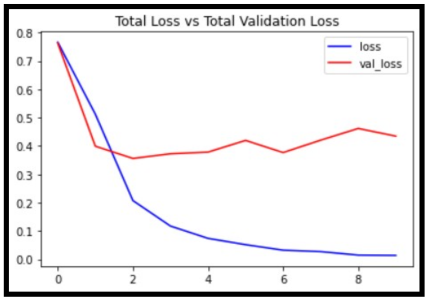

Fig. 12. Performance graph (Loss vs Validation Loss) for GRU Model

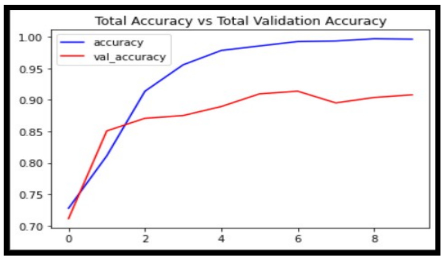

Fig. 13. Performance graph (Accuracy vs Validation Accuracy) for GRU Model

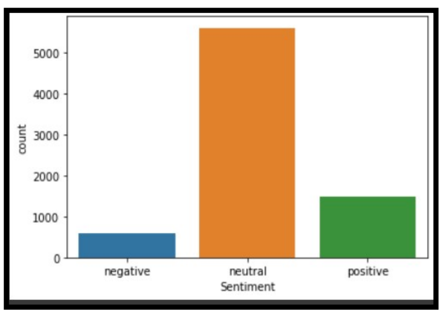

Fig. 14. Graph denoting the distribution of Sentiments across various classes

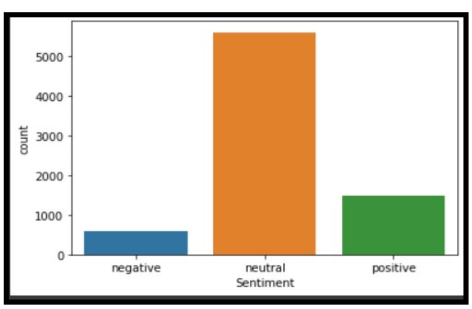

Fig. 15. Graph denoting the distribution of Sentiments across various classes for Twitter

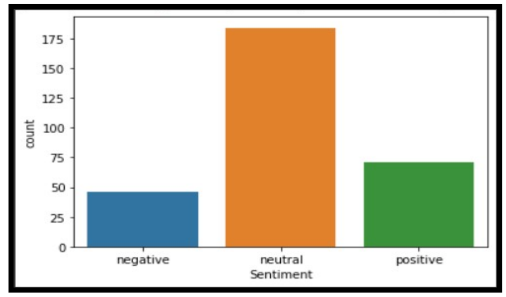

Fig. 16. Graph denoting the distribution of Sentiments across various classes for Google News

CONCLUSION

From the results generated, the problem of overfitting arised in both LSTM and GRU. This is due to high polarization in the dataset towards a particular class label (neutral). The models tended to overfit after running for a few epochs. The models were also found to give reasonable accuracy and loss during testing. The introduction of an early stopping mechanism and an unpolarised dataset can help in building a very coherent sentimental analysis model. In the Machine Learning models, the SVM model outperforms the other two models (Naive Bayes classifier, Decision Tree Classifier), but it is to be noted that the time taken for the SVM model to converge is very high compared to the other two. The Naive Bayes classifier converges the fastest among the three as it assumes each feature is independent of the other. It is to be noted that for sentimental analysis SVM performs better however the decision tree classifier has the highest accuracy.

Finally, the results with a transformer pre-trained model (Distilbert and Roberta) model are calculated. The pre-trained model also generated similar results when trained under a polarised dataset. Thus, further improvements can be made in the collection of datasets and fine-tuning the parameters of each model as per requirement.

REFERENCES

[1] Kosorok MR, Laber EB. Precision medicine. Annual review of statistics and its application. 2019;6:263-86.

[2] Psaty BM, Dekkers OM, Cooper RS. Comparison of 2 treatment models: precision medicine and preventive medicine. Jama. 2018;320:751-2.

[3] Manrai AK, Patel CJ, Ioannidis JP. In the era of precision medicine and big data, who is normal? Jama. 2018;319:1981-2.

[4] Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018;2:719-31.

[5] Zhang X, Yang H, Zhang R. Challenges and future of precision medicine strategies for breast cancer based on a database on drug reactions. Bioscience Reports. 2019;39.

[6] Kroschinsky F, Stölzel F, von Bonin S, Beutel G, Kochanek M, Kiehl M, et al. New drugs, new toxicities: severe side effects of modern targeted and immunotherapy of cancer and their management. Critical Care. 2017;21:89.

[7] https://www.gminsights.com/industry-analysis/precision-medicine-market. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[9] Shraddha Pai, Gary D. Bader, "Patient Similarity Networks for Precision Medicine," J Mol Biol, Elsevier ,2018, 430,pp. 2924-2938.