

# Twitter Tweet Sentiment Analysis

Pratik Nath Tiwari<sup>1</sup> and Meenu Jain<sup>2</sup>

<sup>1</sup>Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

<sup>2</sup>Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

**Abstract:** In this study, an effort was made to make financial-related judgments, such as stock market research and projections, to anticipate the future prices of the company's shares, and Twitter data was evaluated to fill the demand for this. Streaming data continues to be a reliable source of real-time data processing. Streaming information is just a nonstop progression of information that brings data from sources like web journals, PDA applications, worker logs, social sites, exchanging floors, and so forth. The critical attributes of such information are their ease of use and accessibility, which help in the appropriate investigation and forecast of client movement in a consistent way. The grouping model produced using verifiable information can be persistently refined to yield even more exact outcomes, since the outcome is frequently contrasted with the following tick of the clock. Spark streaming has been considered for the handling of humongous information and information ingestion techniques, for example, the Twitter API and Apache Kafka have been additionally contemplated.

**KEYWORDS:** Spark Streaming, Sentiment Analysis, Recurrent Neural Network, Apache Kafka.



Check for updates

\* DOI of the Article: <https://doi.org/10.46501/IJMTST0706029>



Available online at: <http://www.ijmtst.com/vol7issue06.html>



As per UCC guidelines an electronic bar code is provided to secure your paper

To Cite this Article:

Pratik Nath Tiwari and Meenu Jain. Twitter Tweet Sentiment Analysis. *International Journal for Modern Trends in Science and Technology* 2021, 7, 0706149, pp. 169-174. <https://doi.org/10.46501/IJMTST0706029>

Article Info.

Received: 16 May 2021; Accepted: 6 June 2021; Published: 14 June 2021

## INTRODUCTION

Advancements have become a piece of our day by day lives and we can't deny the reality how simple and agreeable our lives have become with the world's gone on the web and rather more virtual. As an ever-increasing number of individuals are deciding to go on the web, it has gotten important to take a gander at the lives they lead on the web. The greater part of us thinks that its more consoling to offer thoughts, or to be general, data on the web. Social destinations furnish us with a stage to do likewise. Twitter, being a mainstream social site, draws in an immense rush hour gridlock. It's somewhat celebrated for where individuals love to impart their insights. To share data on Twitter as generally as could be expected, it gives organizations, designers, and clients with automatic admittance to Twitter information through their APIs (Application Programming Interfaces). The information can be put away in a data set utilizing certain innovations. The put away information can additionally be utilized to investigate certain circumstances, generally regular which is examining the effect of specific gatherings on individuals in a political decision or to break down how certain item, that has been as of late presented on the lookout, is doing. The serious issue, one runs over, while playing out this errand, is dealing with the flood of approaching information. Subsequently, we're utilizing Kafka, an open-source stream-preparing programming stage which gives a brought together, high-throughput, low-inertness stage for taking care of constant information takes care of. The undertaking is made with the point of investigation of tweets for a few informational collections recovered with the assistance of twitter4j API. Thus, the task includes preparing of the application over a given informational collection and afterward utilizing Sentiment Analysis.

Feeling examination (SA)tells client whether the information about the product is agreeable or not before they die. Advertisers and organizations utilize this investigation information to know concerning their product or administrations in such the manner that it is offered according to the client's necessities. Literary information recovery procedures in the principle focus on handling, looking or investigating the authentic information blessing. Realities have A target part anyway there are a unit another printed substance that

downright emotional attributes. These substance region unit in the principal assessments, slants, evaluations, perspectives, and feelings, that type the center of Sentiment Analysis (SA). It offers a few troublesome chances to foster new applications, in the fundamental in light of the monstrous development of possible data on-line sources like sites and interpersonal organizations. For instance, proposals of things arranged by a suggestion framework is anticipated by considering issues like positive or negative feelings about those things by making utilization of Storm Troops.

## TECHNOLOGIES USED & BUILD ENVIRONMENT

**1. Twitter API:** The Twitter API lets you read and write twitter data. Thus, you can use it to compose tweets, read profiles, and access your followers' data and a high volume of tweets on particular subjects in specific locations.

**2. Apache Kafka:** Apache Kafka is an open-source stream-processing software platform developed by the apache software foundation, written in scala and java. Apache Kafka is based on the commit log, and it allows users to subscribe to it and publish data to any number of systems or real-time applications.

