# **Using Twitter Feeds to Analyze People’s Sentiment Towards Popular E-Commerce Sites (Real time)**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Project Team** |  | **Supervisor** |
| 1. Naval Kumar – s11146921 2. Shimneet Chand – s11145829 3. Vishanuel Prasad – s11148140 |  | Kaylash Chaudhary  Course Coordinator |

*Abstract*—the objective of the proposed project is to aid popular E-Commerce sites to enhance their product efficiency by analyzing public sentiment towards their product delivery method and comparing them with their fellow competitors in real time. Twitter is deemed to be one of the most used, liked and informative social media platform in today’s age. At the same time, online shopping from e-commerce sites are becoming more and more successful. Customers would go on Twitter and write all about their experience between them and the popular e-commerce sites (Amazon, Ebay and Aliexpress), such as positive, negative feedbacks and reviews. This will allow the sites to maintain and improve efficiency by working in their weak areas to maintain their place and reputation in the market.

## Introduction

In today’s age, people enjoy to shop in the comfort of their homes and the growing popularity of e-commerce websites provide the medium for them to receive almost any physical product to their doorstep, or even digital content and services. Due to the availability and popularity of a variety of e-commerce websites, people expect to receive the best kind of service from each site and the e-commerce websites intend to deliver the same. However, in order for the product deliverers to deliver an effective product, they need to know what the public sentiments and opinion are towards their product delivery methods and their service. This proposed project intends to use one of the most informative, used and popular social media platforms, Twitter, to analyze public sentiments in real-time for some of the popular e-commerce sites and rank them accordingly. For the purposes of this project, data will be collected and analyzed for three popular e-commerce sites that are Aliexpress, Amazon and Ebay. Analysis of the overall public sentiment in real-time will give the said e-commerce an insight as to where they currently stand in the e-commerce market amongst their competitors and achieve a better understanding of people’s view of their particular products with a high accuracy.

## Methodology

The project was conducted to understand how people view the popular e-commerce websites (Amazon, Ebay and Aliepress). After planning, discussion and research, the team came up with the following steps, which is divided into 4 parts, namely, data gathering, data pre-processing, sentiment analysis and data visualization. These steps are briefly described below.

1. **Data gathering**

In this module:

1. A developer account was created on Twitter to gain access to the Twitter API followed by the creation of a Twitter application with the developer account.
2. Apache Hadoop, HDFS, Apache Hive and Apache Flume was setup and used.
3. To gather real-time Tweets from Twitter, Apache Flume with HDFS sink was used with search words including Aliexpress, Amazon and Ebay and the data stored in Hadoop’s HDFS in JSON format.
4. The keys and tokens created by the Twitter Application for the usage of its API’s was configured in Apache Flume configuration file.
5. A java application was setup for data ingestion from Twitter with apache flume.

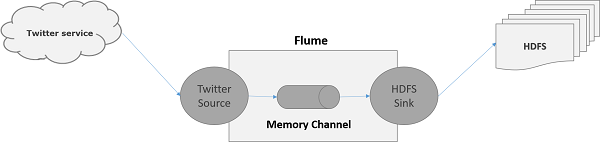


Figure 1.1: Apache Flume architecture for gathering tweets by Tutorialspoint, 2020[6]

1. **Data pre-processing**

After gathering data, the data was processed due to its unstructured nature. By pre-processing the data, we eliminated duplicate tweets, punctuations, stop words, blank spaces and other symbols which might have caused inefficiency and inaccuracy during the sentiment analysis phase and processing. Preprocessing data will also be used to give the Twitter data structure which made it easier for sentiment analysis.

1. As the data stored in HDFS was in JSON unstructured format, a structure was given to the gathered tweets using Apache Hive.
2. By using a Cloudera Json serde, we stored the JSON data to a structured hive table. The created Hive table contained two columns, id(bigint) and text(string).
3. The text in the Hive table was split using Hives split() UDF and stored into another table with two columns, id(bigint) and word(array).
4. Each words were split up into separate rows in a new table with the tweet id using the UDTF called explode.
5. Stop word dictionary was loaded up into a separate Hive table. Stop words were compared with the latest table and stop word was removed. (optional if time permits)
6. **Sentiment analysis**

The processed data was then analyzed to get public opinion from Twitter on the E-commerce website of Aliexpress, Amazon and Ebay. Each public sentiment was categorized in three categories which are neutral, positive and negative. The processed data was analyzed against various types of dictionary to determine if the overall tweet was positive, negative or neutral.

1. Affin dictionary (AFINN is a dictionary which consists of 2500 words rated from +5 to -5 depending on their meaning.) was loaded in a separate table in Hive.
2. The tweet\_word table and dictionary table were joined together so that the rating of the word and the word were together. Whenever there was a match with the word of the tweet in the dictionary, the rating was given to that word else NULL was presented.
3. The ‘groupby’ operation on the tweet\_id was performed so that all the words of one tweet were grouped at a single place.
4. Average operation on the rating of the words for each tweet was performed so that the average rating could be found.
5. **Data Visualization**

The sentiment analysis phase provided the public opinion and view which then was visualized by form of a graph.

1. The id and the rating were queried and pulled from the Hive database.
2. The rating above 0 were treated as positive tweets.
3. The rating below 0 were treated as negative tweets.
4. The 0 rating tweets were treated as neutral.
5. A new table was created in Hive which stored the number of positive, neutral and negative tweets for a given e-commerce site.

A Bar graph was created for the final report to display and visualize the results. The graph automatically refreshed and updated its values as Apache Flume ingested more data into Apache Hive for sentiment analysis.

## Related Works

1. Anisha P. Rodrigues & Niranjan N. Chiplunkar | Marko Robnik-Šikonja (Reviewing editor) (2018) Real-time Twitter data analysis using Hadoop ecosystem, Cogent Engineering, 5:1, DOI: 10.1080/23311916.2018.1534519
   1. Conducted real-time twitter analysis using:
      1. Apache Hadoop
      2. Apache Flume
      3. Apache Pig
      4. Apache Hive
   2. Their research compared the use of Apache Pig and Apache Hive for processing real-time streaming data and proved that Pig provided a much more efficient real-time processing capabilities than Hive as it had a lower execution time in comparison with Hive.
2. Barskar, Anjali & Phulre, Ajay. (2017). Opinion Mining of Twitter Data using Hadoop and Apache Pig. International Journal of Computer Applications. 158. 1-6. 10.5120/ijca2017912854.
   1. Conducted twitter opinion mining (sentiment analysis) using:
      1. Apache Hadoop
      2. Apache Flume
      3. Apache Pig
   2. Their research comprised of conducting twitter analysis with a focus on deep analysis of very large, complex datasets.
   3. Their research proved that Apache Flume was capable of handling very large datasets for fetching tweets.
3. Anusha N, Divya G and Ramya B, 2017. Sentiment Analysis of Twitter Data through Big Data. International Journal of Engineering Research and, V6(06).
   1. Conducted sentiment analysis of tweets regarding a well-known person, Narendra Modi using:
      1. Apache Hadoop
      2. MapReduce
      3. BI (Business Intelligence) tool for visualization of analysis
   2. Their research steps included gathering data, performing real-time sentiment analysis processing, storing the processed data in a database, extracting the sentiments (positive, negative) and finally visualizing the results.
   3. Their research showed that opinions about Narendra Modi were mostly positive (83%).