Near Real Time Twitter Sentiment Analysis and Visualization

1 Wafaa S. Albaldawi, 2 Rafah M. Almuttairi

Rafah@uobabylon.edu.iq ; stud.wafaa.shakir@uobabylon.edu.iq

1Ministry of Higher Education and Scientific Research, Department of Computer Science, College of Science, University of Kerbala, Karbala, Iraq

1,2Ministry of Higher Education and Scientific Research, Department of Software, College of Information Technology, University of Babylon, Hillah, Iraq

Abstract

Twitter can be considered as a large scale network. People's opinions matter a lot to analyze how knowledge spreads impact lives. In this project, we took advantage of the Apache Spark Streaming fast and memory computing platform to retrieve live tweets and perform sentiment analysis. The primary purpose is to provide a tool to evaluate the score of sentiments in streams. This paper reports on the nature of an analysis of emotions, collecting vast numbers of tweets. Results identify the view of users through tweets into positive, neutral and negative about coronavirus.

This project on Spark Streaming to analyze tweets, hashtags or specific keyword/keywords such as (corona) from live twitter data streams. Data is collected from input sources like Twitter and processed downstream using Spark Streaming. Then, how sentiment scores can be generated for tweets and build visualization dashboards on the data using Elasticsearch and Kibana.

Keywords: Twitter, Spark Streaming, Kibana, dashboard, Elasticsearch, Sentiment analysis.

1. Introduction

Twitter is one of the biggest platform in Social network systems. Every second, tweets are released. People tend to express their real feelings openly in Twitter, making it an excellent medium for gathering thoughts on different topics of interest. For example their health, the health care system in their countries, products and more. In this paper, the tweets analysis based on real time. Twitter data is implemented by using Apache Spark streaming and visualization by building the dashboard. Specifically we used the corona china keywords as it is a popular keywords nowadays.

Users are now likely to disseminate information about various aspects of Twitter, using brief 140-character messages called "tweets". Moreover, they track other users sequentially to receive their tweet status updates. Twitter has a broad dissemination instant messaging site in nature and people use it to get updated about world news, current scientific developments, etc. Sentiment is called "a way of expressing one's feelings, views or attitudes, especially based on emotion rather than reason". (1)
In some situations, knowledge about the general inclination of opinion towards a particular topic can be used tremendously. Presumably a mechanical company would be interested in thinking about their customer's impressions of the latest piece, with a specific aim to answer input that will be used in the development of the next gadget.

The issue of opinion inquiry was widely considered in the late years. Many current structures in combined conditions are minimal and are based on characteristic dialect handling techniques and machine learning approaches. Tedious and computationally serious strategies of this kind. Subsequently, underlining arrangements are neither adequate nor suitable for conclusion mining, as there is a huge bungle between their capacity to plan and the production of usable information. It is clear that swing to highly adaptable arrangements is a basic necessity. Distributed innovations in computing provide tools and foundations for making such arrangements and dealing with information distributed across different servers. The Map Reduce programming model (5), created by Google, is the most prominent and remarkably effective tool for handling large-scale information.

2. Previous Work:
Authors aim to develop a real-time remote health status prediction system built around the open source Big Data processing engine, the Apache Spark, deployed in the cloud that focuses on applying the Big Data streaming machine learning model. The user tweets his / her health attributes in this scalable system, and the application receives the same in real time, extracts the attributes and applies machine learning model to predict the health status of the user, which is then immediately sent to him / her for appropriate action.

The authors in aim to create an Apache Spark cloud-based Sentiment Analysis platform for Twitter information that classifies tweets using supervised learning techniques. We are experimenting with binary and ternary classification and concentrating on the shift in accuracy induced by the size of the training dataset and the features derived from the feedback. Naive Bayes, Maximum Entropy and SVM classifiers were used by Pang et al to evaluate the feelings of film reviews.

Naveed et al evaluate tweet posts and estimate the probability that they will be retweeted on their content for a given post. Authors suggest that tweets with negative emotions are more likely than tweets with positive emoticons to be retweeted.

Previous work on emotional content is the one in; it presented the approach to the automatic analysis of tweets and the recognition of the emotional content of each tweet based on the Ekman emotional model, specifying the existence of one or more of the six basic human emotions (Anger, Disgust, Fear, Joy, Sadness and Surprise).

Wang et al used training data from 17000 Tweets to build the U.S. 2012 Presidential Election Cycle's real-time Twitter Sentiment Analysis Program.

