Stock Market Prediction Using Sentiment Analysis

Shubham Raj¹, Sindhu Yadav¹, Md. Meraj Alam¹, Vijay Kumar¹, Pruthvi P R²

¹Students, Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India. ²Assistant Professor, Dept. of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India. Corresponding Author: shubhamrajbest@gmail.com

Abstract: - Social sites like Twitter help millions of people to share their thoughts about the Stock market and what they feel about them. The tweet may be a short and easy sort of expression. Detecting sentiments in-text features a wide selection of applications including identifying anxiety or depression of people and measuring the well-being or mood of a community. Therefore, during this review paper, we focused on Sentiment Analysis of Twitter data. Sentiments are often expressed in some ways, which will be seen like countenance and gestures, speech, and transcription. Sentiment Analysis in text documents is actually a content-based classification problem involving concepts from the domains of tongue processing also as Machine Learning. Using different aspects, the research of Sentiment Analysis of Twitter Data is often performed. We can see the various sorts of Sentiment Analysis and techniques want to perform the extraction of the info. In this paper, we have taken a comparative study of various approaches and techniques of sentiment analysis having Twitter as knowledge.

Key Words: — Stock Market Prediction, Machine Learning, Sentiment Analysis, Twitter API.

I. INTRODUCTION

The most dynamic and advanced means of doing business is the stock market, commonly known as the stock exchange. It is a complex model for little companies, investors and therefore the banking sector to all or any generate revenue and minimize risks. This paper would however plan to use open-source datasets and current data to predict future exchange rates employing a machine-learning algorithm. In the course of years, the share market has been a crucial a part of the expansion of the many companies also as of a country's GDP. In the financial markets of the worldwide private sector, stock markets are given the foremost important position in economic liberalization. There have been a number of flexible impacts on stock markets, the most essential of which are historical data. Many approaches for forecasting stock related data were developed using different techniques and models, which used traditional prices, past revenue growth and dividends, so we all know that we'd like data alongside one among the above factors to effectively predict stocks, in order that the effective market hypothesis are often built.

Manuscript revised July 16, 2021; accepted July 17, 2021. Date of publication July 18, 2021. This paper available online at <u>www.ijprse.com</u> ISSN (Online): 2582-7898 In this paper, the Twitter Application Programming Interface (Twitter API), which offers a streaming API, has been taken under consideration within the study of monetary data and continually returns the info. Each data collected reflects the user's status or attitude with reference to a selected subject. This is available between a basic HTTP documentation and a twitter account. After all data is composed for each line, an interpretation is initiated of the emotions relevant to every tweet then a mood is predicted which features a direct impact on the stock status. Sentiment analysis is essentially a drag of classification during which the info content is categorized with a positive or negative opinion. Various models are developed supported various learning algorithms used for the training results. The flowing data is collected through the flowing API after such a model is developed.

II. LITERATURE SURVEY

Twitter may be a popular social networking website where users post and interact with messages referred to as "tweets". This is a way for people to precise their thoughts or feelings about different subjects. Various different parties like consumers and marketers have done sentiment analysis on such tweets to collect insights into products or to conduct marketing research. With the recent advancements in machine learning algorithms, the accuracy of our sentiment analysis predictions is in a position to enhance. During this report, we had planned to conduct sentiment analysis on "tweets" using several different machine-learning algorithms. We plan to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked because of the final label.

We use the dataset from Kaggle which was dragged and labelled positive or negative. The info provided comes with emoticons, usernames, and hashtags, which are required to be processed and converted into a typical form. It also must extract useful features from the text like unigrams and bigrams, which may be a sort of representation of the "tweet". We exercise various machine-learning algorithms to conduct sentiment analysis using the extracted features.

III. METHODOLOGY

This project consists of four phases:

- Data Collection
- Data Processing
- Data Filtering
- Feature Extraction

A. Data Collection

The input data of raw tweets is recovered by using the Scala library called "Twitter4j" which provides a package for real time twitter streaming API. The API requires us to register a developer account with Twitter and fill in parameters like consumer Key, consumer Secret and Token Secret. Twitter API policy allows the developer to extract 50,000 tweets per month. This API extracts random tweets on a particular subject using keywords as filter. Filters supports to retrieve tweets, which match a selected criterion defined by the developer. We used this to retrieve tweets associated with specific keywords, which are taken as input from users.