**3. Apache Maven:** Maven is a build automation tool used primarily for java projects. Maven can also be used to build and manage projects written in c#, ruby, scala, and other languages. The maven project is hosted by the apache software foundation, where it was formerly part of the jakarta project. Maven is built using a plugin-based architecture that allows it to make use of any application controllable through standard input. A plugin for the .net framework exists and is maintained, and a C/C++ native plugin is maintained for maven.

**4. Apache ZooKeeper:** Zookeeper is a circulated co-appointment administration to oversee enormous arrangement of hosts. Co-ordinating and dealing with an assistance in a dispersed climate is a confounded interaction. Zookeeper addresses this issue with its basic engineering and API. Apache ZooKeeper is an assistance utilized by a bunch (gathering of hubs) to arrange among themselves and keep up imparted information to strong synchronization procedures.

Zookeeper is itself a disseminated application offering types of assistance for composing a dispersed application.

The common services provided by ZooKeeper are as follows –

- i. **Naming service** – Identifying the nodes in a cluster by name. It is similar to DNS, but for nodes.
- ii. **Configuration management** – Latest and up-to-date configuration information of the system for a joining node.
- iii. **Cluster management** – Joining / leaving of a node in a cluster and node status at real time.
- iv. **Leader election** – Electing a node as leader for coordination purpose.
- v. **Locking and synchronization service** – Locking the data while modifying it. This mechanism helps you in automatic fail recovery while connecting other distributed applications like Apache HBase.
- vi. **Highly reliable data registry** – Availability of data even when one or a few nodes are down.

**5.PomXML:**POM means "project object model". It is an XML portrayal of an expert venture held in a document named pom.xml the pom contains all essential data about an undertaking, just as setups of modules to be utilized during the form interaction. It is the explanatory appearance of the "who", "what", and "where", while the form lifecycle is the "when" and "how". This shouldn't imply that that the pom can't influence the progression of the lifecycle - it can. For instance, by designing the expert antrun-module, one can implant apache subterranean insect assignments within the pom. It is eventually a revelation, nonetheless. Though a build.xml guides insect absolutely when it is run (procedural), a pom expresses its arrangement (decisive). On the off chance that some outer power causes the lifecycle to skirt the subterranean insect module execution, it doesn't stop the modules that are executed from doing their enchantment. This is not at all like a build.xml record,

where errands are quite often subject to the lines executed before it.

**6.Hibernate:** Sleep ORM (or essentially rest) is an article social planning apparatus for the java programming language. It gives a structure to planning an Object-Oriented area model to a social information base. Rest handles object-social impedance confound issues by supplanting immediate, persevering data set gets to with undeniable level article taking care of capacities. Rest is freeware. Rest's essential component is planning from java classes to data set tables, and planning from java information types to SQL information types. Rest additionally gives information question and recovery offices. It produces SQL calls and eases the engineer from the manual taking care of and object change of the outcome set.

**7.Spring boot:**Spring boot is an open-source java-based construction used to make a smaller than normal assistance. It is made by essential gathering. It isn't hard to make a free and creation arranged spring applications using spring boot. Spring boot contains a comprehensive establishment support for cultivating a small-scale help and engages you to encourage adventure arranged applications that you can accessible.

**8.Stanford CoreNLP:** Stanford CoreNLP is a Java normal language investigation library. Stanford CoreNLP coordinates all our NLP devices, including the grammatical form (POS) tagger, the named substance recognizer (NER), the parser, the coreference goal framework, and the conclusion examination devices, and gives model documents to investigation of English. Investigating text information utilizing Stanford's CoreNLP makes text information examination simple and proficient. With only a couple lines of code, CoreNLP considers the extraction of a wide range of text properties, for example, named-element acknowledgment or grammatical feature labeling. CoreNLP is written in Java and expects Java to be introduced on your gadget yet offers programming interfaces for a few famous programming dialects, including Python, which I will use in this exhibit. Moreover, it upholds four dialects other than English: Arabic, Chinese, German, French, and Spanish.

