A real-time big data sentiment analysis for iraqi tweets using spark streaming

Nashwan Dheyaa Zaki¹, Nada Yousif Hashim², Yasmin Makki Mohialden³, Mostafa Abdulghafoor Mohammed⁴, Tole Sutikno⁵, Ahmed Hussein Ali⁶
¹University of Information Technology and communications, College of Engineering, Iraq
²Computer Science Dep, College of Education, Al-Iraqia University Baghdad, Iraq
³Department of Computer Science, College of Science, Mustansiriya University, Iraq
⁴Faculty of Automatic Control and Computers, University Polytechnic of Bucharest, Romania
⁵Department of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia
⁶Embedded System and Power Electronics Research Group, Yogyakarta, Indonesia

1. INTRODUCTION

The much attention has been given to real-time data processing and big data analytics in the recent era [1, 2]. The increasing volume of information generated daily has made it necessary that organizations must devise ways of handling this information since the existing techniques are not efficient in handling such volume of data created at such a high rate [3-6]. The concept of big data does not only relate to the data volume; it also encompassed data velocity and variety[7]. Data can be either structured or unstructured; it can also be in the forms of a dumped file or in real-time streaming format with high velocity [8-10]. The recent advancements in social network websites have made sentiment analysis and opinion mining a trending research area [11-13]. Twitter is one of the social network applications which is used daily by millions of people to express their opinion. Iraqi dialect stream data has witnessed a tremendous level of growth [14-17] and social media applications remain the commonest source of this volume of stream data. Iraqi users express their opinion and attitude towards issues using different applications [18, 19]. The current opinion of Arabic users can be understood through real-time sentiment analysis of social media streams. Sentiment analysis

Keywords:
Big data
Online processing
Sentiment analysis
Spark streaming
Twitter platform

ABSTRACT

The scale of data streaming in social networks, such as Twitter, is increasing exponentially. Twitter is one of the most important and suitable big data sources for machine learning research in terms of analysis, prediction, extract knowledge, and opinions. People use Twitter platform daily to express their opinion which is a fundamental fact that influence their behaviors. In recent years, the flow of Iraqi dialect has been increased, especially on the Twitter platform. Sentiment analysis for different dialects and opinion mining has become a hot topic in data science researches. In this paper, we will attempt to develop a real-time analytic model for sentiment analysis and opinion mining to Iraqi tweets using spark streaming, also create a dataset for researcher in this field. The Twitter handle Bassam AlRawi is the case study here. The new method is more suitable in the current day machine learning applications and fast online prediction.

This is an open access article under the CC BY-SA license.

Journal homepage: http://beei.org
(SA) refers to the identification of the hidden subjectivity in a text and highlighting the human emotions. During SA, the opinions’ contextual polarity (such as positive, neutral, or negative) is portrayed [14, 20-22]. The current researchers on Iraqi sentiment depend on supervised or lexicon-based methods; however, until now, real-time SA has not been embarked by the existing Iraqi studies.

There are some key challenges before developing any online analytics framework; first is developing reliable and efficient frameworks to distributed data without losing accuracy [23, 24]. Example of this framework is Apache Storm and Apache Kafka [25-27]. The problem of streaming data is that it contains high-velocity information in continues form. Therefore, the process of text analysis in current machine tools will be a bottleneck. All the previous work on sentiment analysis are limited to batch analysis of data. Hence, this study used Spark and decision tree (DT) to propose and develop SA solution for Iraqi dialect. In this study, we want to design an online analytics framework that ingests data from Twitter API with fast processing and ready for predictions at any time [28-30]. The above system is missing in the current Arabic research of SA. This study aims to develop a real-time analytic model for sentiment analysis and opinion mining of Iraqi tweets using Spark streaming (Bassam Al-Rawi is the case study).

This organization of this paper is as follows: a general description of Spark Streaming was presented in section 2 while section 3 presented the concept of resilient distributed dataset (RDD). Section 4 described sentiment analysis (text analytics). Section 5 explained the proposed method in this study while the experiment and results were detailed and discussed in section 6. The last section (section 7) presented the conclusion drawn from the study.