B. Data Processing

Data processing is that the process of splitting the tweets into individual words called tokens. Tokens are commonly divided using whitespace or punctuation characters. It can be single word or cluster of words depending on the classification model used. The bag-of words model is one of the majorly used model for classification. It is supported the very fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest thanks to incorporate this model in our project is by using unigrams as features. It is just a set of individual words in the text to be classified, so, we split each tweet using whitespace. Tweets are simplified by converting it to lowercase which makes its comparison with a dictionary easier.

ί.	A	- 1	¢	0		1.1	
	- 3	d	cented_at	bed.	probability	settinet	
	0	1413075-08	Thu Jul 08 09:30:59 +0000 2021	BHP's lot new Kang; BHP's also how the rest of the UK perceive you too. English ecoptionalism is the problem. England is a lovely country with ma	0.990991	ISS NEGATIVE	
	1	1403075-08	The lot 08 09/24/03 +0000 2021	Aules Britannia pol save the queer	0.838230	119 POSTIVE	
	2	1403075-08	The Jul 28 09 20 26 +0000 2021	والم معرفة الجارية مراجعة ومترك والمراحة و	0.634218	DE NEGATIVE	
	3	1413078+38	Thu Jul DE 09:19:59 +0000 2021	Penalty my falorg ania. Stop lying to unselver, he connect the net WM needs binned it's hopeless. BPI -BPI(BPI & \$6.00 % it's 10 's (BPI -BPI) Rule Britannia	0.999875	7/1 NEGATIVE	
	4	141306+18	The tel 18 09:54:48 -0000 2021	The Britamia-Rotel? It's not the nicest looking building though.	0.99826	ECS NEGATIVE	
	5	141306-08	The Jul DE 09:13:23 +0000 2021	When would you like me to leave a seriously disabled person on their own in order to regin society. Are you willing to rouch for his safety. Are you will	0.995638	ISD NEGATIVE	
	6	140306-08	The Jul 08 09:52:39 +0000 2021	Britannia Industries Big block of \$7,925 shares at 3480 worth worth 34.07 store rupees Milifly	0.99789	SHE POSTIVE	
	2	141306-08	The tol 08 09:11:09 40000 2021	RT : Anyone remember the relaxed patriotic self-confidence of "Cosi Britannia", Brit Pop, the Condon Olympics, TursiAl,	0.95385	NI ROSTINE	
		1.413062+08	The tel 08:09-36-40000 2021	KT : Britania 7,25 %	0.88602	Std ROSTINE	
	9	141306+18	Thu to/ 08 09:08:59 +0000 2021	We love to see it? Absolute scenes live onboard Britannia (#7 as took its penalty in last sightife"s #EURO3021 sami finall Kport24 RiveOnboard Ki-	0.999034	421 POSTNE	
	30	14006-0	The Jul DE 09:08:34 +0000 2021	No. Irith Karry: Scottish cartolics hate England. Pin Scottish Karry: British and prouf to be so. #14-#19(#7-6)(0-13-7-6)(-0-2#14-#14) Aule Britannia	0.829933	SA ROSTINE	
	31	140306-08	7% (4/18/09/07/47 -0000 2021	Rule Bittamia)	6.90047	ISS NEGATIVE	
ł	32	1413062-18	The Jul 28 09:02:56 -0000 2021	Seriously, I don't like italy (political masors) but I can't bear an England win	0.982399	DEG NEGATIVE	
	18	141306-18	The Av108 09:01:36 +0000 2021	Britannia will set sail for Local Coastal this afternoon at 610 prs. Abritannia	0.996100	DER POSITIVE	
	38	140306-08	The last 08 09:00:05 +0000 2021	Britannia The Bear w Emirs Ty Beanie Bables NB D468ROOK rave, "All 4 Completer" eBay	0.82464	IOB NEGATIVE	
	15	141306-18	Thu Jul DE 08:59:15 +0000 2021	RT : Multiple Britannia use originally a nostalge call to arms, reflecting on a modern precalisaaness in contrast to long past victil?	\$9437	IST POSITIVE	
	36	141306-18	The tail 08 08:57:47 +0000 2021	til/"s coning hone. Rule Ditamia	0.994751	WTROF CC	
	17	141306-08	The Jul 28 08:56:47 +0000 2021	1. Anul 2. Knan 3. Britannia Following u on all platforms	0.99687	3VTRON 521	
		1413062-38	The lot DE DE 56:00 +0000 2021	RT: 1. Parle 3. Britannia 3. Nestle #13star/schlasidine Join In: 341	0.989432	INTROFT IN	
	39	5.45306-08	The Jul DE 08:53:50 +0000 2021	87: 8740 Over 10,000 excited people have already made their way through the doors of the state of the wit new 'Initiania Leisaria',	0.953962	266 205714	
	20	1.413066+38	The Jul DE 08:54:07 +0000 2021	des partices 7 Rule, Britannia/ Britannia, nue the waves?	0.99578	NT POSTNE	
	21	1413062+18	The Jul DE OR 53:50 +0000 2025	Mathate Britamia was originally a nostalgic call to arms, reflecting on a modern precariosones in contrast to long past victories. After major victories it	0.993555	BY ROSTINE	
	22	1413068-08	The Jul DE OR 52:49 +0000 2021	In a way, KHP*s best that England play Italy in the final. 50 conflicted about this England team. Some great players, but above all genuinely good human	0.955866	IN NEGATIVE	
	23	141306-11	Thu Jul 28 GB 49-KB +2000 2021	Too right, once again CED yould?" we captured the mood of the nation, I think the only time these young people should be allowed to make a noise is too	0.990049	75 ROSTINE	
	34	141306-08	Thu Jul 08 08:46:57 +0000 2021	#1: Rule Britannia, Britannia nairs the wavent #13-#135 #17 Go on England-#126 KMC*s coming home #126	0.996063	ANT POSITIVE	
	25	14106-08	The Jul 28 08:42:59 +0000 2021	Used Car salesman and an Adult child. These are the mandards we take high. Rule Britannia, teep:	0.909047	NTRON SHE	
	28	1413062-08	The Jul DI OR 40:05 +0000 2021	Activenia go for it good log at CAP	0.965498	HIP POSTINE	
	37	141306-18	The Jul DI OR 39-94 -0000 2021	Will the Roman team to be "taking the kneede" when they come to play east Britannia?	0.9687	THE NEGATIVE	
	28	143854-0	The 3x1 04 08:33:52 +0000 2021	Rule Britannial (FT-dTa)	0.967253	26 205734	
	29	1413056-08	The Jul DI 08/32/23 +0000 2021	87: #Britannia rear support of 3380-3430 with PSI new 30: possible that it will turn from here. Low risk high reward trade. VMC	0.990588	ISB NEGATIVE	

