International Journal of Intelligent Information Technologies
Volume 17 • Issue 1 • January-March 2021

Complex Events Processing on Live News Events Using Apache Kafka and Clustering Techniques

Aditya Kamleshbhai Lakkad, Vellore Institute of Technology, India
Rushit Dharmendrabhai Bhadaniya, Vellore Institute of Technology, India
Vraj Nareshkumar Shah, Vellore Institute of Technology, India
Lavanya K., Vellore Institute of Technology, India

ABSTRACT

The explosive growth of news and news content generated worldwide, coupled with the expansion through online media and rapid access to data, has made trouble and screening of news tedious. An expanding need for a model that can reprocess, break down, and order main content to extract interpretable information, explicitly recognizing subjects and content-driven groupings of articles. This paper proposed automated analyzing heterogeneous news through complex event processing (CEP) and machine learning (ML) algorithms. Initially, news content streamed using Apache Kafka, stored in Apache Druid, and further processed by a blend of natural language processing (NLP) and unsupervised machine learning (ML) techniques.

KEYWORDS

Apache Druid, Apache Kafka, Data Streams, DBSACN, DGStream, K-Mean, Macro-Cluster, Micro-Cluster, News, Python

INTRODUCTION

The present world is changing over to the advanced world. Creation of news content is developing at an astounding rate. Many news destinations are there, which gives all reports pretty much all around the globe. Numerous individuals are changing the way they expend news, replacing the conventional physical papers and magazines with their virtual online adaptations (M. I. Rana, 2014). Two major key highlights of online news: Interactivity and immediacy. Interactivity identifies with how individuals will expend the news they are keen on, while immediacy expresses that individuals hope to be educated towards the most recent news with for all intents and no delay. That made the online news industry competitive. In the first place, a progressive online news locales and second, in contrast to physical papers, online news destinations do not have any physical limitation on the measure of data they can place in, accordingly they can distribute. Given that, individuals are ready to invest a restricted time for expending news. News locals expected a successful system to grab individuals' eye and pull in their snaps. Regardless of the supreme significance of news generation and utilization, little is thought about them. Motivation of this paper is to find the most interactive and the most immediate news content out of a vast amount of journalism content over the globe. This paper tries to find interactive news content by groping several news articles concerning news's literature. Most immediate news filtered by using the uniqueness of those articles. Hence, the principal

DOI: 10.4018/IJIIT.2021010103

aim for given research work is to create superior comprehension of the sentiments communicated in headlines, the popularity of news stories and the remarks news things trigger. Computation of above algorithm needs a consistent data processing pipeline, as the quantity and time of data generation are uncertain. Thus, paper intended to design the most reliable automatic data fetching and pre-processing architecture. These efforts depend on utilizing unsupervised clustering analysis as an intent to catch how the news reflects present time. Besides, it draws attention to the unique and interesting news feeds in the current scenario. It additionally looks at how the nature of news changes at each moment. To perform the analysis, this study has taken live news channels from everywhere throughout the world through explicit Application Programming Interface (API). Apache Kafka accepts the news events and spills it. Apache Kafka an open-source stream handling software that functions as a Publish/Subscribe broker. (Carina Andrade, 2019) The reason for utilizing the Apache Kafka stage is that it intends to give a unified, high-throughput, low-latency stage for dealing with continuous information feeds. Newsfeeds published to Kafka as back-to-back events. Apache Druid stores those events and liable for a streaming analytics data store, perfect for powering user-facing data applications. Druid explores events following they happen and to join ongoing outcomes with historical occasions of news. (Rowanda Ahmeda) As the online-offline, method has incorporated effectively with many streams clustering algorithms. The authors have likewise used an online-offline handling structure. During online stage, the Druid dynamically maintains the necessary information of the uninterrupted arriving data records. While in the offline stage, it utilizes a centroid based and density-based clustering calculations for evaluating better insights of events. As online mode brings vital events from the live stream and those filtered events than transferred to the offline mode it improves the performance of the clustering algorithm and accelerates speed, the need for storage memory, and reduce the time complexity of event processing.

The research work contributes towards developing an automated news processing system, which will give a valuable insight from a large amount of data. Initial section describes the background and related research work for the related fields. It gives an already developed key research topic and a building block for this article. In the following section, the primary method and approach is discussed. The method section further divided into two sections implementation and approaches. Implementation shows throughout the data life cycle and approaches show how different algorithms and data analytics techniques utilized to come up with some valuable results. The following section of the article gives some results of the implemented algorithm and tests those algorithms using some hypothesis testing techniques. At last, we have derived the future works and conclusion.

BACKGROUND

Online news has been widely examined in various spaces including various streams of computer engineering. Here it quickly describes those efforts in regards to stream data analysis and machine learning techniques on news items.

(Julio Reis, 2015) Studies the polarity of news headlines and how it differs across topics and changes over time. After that, it looks at how the sentiment of Headline relates to the popularity of the news article. Lastly, characterize the news comments by their sentiments and how it associates with headlines and news articles and conclude with a discussion about the implication of the findings. (M. Tarik Altuncu, 2018)Shows consistent groupings of documents according to content without prior assumptions about the number or type of clusters to be found. The multilevel clustering reveals a quasi-hierarchy of topics and subtopics with increased intelligibility and improved topic coherence as compared to external taxonomy services and standard topic detection methods. (Rowanda Ahmeda) Proposed DGStream, a new online-offline grid, and density-based stream clustering algorithm. The study conducted many experiments and evaluated the performance of DGStream over different simulated databases and for different parameter settings where a wide variety of concept drifts, novelty, evolving data, number and size of clusters and outlier detection are considered. An algorithm