2. SPARK STREAMING

Apache Spark is an open source framework which consists of an engine for programs distribution across machine clusters and a sophisticated model for writing programs [31-33] that has contributed to the Apache Software Foundation, making it possible for a data scientist to access distributed programming. The initialization of the Spark engine is shown in Figure 1.

![Figure 1. Initializing spark](Image)

The first thing a Spark program must do is to create a SparkContext object, which tells Spark how to access a cluster. To create a SparkContext you first need to build a SparkConf object that contains information about your application Spark as shown in Figure 1. Streaming is a component of Spark which facilitates live stream data processing [34, 35]. Instances of data streams are the generated log files by production web servers or message queues which contains status updates posted by a web service user [36, 37]. The API for the manipulation of data streams which closely matches the RDD API of
the Spark Core is provided via Spark Streaming, giving programmers the chance of learning the project and switching between apps which manipulate the stored data in memory, on disk, or data arriving in real-time. Spark Streaming was also developed to provide an equivalent level of throughput, fault tolerance, and scalability as Spark Core [38-43]. Stream processing at a high level is all about the incessant processing of unbounded data streams. However, it is a difficult task to do this in a consistent and fault-tolerant manner [44]. However, there have been improvements in the stream processing engines such as Spark, Heron, Kafka, Flink, and Samza over the past few years which enables the development and operation of complex stream processing apps by businesses [45, 46]. Spark revolves around the concept of a resilient distributed dataset (RDD), which is a fault-tolerant collection of elements that can be operated on in parallel. The number of data pieces in a batch depends on the rate of incoming data and on the batch interval. The concept of DStream at a high level is shown in Figure 2.

![Figure 2. DStream RDD](image)

3. RESILIENT DISTRIBUTED DATASET (RDD)

The concept of RDD is pretty unique in the domain of distributed data processing as they are introduced to address the problems of complexity and efficiency of both interactive and iterative data processing instances [47-50]. Spark 2.0 gives Spark users the leverage of not having to be having a direct interaction with RDD, but it is important to provide them with the robust mental model of the concept of RDD. Figure 3 showed the Spark Executor window. In brief, Spark depends on the RDD concept where both the idea of a large dataset representation in Spark and the idea for working with it are presented. As immutable, fault-tolerant, parallel data structures, RDD allows users to clearly persist intermediate results in memory, optimize data placement via partitioning, and use a set of rich operators to manipulate them [51, 52].

![Figure 3. Spark Executor](image)
As showing in Figure 3 Spark Execute, so to execute jobs, Spark breaks up the processing of RDD operations into tasks, each of which is executed by an executor. Prior to execution, Spark computes the task’s closure. The closure is those variables and methods which must be visible for the executor to perform its computations on the RDD (in this case for each()). This closure is serialized and sent to each executor.

4. SENTIMENT ANALYSIS (TEXT ANALYTICS)

Text analytics refers to the ways of extracting information from a text collection [45, 53]. The patterns and themes in a given dataset can be uncovered using several data processing and analysis algorithms and techniques. The major aim of this process is to make the unstructured text meaningful in order to extract the relationships and contextual meaning [54]. The analysis of peoples’ political opinions on social networks is a perfect instance of sentiment analysis. The recent trend of tweets is shown in Figure 4. Similarly, the analysis of restaurants reviews on Yelp is another instance of SA [55, 56]. Sentiment analysis is typically implemented using Natural Language Processing (NLP) libraries and frameworks, such as OpenNLP and Stanford NLP.

Figure 4. The recent trends of tweets

5. PROPOSED METHOD

The proposed method is to design a fast analysis framework for gathering, processing, prediction, and visualization of Twitter data. There are many advantages of the proposed method that distinguished it from previous sentiment analysis frameworks. The proposed system breaks the challenge of analyzing hundreds of tweets that arrived the system memory per second. The framework provides a solution to the volume of data by HDFS storage of Spark. We provide parallel data gathering nodes and parallel processing nodes for scalable stream data. The challenge of managing the arrived tweets was solved by Kafka. The proposed framework is shown in Figure 5. The tweets arrive from twitter API in batches and are arranged by Kafka into data streams. The data stream is forwarded to Spark engine (Spark Streaming). The new method is based on lexicon-based algorithm using Apache Spark. The case study of the new method is Bassam Al-Rawi Tweets through the hashtag Bassam Al-Rawi on Twitter.