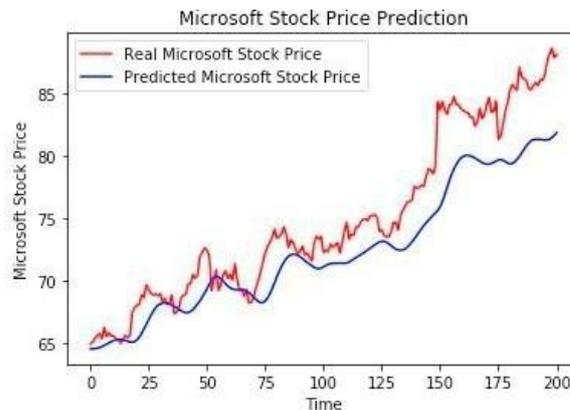
**METHODOLOGY**

1. A Twitter Developer account is created.
2. Integrate Apache Kafka with Twitter API to retrieve Real-time data streams.
3. Build a Java program on Maven and integrate it with Kafka, that further extends to a database to store the required data-set from the acquired data-stream retrieved from Twitter by Kafka.
4. Use queries to obtain the further sub-data from the database.
5. Use Stanford’s CoreNLP library to do further processing of the data set and analyze the data.
6. For every search term, find the tweets and perform sentiment analysis.
7. Finally, return the results list to the user.

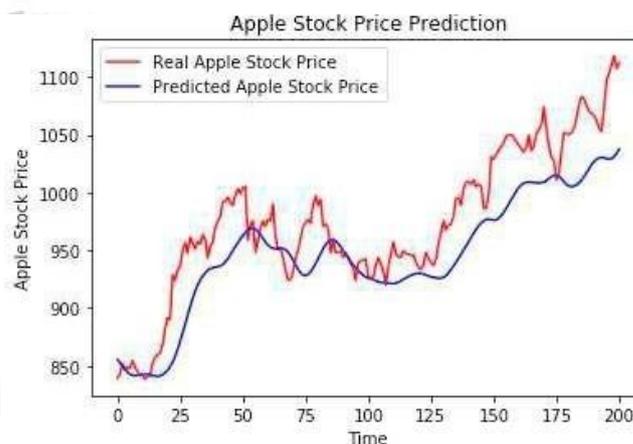
**RESULT AND DISCUSSION**

Google, Microsoft and Apple’s genuine stock costs have been accumulated utilizing the Yahoo Finance site, who’s yield pointer is seen as ground reality. The normal stock cost is contrasted and the extended stock cost for the testing information. The secured dataset incorporates 5,60,000 tweets spreading over 13 years of Twitter.

The guidance set includes a similar arrangement of Appropriately classified as sure and negative criticism of the organization viable. The Twitter information gathered the most recent 200 days are taken as a test dataset. Expectation diagram for every one of the organization-based history datasets names were planned along with a plot showing the weighted extremity of the tweets made on that very day.



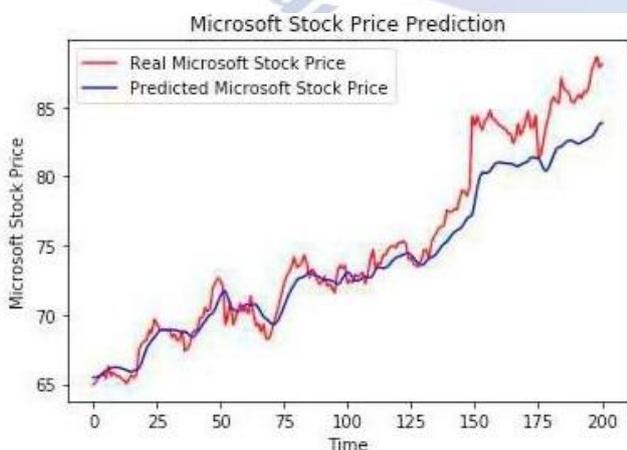
(b) Prediction for 300 Epoch



(a) Prediction for 150 Epoch



(b) Prediction for 300 Epoch



(a) Prediction for 150 Epoch

**CONCLUSION**

This venture holds a wide scope of utilization. It, for the most part, is created for investigating tweets and building prescient models. The prescient models are worked to help political race/promotion/showcasing efforts dive into online media discussions (popular feelings) to get bits of knowledge for settling on canny choices. Fundamentally, this task fosters the premise of Sentiment Analysis. Anticipating future upsides of stock costs is an intriguing examination region,

generally associated with the investigation of public state of mind. Given that an ever-increasing number of genuine beliefs are made accessible on the web, different examinations show that these sorts of investigations can be mechanized and can deliver valuable outcomes. The investigation shows that estimation examination of public temperament got from twitter channels can be utilized to in the end gauge developments of individual stock costs. Besides, the procedure was adjusted to a stream-based setting utilizing the gradual dynamic learning approach, which furnishes the calculation with the capacity to pick new preparing information from an information stream for hand labeling. Stream based dynamic learning for assumption investigation of microblogging messages in the monetary space may contribute both to conclusion examination and the dynamic learning research region. In addition, possibility concentrate through cluster handling has additionally been performed with this trial by taking the guide of RNNs Long Short-Term Memory(LSTM). This assists with advancing burrow upon the gushing of online stock information accessible in different monetary sites which will give a superior model to investigate and anticipate the future stock costs for a specific organization. In this way, a cross breed model dependent on investigation of the assessments and the current stock pattern in the ascent or fall of costs can be presented which will improve its unwavering quality just as trust-value of the expectation. In future, incorporation of mixed AI calculations should be possible for the expectation of stock information, for example, profound learning models. Different strategies for information ingestion, for example, information ingestion through Apache Flume or NodeJs can likewise be carried out and their exhibition and exactness might be thought about for the equivalent.