C. Data Filtering

A tweet received after processing still have a portion of raw information in it which we may or might not be useful for our application. Therefore, these tweets are further filtered by pulling out stop words, numbers and punctuations. Stop words: For example, tweets fulfill stop words which are enormously common words like "is", "am", "are" and holds no valuable information. These words set out no purpose and this property is implemented employing a list stored in stopfile.dat. We then compare each word in a tweet with this list and delete the words matching the stop list as Code snippet for stop words removal removing non-alphabetical characters: Symbols like "#, @" and numbers hold no relevance just in case of sentiment analysis and are removed using pattern matching. Regular expressions are used to match alphabetical characters only and rest is ignored. Code snippet for removing non-alphabets this helps to scale back the clutter from the twitter stream. Stemming: it's the method of reducing derived words to their roots.

1	A	8 C	D	E	F
5	3	141193E+18 Mon Jul 05 06:24:23 +0000 2021	H, we reget the incomesience caused. Rease DM as your contact details by clicking on the link below and our official will contact you.	0.99563884	7 NEGATIVE
6	4	1.41193E+18 Mon Jul 05 06:24 10 +0000 2023	I applied for a credit card. Video EVC was done. I received the application form with acknowledgement number but an unable to track it due to so-	1.99704319	2 NEGATIVE
7	5	14019/E+18 Mon Jul 05 06/23 46 +0000 2021	#1: On #NationaEostorsDay, ICO Bank introduces "Salute Eostors" banking solution for medical doctors. An extension of "ICOLH"	0.60988771	9 POSITIVE
8	6	1.4119/E+18 Mon 3/4/05 06:23:42 +0000 2021	What details and which email ?	1,9929638	8 NEGATIVE
9	1	1.413936+18 Mint Jul 05 06:23:35 +0000 2021	W: Right from exploring the local tourist attractions to experiencing the culture from a closer lens, it was just perfect And whalk?	1.99934691	2 POSTINE
10	8	1.43393E+18 Mice 34-05 06-22-59 +0000 2025	Hurry up its on tonomou look your seat now	1.143678	PATRON 7
11	9	1.41193E+18 Man Jul 05 06:22:21 +0000 2021	could you please start taking your customers seriously and stop blaning your customers just because your security system is not robust and intelli-	0.99995648	8 NEGATIVE
17	22	1.4119/E+18 Mon Jul 05 06/22:16 +0000 202	I am not able to contact customer care. I need my passbook and cherpe look delivered ASAP to my permanent address. Please contact me	0.99985390	A NEGATIVE
13	11	1.41193E+18 Mon Jul 05 06:22:54 +0000 202	Amost Have been 16 Days and My money still not credited into my bank. What You have investigate till now. I am doing frustrate. No Call Review	0.99955445	S NEGATIVE
14	12	141193E+18 Mon Jul 05 06 22:13 +0000 2021	Really nice plans for Current account by Acpefully they do a online/ email sign up instead of an unnecessary branch vist	0.73328405	6 POSTINE
15	13	1.4129/E+18 Mon Jal 05 06/22-12 +0000 2021	MY CRAD WHS CLOSED FEW CARS BICK AND THERE IS NO OS amount and nill my account got auto debited for 2000+ amount as a auto debit. How	1,9992340	8 NEGRITIVE
16	14	1.41193E+18 Mon Jul 05 06:21:57 +0000 2021	ket f	0.98957219	4 POSTM