Furthermore, authors present a novel approach for Sentiment Learning within the Spark framework; the proposed algorithm uses the hashtags and emoticons within a tweet as sentiment identifiers, and proceeds to a parallel and distributed classification procedure for various types of sentiment.

The sentimental evaluation method was discussed by Achin et al. They conducted Renewable Energy's sentiment analysis and identification function of tweets. To classify tweets into three groups, they implemented five separate machine learning algorithms. Both feature selection strategy and feature selection techniques, they bore classification. They have used choice methods for CfsSubsetEvaluation and Information Gain to reduce the number of data set features. Their results show that with feature selection methods, the accuracy of classification of feelings is better. With Support Vector Machine (Using PUK Kernel) and CfsSubsetEval feature selection method, the best accuracy (92.96 percent) is achieved.

Under the framework of Apache Spark, Al-Saqqa et al collected 4 million Amazon customer review datasets for broad-based sentiment analysis. The dataset was tested for the supervised algorithm of machine-learning, where the model was trained using the labeled training set. It
applied classification techniques where vector support outperforms Naïve Bayes and logistical regression, achieving 86 percent accuracy.

Analyzing the sentiment of diabetes patients on the twitter platform.\textsuperscript{16}

To develop a real-time remote health status prediction system using open source big data technologies and cloud.\textsuperscript{17}

To examine the opportunities from big data in retailing for predictive analytics with Bayesian techniques.\textsuperscript{18}

Van et al\textsuperscript{19} proposed an approach that enabled the entire tweet text to be analyzed and thus helped to reduce the risk of sampling errors. Author developed that, using multiple methods, it could also be extended to other social media content for varied request topics. SVM outperformed methods such as NB, LogR and, when applied to it, provided appropriate and higher level values for all performance parameters, resulting in the highest reliability statistics.

Nair and Shetty\textsuperscript{20} developed a health monitoring application based on spark cluster ML model for the prediction of heart disease using DT methodology to predict an individual's health status by using Twitter real-time data. The software was deployed in "Amazon Elastic Compute Cloud (EC2)" using the cloud.

Xiaomei et al\textsuperscript{21} used five-fold cross-validation reference datasets for SA, including Sanders, Presidential Debate Corpus and Healthcare Reform (HCR). The tweets belonged to different domains: Education, Trump, Republicans, Democrats, Conservatives, Liberals, Voting, Politics, etc. SVM gave better accuracy.

The research was performed in 2017 by Singh et al\textsuperscript{22} where the researcher applied SVM and NB to examine the emotional polarity of health domain tweets. SVM's performance was again observed to be higher than NB.

3. Spark Framework

Apache Spark is a fast and general motor for extensive handling of information on a scale.\textsuperscript{23} Essentially, it is the advance of the development of Hadoop.\textsuperscript{24} Hadoop is the MapReduce demonstrate open source execution and is widely used for conveyed preparation by different servers. It is ideal for cluster-based procedures when all the knowledge has to be encountered. Whatever it may be, its execution drops rapidly for writing other problems (e.g. when iterative or chart-based calculations need to be managed). Spark is a series of numerous tightly organized modules, and conquers Hadoop's problems. It has an execution motor for Directed Acyclic Graph (DAG) that supports cyclic information stream and recording in memory. It can then run programs up to 100x faster in memory than Hadoop, or 10x faster on the network. Begin integrates a heap of libraries that combine the handling of SQL, gushing, machine learning and diagram in a solitary engine. For example, Start offers some abnormal state systems to reserve and makes building circulated applications in Java, Python, Scala and R easy. The applications are converted into employments from MapReduce and continue to run in parallel. In addition to this, Spark can access various sources of information, such as HDFS or Hbase.\textsuperscript{25}

4. Elasticsearch

Elasticsearch is a full-text program written in Java, distributed, that is powerful and freelance from the platform. Combined with restricted skillfulness and quick growth decisions, these options are useful for period huge knowledge analysis.\textsuperscript{26}

Elasticsearch uses a running instance known as a node among the overall functioning of the program which will play one or a lot of functions together with a master or data node. Knowledge set clusters among Elasticsearch need a minimum of one master and one slave node, however a cluster might incorporates one node as a node will assume multiple roles. The sole knowledge
storage format compatible with Elasticsearch is JSON and so, thanks to the unstructured nature of
twitter knowledge, knowledge mapping is needed to supply usable analysis and visualizations.
Supported the JSON format, the framework is a lot of versatile than MySQL and alternative
RDBMS, however less versatile than MongoDB. Whereas a traditional server like RDBMS uses
knowledge storage tables, MongoDB uses BSON (like JSON) format, Associate in Nursing
Elasticsearch uses an inverted index to store knowledge.  
At the instant, Elasticsearch, Hadoop and Spark are the highest three strategies to examine
massive databases.  
Elasticsearch is a distributed search and analytical engine that permits comparatively high speed
knowledge transformations in real time, search queries, document stream process and
classification. Moreover, whereas supporting multiple languages (i.e., Python, Java, and Ruby),
Elasticsearch will index numbers, geographical coordinates, dates and virtually any datatype.  