## REFERENCES

- Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10, 2010
- Xue Zhang, HaukeFuehres, and Peter A Gloor. Predicting stock market indicators through twitter i hope it is not as bad asi fear. *Procedia - Social and Behavioral Sciences*, 26:55–62, 2011.
- Michael Stonebraker, UurC,etintemel, and Stan Zdonik. The 8 requirements of real-time stream processing. *ACM Sigmod Record*, 34(4):42– 47, 2005.
- Michael Hausenblas and Nathan Bijmens. Lambda architecture. URL: <http://lambda-architecture.net/>. *Luettu*, 6:2014, 2015.
- Philip Russom et al. Big data analytics. *TDWI best practices report*, fourth quarter, 19:19–40, 2011.
- Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and MarimuthuPalaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, 29(7):1645–1660, 2013.
- Gema Bello-Organ, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.
- B. G. Malkiel. The e cient market hypothesis and its critics. *Journal of economic perspectives*, 17:59–82, 2003.
- De Rijke M. Mishne, G. Capturing global mood levels using blog posts. *AAAI spring symposium: computational approaches to analyzing weblogs*, 6:145–152, 2006.
- Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004.
- EijiAramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.
- Carl Sabottke, Octavian Suci, and Tudor Dumitras. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In *USENIX Security Symposium*, pages 1041–1056, 2015.
- AibekMakazhanov, DavoodRafiei, and Muhammad Waqar. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4(1):193, 2014.
- Nishant Garg. *Apache Kafka*. Packt Publishing Ltd, 2013.
- Pieter Hintjens. *ZeroMQ: messaging for many applications*. " O'Reilly Media, Inc.", 2013.
- Mayur R Palankar, Adriana Iamnitchi, MateiRipeanu, and Simson Garfinkel. Amazon s3 for science grids: a viable solution? In *Proceedings of the 2008 international workshop on Data-aware distributed computing*, pages 55–64. ACM, 2008.
- Ranjan Kumar Behera, Abhishek Sai Sukla, Sambit Mahapatra, Santanu Ku Rath, Bibhudatta Sahoo, and Swapan Bhattacharya. *Mapreduce based link prediction for large scale social network*. 2017.

18. Abdul GhaarShoro and Tariq Rahim Soomro. Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology*, 15(1), 2015.
19. Ranjan Kumar Behera, SK Rath, and Monalisa Jena. Spanning tree-based community detection using min-max modularity. *Procedia Computer Science*, 93:1070–1076, 2016.
20. Matei Zaharia, M Chowdhury, T Das, A Dave, J Ma, M McCauley, M Franklin, S Shenker, and I Stoica. Resilient distributed datasets. A Fault Tolerant Abstraction for In-Memory Cluster Computing, nd <http://www.cs.berkeley.edu/~matei/talks/2012/nsdirdds.pdf>, accessed April, 2014.
21. Matei Zaharia, Tathagata Das, Haoyuan Li, Scott Shenker, and Ion Stoica. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters. *HotCloud*, 12:10–10, 2012.
22. Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanfordcorenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
23. Leishi Zhang, Andreas Stoel, Michael Behrisch, Sebastian Mittelstadt, Tobias Schreck, René Pompl, Stefan Weber, Holger Last, and Daniel Keim. Visual analytics for the big data era: comparative review of state-of-the-art commercial systems. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 173–182. IEEE, 2012.
24. Thomas Niederkrotenthaler and Gernot Sonneck. Assessing the impact of media guidelines for reporting on suicides in austria: interrupted time series analysis. *Australian & New Zealand Journal of Psychiatry*, 41(5):419–428, 2007.
25. Nikolaos Nodarakis, Spyros Sioutas, Athanasios K Tsakalidis, and Giannis Tzimas. Large scale sentiment analysis on twitter with spark. In *EDBT/ICDT Workshops*, pages 1–8, 2016.