17	15	14119/E+18 Mon Jul 05 06:28:52 +0000 2021	H, we repet the inconverience caused. We will have our official get in touch with you.	0.82993848	A NEGATIVE
18	15	1.40193E+18 Mon Jul 05 06:19:32 +0000 2023	Wy loan account LAGH200043E38175 It is a car loanand the instalment of Pt 34999 has been already debited from my account, as I did not took i	0.9998333	8 NEGATIVE
19.	12	1.41193E+18 Mon Jul 05 06:18:33 +0000 2021	87 : On MinistraelDoctorsDay, ICIC Bank introduces "Salute Doctors' tanking solution for medical doctors. An extension of ICICIAE]	0.60988771	9 POSTINE
20	18.	14019/E+18 Mph Jol 05 06:25:07 +0000 2021	Dear Castomer, your transaction of NR oox on ICO. Bank Credit Card XI is withheld due to security reasons. Please call our Castomer Care. Ph hel	0.99536848	9 NEGATIVE
28	19	1.42199E+18 Mice 34/05 06:34:24+0000 2021	all your ICO BANK SERVICES related queries request may you to please tog , or unite to headernicequality con Alternatively you may please visit	1.98866321	7 NEGATIVE
22	21	1.41193E+18 Mon Jul 05 06:12:52 +0000 2021	Achta, Rupye card other countries me nahi chalti hogi to? 8%*	1.99996244	9 NEGATIVE
28	25	1.41193E+18 Man Jul 05 06:12:29 +0000 2021	Tranks for reaching out. Peace send as a DM with your contact information and Merchant ID or Terminal ID. Our team will follow up with you to	1.95962007	2 NEGATIVE
24	22	1.4119/E+18 Mon Jul 05 06:11:22 +0000 2021	IT : Mustin This MusionaDoctorsDay, ICO Bank introduces Salute Doctors', to express gratitude Bang, salute our illesavers forM (0.98343928	6 POSTINE
Z	23	1.4119/E+18 Mon Jul 05 06:11:06 +0000 2011	VSA international card tha	0.70014030	1 NEGATIVE
26	34	1.41158E+18 Mon 34/05 06:00:50 +0000 2021	Forotable. There rever requested till now for any services from Icic Direct side. The only thing which I did was feedbacks to that clients damp; Cor-	1.84533505	2 POSTINE
27	z	1.41193E+18 Mon.bil 05 06:10:48 +0000 2021	your executive rodely used if payment missed then they will do Samp, call my dashing numbers, is this the unettical policy by 100 Bank, it's nothing b	0.99953520	B NEGATIVE
28	8	1.4119/E+18 Max 14/05 06:02:00 +0000 2021	#OOBackInstalEbLoan is the one-stop solution to finance your higher education dreams. Pre-approved customers of the Back can easily get an w	0.87832564	L NEGATIVE
28	27	1.4119/E+18 Mon Jol 05 06:09:39 +0000 2021	H, we repet the inconvenience caused. We will have our official get in touch with you.	1.82993848	A NEGATIVE
30	28	1.41193E+18 Mon 1xF05 06:09-25 +0000 2025	87: On #NationalDoctorsDay, ICO Bank introduces 'Salute Doctors' banking solution for medical doctors. An extension of 'ICOLIAC'	0.60588771	S POSTNE
31	29	1.40193E+18 Mon Jul 05 06-09-03 +0000 2023	H, we repet the inconvenience caused. We will have our official get in touch with you.	0.82953848	A NEGATIVE
12	30	1.4019/E+18 Mon Jul 05 06:08:22 +0000 2021	afters are your employees (CO?? Area/K**) they interested in handing customes/K**a issues/?	1.99803847	1 NEGATIVE
12	25	1.8179/F+18 Mex 34/05 06/08/12 +0000 2021	If we deady separative reconcertience record to you. We will have for official art in touch with you.	0.90512921	R AFGETHE