5. Coronaviruses

Coronaviruses (CoV) are positive-sense, single-stranded RNA viruses that infect animals and
humans. Based on their host specificity, these are categorized into 4 genera: Alpha coronavirus,
Beta coronavirus, Delta coronavirus, and Gammacoronavirus.  
A coronavirus outbreak was recorded in December 2019 from Wuhan, China, which transmitted
from animals to humans. The World Health Organization (WHO) has temporarily termed this
new virus as 2019-novel Coronavirus (2019-nCoV).  
Signs of infection include fever, cough, shortness of breath and difficulty breathing, according to
the WHO. This can lead to pneumonia, SARS, kidney failure and even death in more severe
cases. The period of coronavirus incubation remains unknown. Some sources say this could be 10
to 14 days.

6. Proposed project:

This study is performed several steps. Fig 1 shows the steps of the study. Fig1 is drawn by using
Dia Diagram Editor. It is installed in Ubuntu 16.04. It is an open source graphic design and
editing application and a nice open source Visio alternative for Linux.

7. Implementation of Twitter analysis system

In this paper, we chose tweets as streaming text. We need an authentication framework to get
twitter streaming data. To create services which act on behalf of users’ accounts and make it
really secure and easy to develop, we need three things, which are the Twitter application, REST
API and access to the user account.

We utilized Twitter streaming data from Twitter API. To begin with, we have to be make a new
application in twitter application web site. After that, we are ready to urge authentication secret
key from the site.
In order to get the Twitter feed working, we need four keys; the Consumer Key, Consumer
Secret, Access Token and Access Token Secret. Below are the steps to get those 4 keys:

Go to https://apps.twitter.com/app/new and log in.
Enter the desired Application Name, Description and the website address making sure to enter the
full address including the http://. We can leave the callback URL empty.
Accept the TOS and submit the form by clicking the Create your Twitter Application.
After creating the Twitter Application click on the tab that says Keys and Access Tokens, then we have to give access to the Twitter Account to use this Application. To do this, click the Create my Access Token button.

5. Lastly copy the Consumer key (API key), Consumer Secret, Access Token and Access Token Secret from the screen into our plugin’s Twitter Options page and test.

The experiment is done in Linux environment (Ubuntu 16.04 LTS), hadoop version is 2.7.3, Spark version is 1.5.2 and Scala 2.10.4 version. Scala IDE (IntelliJ IDEA) to develop Scala programming.
We explain the real-time streaming data analysis approach used in detail. First, we need to add some external jar files to develop the model. Our model is based on real-time streaming data. So we need jsonic-1.2.0 and langdetect jar files to be added in lib directory. Also, some dependencies have to be added in the build.sbt file such as Twitter4j to make it possible to load Twitter data and use their function. Now create a scala sbt project and update the build.sbt with dependencies given below:

```scala
libraryDependencies += "org.apache.spark" %% "spark-core" % "1.5.2" % "provided"
libraryDependencies += "org.apache.spark" %% "spark-streaming" % "1.5.2" % "provided"
libraryDependencies += "org.apache.spark" %% "spark-streaming-twitter" % "1.2.0"
libraryDependencies += "org.twitter4j" % "twitter4j-stream" % "3.0.3"
libraryDependencies += "edu.stanford.nlp" % "stanford-corenlp" % "3.5.1"
libraryDependencies += "edu.stanford.nlp" % "stanford-corenlp" % "3.5.1" classifier "models"
libraryDependencies += "org.elasticsearch" % "elasticsearch-spark_2.10" % "2.2.0-m1"
resolvers += "Akka Repository" at "http://repo.akka.io/releases/"
assemblyMergeStrategy in assembly := {
  case PathList("META-INF", xs @ _*) => MergeStrategy.discard
  case x => MergeStrategy.first
}
```

In the program, A DStream is created every 5 seconds. Extract required columns from json object and also generate sentiment score for each tweet. Sentiment score is generated for each tweet using natural language processing library – Stanford core-nlp. Since elasticsearch requires content that can be translated into a document, each RDD is transformed to a Map object before storing it in elasticsearch index twittr/tweet.

