Twitter Data Streaming

Ritesh Kumar[1] and Vasudha[2]

[1]UG student, IT, Maharaja Agrasen Institute Of Technology, Delhi, India.
[2]Assistant Professor, Department of IT, Maharaja Agrasen Institute of Technology, Delhi, India.

Abstract: In this report, an effort was made to make financial-related decisions, such as stock market research and forecasts, to estimate the future prices of the company's stock, and to satisfy the need for this, Twitter data was considered to give the firm an idea. Streaming data continues to be a perennial source of data processing obtained in real time. Streaming data is simply a continuous flow of data that brings information from sources such as blogs, cell phone apps, server logs, social websites, trading floors, etc. The key characteristics of such data are their usability and availability, which aid in the proper analysis and prediction of user activity in a continuous manner. The classifying model made from historical data can be continuously refined to yield still more precise results, since the result is often compared to the next tick of the clock. Spark streaming has been considered for the processing of humongous data and data ingestion methods such as the Twitter API and Apache Kafka have been further studied.

KEYWORDS: Spark Streaming, Apache kafka, Recurrent Neural Network, Apache Maven, ZooKeeper

DOI of the Article: https://doi.org/10.46501/IJMTST0706054

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

As per UGC guidelines an electronic bar code is provided to seure your paper

To Cite this Article: Ritesh Kumar and Vasudha. Twitter Data Streaming. *International Journal for Modern Trends in Science and Technology* 2021, *7*, 0701074, pp. 324-328. https://doi.org/10.46501/IJMTST0706054

Article Info. Received: 15 May 2021; Accepted: 14 June 2021; Published: 22 June 2021

INTRODUCTION

Technologies have become part of our everyday lives and we can't ignore the fact that with the world going online and rather more virtual, our lives have become simple and relaxed. As more and more people choose to go online, looking at the lives they lead online has become important. It is more comforting for most of us to offer opinions or, to be general, information onlineTo do the same, social media provides us with a forum. Being a popular social network, Twitter receives a large amount of traffic. It's fairly common for a place where people love to express their views. It provides businesses, developers, and users with programmatic access to Twitter data through their APIs to exchange information on Twitter as broadly as possible (Application Programming Interfaces). Using some technology, the data can be stored in a database. In addition, the stored data can be used to evaluate certain circumstances, the most common of which is to analyze the influence of certain parties on individuals in an election or to analyze how certain goods that have recently been launched on the market perform. The key problem that one faces while performing this role is managing the stream of incoming information. We therefore use Kafka, a software framework for open-source stream-processing that offers a single, high-throughput, low-latency platform for handling real-time data feeds.

TECHNOLOGIES USED & BUILD ENVIRONMENT

1.Apache Kafka: Apache Kafka is an open-source stream-processing software framework written in scala and java, created by the apache software foundation. Based on the commit log, Apache Kafka enables users to subscribe to it and publish data to any number of systems or real-time applications.

2.Twitter API: The Twitter API helps you to read and write data about Twitter. You can therefore use it to write messages, read profiles, and access the data of your followers and a large volume of tweets in specific locations on specific topics.

3.Apache ZooKeeper: ZooKeeper is a distributed coordination service designed for the management of large hosts. It is a dynamic process to co-ordinate and operate a service in a distributed environment.With its simple architecture and API, ZooKeeper solves this problem. Apache ZooKeeper is a cluster (group of nodes) service used with robust synchronization techniques to coordinate between themselves and maintain shared data. ZooKeeper itself is a distributed application that offers distributed application writing services.

Some of the common services provided by ZooKeeper are –

i. **Cluster management** – Real-time joining/leaving of a node in a cluster and node status.

ii. **Leader election** – Electing a node for communication purposes as a representative.

iii. **Highly reliable data registry** – Data availability even when one or a couple of nodes are down.

iv. **Naming service** – Identifying the nodes by name in a cluster. It is close, except for nodes, to DNS.

v. **Configuration management** – The most current and up-to-date device configuration information for the joining node.

vi. **Locking and synchronization service** – Locking the data while it is being updated. This framework allows you to automatically recover from failure when other distributed applications such as Apache HBase are linked.