D. Feature Extraction

This Method Utilized In Text Mining to Seek Out the Importance of a Term to a Document within the Corpus. The Recommended API Is That The Data Frame Based API. This Feature Is Beneficial for a Case Where We'd like to Seek out Trending Topics or to Make Word Clouds. However, Our Project Is More Concentrated Towards Finding Sentiment In Twitter Streams So TF-IDF Is Not Implemented.



IV. RESULT AND DISCUSSION

Twitter API allows 50,000 tweets to be extracted per month for developing purposes.

In our project, in each iteration 100 real time tweets are extracted. The keywords are the specific subjects about what we have to check the prediction.

Keywords like ICICI will work as a filter to extract tweets related to that subject only.

After the filtration and processing module a tweet is becomes a collection of unigrams and multigram.

Sentiment analysis id performed on the collected processed data.

After sentiment analysis, each word in the sentence is classified as positive and negative.

After considering the polarity of each word, the polarity of whole sentence is calculated.

The final result is obtained by taking the average of all the negative tweets and all the positive tweets separately.

The probability distribution of polarity about a specific subject gives us the inference whether it is profitable to buy a stock or sell it.

😥 vjeykumar	=	
🙆 Stocks Prediction	Stocks Prediction	
🖌 Logout	DMSLA8	
	PROCT	
	BRITANNA Buy Probability: 54.09 BRITANNA Sel Probability: 55.35	
	DOWINGIAL STREETS	

V. CONCLUSION

During this project, we achieved two goals within the areas of machine learning and interactive visualizations. We developed an honest model of a Twitter sentiment prediction system using supervised text -classification technique and support vector machines. a mix of proven-published ideas (bag-of-words, lexical features, etc.) and novel ideas (polarity buckets, weightadjusted negation, etc.) contributed to achieving a highperformance system. Several aspects of the system are often improved with a further commitment to the present research. (1) In building a prediction model, a neighbourhood of ensemble machine learning is trending. Similar tasks have reported performance boosts using this system. (2) Hashtags can remarkably help in identifying tweet sentiment. However, since they are compound words, the system was not ready to leverage their full potential in indicating sentiment. (3) We reported a high-performance increase using sentiment lexicons. Since a number of these lexicons are automatically compiled and contain some noise, there is an opportunity for further performance boost if this noise is often reduced. (4) A comparative study of the proposed system on diverse short-text datasets would be required so as to utilize this research in other mediums of sharing sentiments online

REFERENCES

- [1]. Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012).
- [2]. Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000-rupee banknotes by Indian government." ICT Express (2017).

- [3]. Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.
- [4]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015).
- [5]. Amolik, Akshay, etal. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016).