Create a pipeline with NLP properties. Run message through the Pipeline. An Annotation is a Map and you can get and use the various analyses individually. For instance, this gets the parse tree of the first sentence in the text. Iterate through tweet. Create a RNN parse tree and Detect Sentiment.

Using Stanford coreNLP – the natural language processing library provided by standford university, parse and detect the sentiment of each tweet.

Stanford coreNLP provides a tool pipeline in terms of annotators using which different linguistic analysis tools may be applied on text. Following annotators are included:

- tokenize: Divides text into a sequence of words
- ssplit: Split the text into sentence. Identify full stop, exclamation etc and split sentences
- **pos**: Reads text and assigns parts of speech to each word such as noun, verb, adjective, etc.
- **lemma**: Groups together forms of a word so they can be analyzed as a single item.
- **parse**: Provides syntactic analysis
- **sentiment**: Provides model for sentiment analysis. Attaches a binary tree of the sentence. The nodes of the tree then contain the annotations from RNNCoreAnnotations indicating the predicted class and scores for that subtree. The sentiment values of the individual words are aggregated at the root of the binary tree.

Sentiment score is then averaged based on length of each sentence as longer sentence must carry more weight in the overall sentiment of the text.

8. Result

Once the implementation is complete, Elasticsearch and Kibana can visually display the results. The contents of the output are transformed into Elasticsearch using the index created before by twitter / tweet.

The Spark calculates sentiment score for each tweet data from the total number of tweets data and is classified as either positive, negative, neutral, not understood and very negative categories based on the score.

![Elasticsearch view of results based on the created index twitter.](image)

The results are shown below with total counts of analytical tweet data, the list of one sentiment-based classified tweets, the quarter-hour analysis of classified tweets and the view of classified tweets, the list of trending hash tags.

This work of analyzing emotions using Spark coreNLP thus provides a clear representation of classified tweet data and more accurate results with less time consuming.
Fig. 3 The total number of analyzed tweets. There are 139,634 tweets.

| Sentiments       | Count |
|------------------|-------|
| negative         | 119,311 |
| neutral          | 17,762 |
| positive         | 1,184  |
| not_understood   | 920    |
| very_negative    | 446    |

Fig. 4 List of Sentiment-based categorized tweets.
In this given tweet data 139,634 tweets are classified into 119,311 of negative, 1,184 of positive, 17,762 of neutral, 920 of not understood and 446 of very negative categories.
Fig. 5 Classified Tweet Created at per 15 minutes Analysis.

Fig. 6 Classified Tweets PIE Chart. It that make the sentiments of classified tweets are easy to understand.

Fig. 7 Pie Chart for Hash tags of patterns. It shows the large number of hash tags for coronavirus keywords.
9. Conclusion

In the Spark framework, we presented a novel sentiment learning technique and a visualization of the results using Elasticsearch and Kibana. The coreNLP natural language processing library in Stanford is very efficient and most useful for processing the text using the NLP functions and it helps us to classify the text based on the scores of sentiments. Because this classification used the Apache Spark faster distributed and parallel computing engine framework, the performance is much better than other works discussed in this paper before. The Elasticsearch and Kibana visualization frameworks are used to extract the data from Spark and are presented with the index generated in different formats. Users can therefore easily understand the results and recognize which opinion (positive, negative and neutral) has received more tweets on the corona subject or domain. It also offers the topic's most trending hashtags. Work is spread across multiple cluster nodes to achieve less time-consuming and more accurate results for large data sets. Spark offers cluster computing for a comfortable and powerful network. Taking advantage of this forum, we successfully develop Scala and spark application infrastructure that analyzes the feelings of the corona tweets, and include visualization in the article. Similar to the positive and neutral tweets, most of the tweets that came in at every run of the request were negative tweets.

10. References

1. Davidov D, Rappoport A. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL 06. 2006;

2. Davidov D, Tsur O, and Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241-249, 2010.

3. Wang X, Wei F, Liu X, Zhou M, Zhang M. Topic sentiment analysis in twitter. Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM 11. 2011;

4. Mei Q, Ling X, Wondra M, Su H, Zhai C. Topic sentiment mixture. Proceedings of the 16th international conference on World Wide Web - WWW 07. 2007;

5. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. In Proceedings of the 6th Symposium on Operating Systems Design and Implementation, pages 137-150, 2004.