Before data processing through the Spark engine, Spark Streaming will convert the data into RDD form and transfer it to the system memory. The sentiment based on tweets will be categorized into positive, negative, and neutral as Figure 6. If the summations of sentiment of all the words of a tweet are positive, the event will be categorized as positive and placed in the positive bucket in HDFS. The negative sum of sentiment is placed in the negative bucket of HDFS, and the same for neutral. With this process, sentiment of the whole messages can be easily captured and processed and used eventually to derive the live dashboard for monitoring the trends as Figure 7. Apache Kafka was preferred in this study than the original Twitter
streamer in Apache Spark because of the lack of support for many options by the original streamer; hence, we strive to include support for buffering incoming stream for later processing or for when a certain condition is met. It also does not support streaming tweets in certain languages.

Figure 5. The proposed framework

Figure 6. DT of tweets
Figure 7. K-Means clusters of tweets

Figure 7 showing the ability of K-Means algorithm can be used to identify unknown groups in complex and unlabeled data sets. The general parameters that were used in this experiment were as follows: the number of clusters 2, the distance measure is squared Euclidean distance, the average cluster distance is 9.823, and the Davies-Bouldin index is 0.805.

6. EXPERIMENT AND RESULTS

The experiment of our proposed system was based on two implementations. First, we implemented the framework in Weka platform on 30% of Bassam Al-Rawi Twitter dataset. Second, we implemented the same percentage of data in the Spark platform. We also made a comparison in terms of time consumption through the processing stage and ingestion stage. The real data sets collected from Twitter API according to Bassam Al-Rawi Hashtag. The Qatari player of Iraqi origin, Bassam Al-Rawi, raised controversy in the social networking sites after scoring the only goal of Qatar's quarter-final team that led to the exclusion of Iraq from the Asian Cup 2019 in Emirates as Figure 8.

This method ensures that all nearby points are in the same cluster. According to the results of k-means, agglomerative clustering, and other clustering methods, we can claim that the accuracy of EM method is higher than others. However, the computational time is more, especially for the dataset with higher dimension. The future work to be done is to reduce the computational time of this algorithm to make it more suitable for high dimensional datasets. As well as testing it on more clustering problems and comparing its
performance with other clustering methods. In this study we implement different data clustering algorithms for online dataset (Bassam Al-Rawi dataset) acquired from twitter API. We made a comparison of this algorithms intern of accurarcy and computational time. Figure 9 shows the clustering results. Figure 9 (a) and Figure 9 (b) show the clustering result accuracy and the clustering result in term of computational time of EM-clustering, DBSCAN, mean-shift clustering, agglomerative clustering and K-means function from statistical and Machine Learning toolbox of MATLAB, respectively.

![Clustering accuracy](image1)

(a)

![Computational time](image2)

(b)

Figure 9. (a) The clustering result accuracy, (b) The clustering result in term of computational time

7. CONCLUSION

Sentiment analysis is one of the most significant areas of text analysis. This study suggested a distributed approach to real-time sentiment analysis for Iraqi dialect sentiment on Twitter using a lexicon-based algorithm. This framework was proposed to gather, filter, and mine streams of data in three main phases of ingestion, processing, and visualization. All the components of the proposed framework have been tested and discussed based on Bassam Al-Rawi dataset. The significant improvement is the speed of processing tweets and implementation of some machine learning algorithms, such as DT and K-Means clustering based on lexicon algorithm in Weka compared to Spark. Another output of this framework is the presentation of a method to collect Twitter datasets for future studies in data analysis and other machine learning approaches.

REFERENCES


A real-time big data sentiment analysis for iraqi tweets using spark streaming (Nashwan Dheyaa Zaki)