4.Apache Maven: Maven is a construction automation tool that is mostly used for Java projects. Maven can also be used to develop and manage c#, ruby, scala, and other language-written projects.The Apache Software Foundation, which was previously part of the Jakarta Project, hosts the Maven project. Maven is designed using a plugin-based architecture that allows standard input to monitor any program. A plugin for the.net framework exists and is maintained, and a native C/C++ plugin for maven is maintained.

5.PomXML: POM stands for "project object model". It is an XML representation of a maven project held in a file named pom.xml the pom contains all necessary information about a project, as well as configurations of plugins to be used during the build process. It is the declarative manifestation of "who" "what" and "where" while "when" and "how" are the building lifecycle.This does not mean that the pom is unable to influence the lifecycle flow - it can. For instance, one can insert apache ant tasks inside the pom by setting up the maven-antrun-plugin. However, it is ultimately a declaration. While a build.xml tells ant exactly what to do (procedural) when it is run, a pom states its configuration (declarative).If any external force causes the lifecycle to miss execution of the ant plugin, it does not stop the executed plugins from doing their magic. This is unlike a file called build.xml, where tasks almost always rely on the lines that are executed before them.

6.Hibernate: Hibernate ORM (or simply hibernate) is an object-relational mapping tool for the java programming language. It provides a basis for a relational database to map an Object-Oriented domain model. If the execution of the ant plugin causes some external force to skip the life cycle, it does not stop the executed plugins from doing their magic.This is unlike a file called build.xml, where tasks almost always rely on the lines that were executed before it.This creates SQL calls and releases the developer from the result set's manual handling and object conversion.

7.Spring boot: Spring boot is a java-based open source platform used to build a micro service.. Using a spring boot, it is simple to build stand-alone and production ready spring applications. Spring boot provides extensive support for the development of a micro service infrastructure and helps you to build enterprise-ready applications that can be on-hand. It is developed by pivotal team.

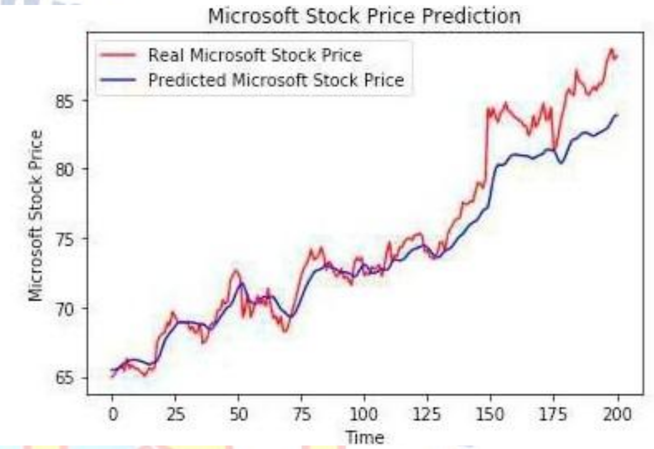
METHODOLOGY

1. A Twitter Developer account is created.
2. To retrieve real-time data sources, integrate Apache Kafka with the Twitter API.
3. Create and integrate a Java program on Maven with Kafka, which further extends to a database to store the appropriate data set from the acquired data stream obtained by Kafka from Twitter.
4. Using queries to access additional sub-data from the database.