6. Banerveld MV, Le-Khac N-A, Kechadi M-T. Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation. Future Data and Security Engineering Lecture Notes in Computer Science. 2014;:29–43.

7. Nair LR, Shetty SD. Applying spark based machine learning model on streaming big data for health status prediction. Computers & Electrical Engineering. 2018;65:393–9.

8. Baltas A, Kanavos A, Tsakalidis AK. An Apache Spark Implementation for Sentiment Analysis on Twitter Data. Algorithmic Aspects of Cloud Computing Lecture Notes in Computer Science. 2017;:15–25.
9. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP 02. 2002;

10. Naveed N, Gottron T, Kunegis J, Alhadi AC. Bad news travel fast. Proceedings of the 3rd International Web Science Conference on - WebSci 11. 2011;

11. Kanavos A, Perikos I, Hatzilygeroudis I, Tsakalidis A. Integrating User’s Emotional Behavior for Community Detection in Social Networks. Proceedings of the 12th International Conference on Web Information Systems and Technologies. 2016;

12. Wang, H Can, D Kazemzadeh, A Bar, F Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: ACL System Demonstrations.2012;; pp. 115–120.
13. Nodarakis, N Sioutas, S Tsakalidis, A Tzimas, G. Large scale sentiment analysis on Twitter with Spark. In: EDBT/ICDT Workshops. 2016;
14. Jain A, Jain V. Sentiment classification of twitter data belonging to renewable energy using machine learning. Journal of Information and Optimization Sciences. 2019;40(2):521–33.

15. Al-Saqqa S, Al-Naymat G, Awajan A. A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Spark. Procedia Computer Science. 2018;141:183–9.

16. Salas-Zárate MDP, Medina-Moreira J, Lagos-Ortiz K, Luna-Aveiga H, Rodríguez-García MÁ, Valencia-García R. Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computational and Mathematical Methods in Medicine. 2017:2017:1–9.

17. Nair, L Shetty, S. Applying spark based machine learning model on streaming big data for health status prediction. Computers & Electrical Engineering. 2017;65, 393–399.

18. Bradlow ET, Gangwar M, Kopalle P, Voleti S. The Role of Big Data and Predictive Analytics in Retailing. Journal of Retailing. 2017;93(1):79–95.

19. Zoonen WV, Toni GVDM. Social media research: The application of supervised machine learning in organizational communication research. Computers in Human Behavior. 2016;63:132–41.

20. Nair LR, Shetty SD, Shetty SD. Applying spark based machine learning model on streaming big data for health status prediction. Computers & Electrical Engineering. 2018;65:393–9.

21. Xiaomei Z, Jing Y, Jianpei Z, Hongyu H. Microblog sentiment analysis with weak dependency connections. Knowledge-Based Systems. 2018;142:170–80.

22. Singh A, Shukla N, Mishra N. Social media data analytics to improve supply chain management in food industries. Transportation Research Part E: Logistics and Transportation Review. 2018;114:398–415.

23. Banerveld MV, Le-Khac N-A, Kechadi M-T. Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation. Future Data and Security Engineering Lecture Notes in Computer Science. 2014;;29–43.
24. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT 05. 2005;

25. Manikanthan S, Padmapriya T. Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g. International Journal of Pure and Applied Mathematics, ISSN NO: 1314-3395, Vol-115, Issue -8, Sep 2017.

26. Gormley C, Tong Z. Elasticsearch: The definitive guide: A distributed period search and analytics engine. Sebastopol: O'Reilly Media, Inc, 2015;

27. Elasticsearch-elastic.co. Retrieved Apr thirty, 2019, from https://www.elastic.co/guide/en/elasticsearch/reference/6.2/index.html.

28. Kumar P, Kumar P, Zaidi N, Rathore VS. Analysis and Comparative Exploration of Elastic Search, MongoDB and Hadoop Big Data Processing. Advances in Intelligent Systems and Computing Soft Computing: Theories and Applications. 2017;:605–15.

29. Cea D, Nin J, Tous R, Torres J, Ayguadé E. Towards the Cloudification of the Social Networks Analytics. Modeling Decisions for Artificial Intelligence Lecture Notes in Computer Science. 2014;:192–203.

30. Snijder EJ, Meer YVD, Zevenhoven-Dobbe J, Onderwater JJM, Meulen JVD, Koerten HK, et al. Ultrastructure and Origin of Membrane Vesicles Associated with the Severe Acute Respiratory Syndrome Coronavirus Replication Complex. Journal of Virology. 2006;80(12):5927–40.

31. Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes & Infections. 2020Jan;9(1):221–36.