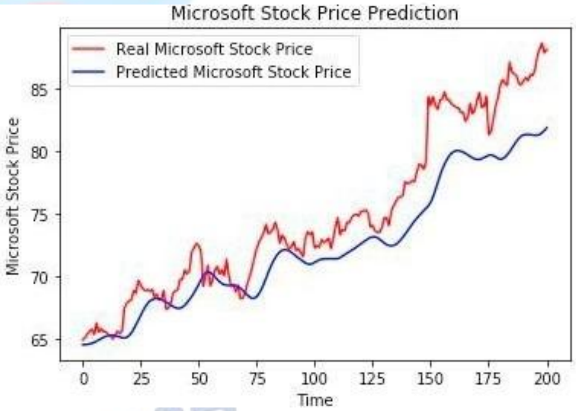
RESULT AND DISCUSSION

The actual stock prices of Google, Microsoft and Apple have been compiled using the Yahoo Finance website, which views the performance indicator as ground fact. The estimated stock price for the testing data is compared with the predicted stock price. 5,60,000 tweets spanning Twitter's 13 years are included in the procured dataset.

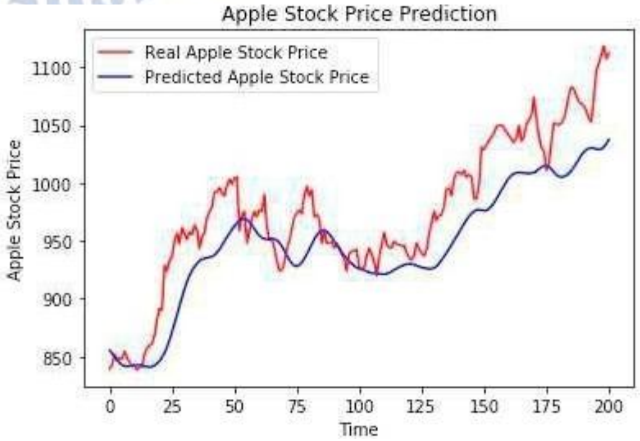
The set of guidelines contains the same set of correctly listed as positive and negative reviews from the organization under consideration. Collected Twitter data The last 200 days was taken as a dataset of studies. The prediction graph was mapped along with a plot showing the weighted polarity of the tweets made on that very day for each of the company-based background dataset names.



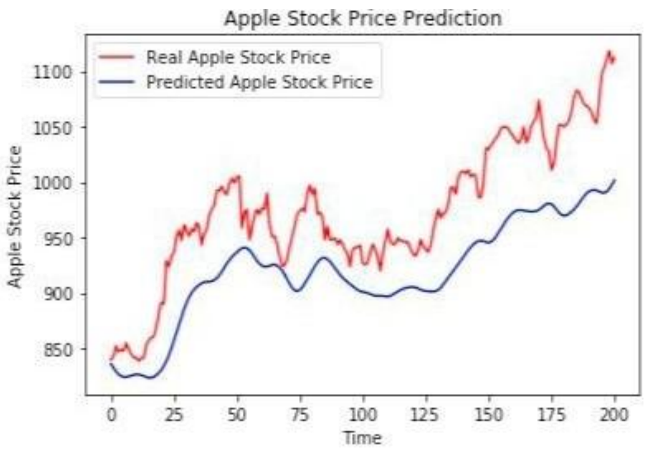
(a) Prediction for 150 Epoch



(b) Prediction for 300 Epoch



(a) Prediction for 150 Epoch



(b) Prediction for 300 Epoch

CONCLUSION

This project has a broad variety of applications. It is typically developed for the study of tweets and the development of predictive models. To help election/ad/marketing campaigns dig into social media discussions (public opinions) in order to get insights into intelligent decision-making, predictive models are developedDifferent studies suggest that these forms of research can be automated and can yield useful results, as more and more personal opinions are made accessible online. The study shows that public mood sentiment analysis derived from Twitter feeds can essentially be used to predict individual stock price movements.Different studies suggest that these forms of research can be automated and can yield useful results, as more and more personal opinions are made accessible online. The study shows that public mood sentiment analysis derived from Twitter feeds can essentially be used to predict individual stock price movements.

In addition, using the incremental active learning approach, the technique was adapted to a stream-based environment, which allows the algorithm the opportunity to select new training data from a data stream for hand labeling. Stream-based active learning for sentiment analysis of financial microblogging messages can lead to both sentiment analysis and active learning research. In addition, this experiment has also been conducted with the assistance of RNNs Long Short-Term Memory feasibility analysis via batch processing (LSTM). This allows to further explore the streaming of online stock data accessible on different financial websites, which will provide a better model for a specific business to evaluate and forecast future stock prices.Thus, it is possible to implement a hybrid model based on an interpretation of the emotions and the current stock trend in the rise or fall of prices that will boost its reliability as well as the prediction's trustworthiness. In the future, it is possible to use eclectic machine learning algorithms for stock data prediction, such as deep learning models.Other data ingestion strategies such as data ingestion via Apache Flume or NodeJs can also be introduced and can be compared for the same performance and accuracy.

REFERENCES

1. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. LREc, 10, 2010
2. NRC/National Research Council. Management of Technology: The Hidden Competitive Advantage . Washington DC: National Academy Press;1987.
3. Liao S. Technology management methodologies and applications. A literature review from 1995 to 2003. Technovation 2005;25:381–393
4. Michael Hausenblas and Nathan Bijnens. Lambda architecture. URL: http://lambda-architecture. net/. Luettu, 6:2014, 2015.
5. Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. Future generation computer systems, 29(7):1645–1660, 2013.
6. Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. Information Fusion, 28:45–59, 2016.
7. B. G. Malkiel. The e cient market hypothesis and its critics. Journal of economic perspectives, 17:59–82, 2003.
8. De Rijke M. Mishne, G. Capturing global mood levels using blog posts. AAAI spring symposium: computational approaches to analyzing weblogs, 6:145–152, 2006.
9. Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. IEEE Transac-tions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(3):334–352, 2004.
10. Eiji Aramaki, 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.
11. Carl Sabottke, Octavian Suciu, 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.

12. Aibek Makazhanov, Davood Rafiei, and Muhammad Waqar. Predicting political preference of twitter users. Social Network Analysis and Mining, 4(1):193, 2014.
13. Nishant Garg. Apache Kafka. Packt Publishing Ltd, 2013.
14. Tschirsky HP. The role o f technology forecasting a nd assessm ent in technology management. R &D Management 1994;24(2):121–129.
15. Łunarski J. Zarządzenie technologiami. Ocena i doskonalenie [Technology management. Evaluation and improvement]. Rzeszów: Oficyna Wydawnicza Politechniki Rzeszowskiej; 2009.
16. Pieter Hintjens. ZeroMQ: messaging for many applications. " O'Reilly Media, Inc.", 2013.
17. Mayur R Palankar, Adriana Iamnitchi, Matei Ripeanu, 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.
18. 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.
19. Abdul Gha ar Shoro and Tariq Rahim Soomro. Big data analysis: Apache spark perspective. Global Journal of Computer Science and Technology, 15(1), 2015.
20. Ranjan Kumar Behera, SK Rath, and Monalisa Jena. Spanning tree based community detection using min-max modularity. Procedia Computer Science, 93:1070–1076, 2016.
21. 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/nsdi rdds. pdf, accessed April, 2014.
22. Leishi Zhang, Andreas Sto el, Michael Behrisch, Sebastian Mittelstadt, Tobias Schreck, Rene´ Pompl, Stefan Weber, Holger Last, and Daniel Keim. Visual analytics for the big data eraa 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.
23. 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.
24. 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.
25. Badawy MK. Technology ma nagement education: alternatives models. California Management Review 1998;40(4):94–116.
26. Klincewicz K. Zarządzanie technologiami. Przypadek niebieskiego lasera [Technology management. The case of blu e laser]. Warszawa:Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego; 2010.
27. Cetindamar D, Wasti SN, Ansal H, Beyhan B. Does technology management research diverge or converge in developing and developed countries Technovation 2009;29:45–58.
28. Va n Eck NJ, Waltman L. VOSviewer Manual. Manual for VOSviewer version 1.3.0. software documentation; 2011.