Research Article

Twitter Sentiment Analysis In Diabetes Domain Using Apache Flume And Hive

Harbhajan Singh[1], Vijay Dhir[2]
[1]Research Scholar, Dept. of CSA, Sant Baba Bhag Singh University, Jalandhar, 144030
[2]Director, R&D, Sant Baba Bhag Singh University, Jalandhar, 144030

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Twitter is a social media platform used by millions of people around the world. People express their feelings by posting tweets related to various topics and products. Sentiment Analysis on these tweets can be performed to analyse their opinion. This paper endeavours to perform Sentiment Analysis by extracting tweets related to diabetes domain from twitter.com via Apache Flume and store them in JSON format. By using Apache Hive, the tweets are transferred from the text file to a table and are analysed by comparing the sentiments expressed in the tweets with the AFINN dictionary. Each individual tweet is scored based on a scale from -5 to +5, where a score having value less than zero indicates a negative sentiment; a zero indicates a neutral sentiment, and a score which has a value greater than zero indicates a positive sentiment. This study can benefit the people suffering from diabetes by making them aware about diet, lifestyle and precautionary measures required to manage their condition in a better way. Also the heath care organizations can utilize the results of this research and improve their strategies to benefit the society.

Keywords: AFINN Dictionary, Apache Flume, Diabetes, Hadoop, Hive, Sentiment Analysis, Twitter

1. Introduction

In recent years, millions of users have started using social media platforms such as Twitter, Facebook and Instagram. People express their opinions using these social media websites and these opinions can be analysed to form future policies by different organizations. This paper endeavours to perform Sentiment Analysis by extracting tweets related to diabetes domain from twitter.com via Apache Flume and store them in JSON format. By using Apache Hive, the tweets are transferred from the text file to a table and are analysed by comparing the sentiments expressed in the tweets with the AFINN dictionary.

1.1 Sentiment Analysis

Sentiments are internal feelings and emotions of a person towards an entity in real world. Sentiment analysis is a process of extracting and stating the opinions from very large files. The major sources include reviews, comments from social networking sites and political judgements made by people. The fields such as natural language processing and data mining handle the process of sentiment analysis by applying various methods and algorithms in sophisticated manner. Sentiment analysis is also referred as opinion mining because data mining is applied to extract and segregate opinions. The sentiments based on the nature of an opinion can express positive, negative or neutral attitude of a person towards an object. The reviews, comments and judgements provided for products and services offered by an organization are valuable assets for future policy formation. Manual reading and extracting useful opinions from large number of reviews and comments seems very tedious and time consuming [1] [2]. The sentiment analysis replaces the process of manual opinion handling with a set of well-defined NLP and data mining algorithmic techniques that work from automatic data collection to presentation of results in efficient and interactive ways. The approach adopted in the analysis is a series of steps to establish the results based on the polarity of opinions provided by contributors. The first step generally specifies identification of sentiments from a given sentence or text. The basic approach stores all or maximum opinion words into a file or database for comparison purpose. The reviews and comments are processed against these database opinion words to determine whether a review or comment contains any opinion or not. Other methods such as dictionary-based and corpus-based use online references such as WordNet to find the polarity of opinion. Polarity states
whether an opinion shows favourable or unfavourable attitude about any object. Polarity can be ranged on a given scale to conclude the overall rating of an object [3] [4].
Dictionary-based methods sometimes appear less effective to categorize the nature of an opinion. For example, the word ‘long’ has different polarity in a given context. “This cell phone has long battery life” states positive opinion whereas “It takes so long to boot up the system” states negative opinion. The nature of same word can be different on the basis of context. The corpus-based methods are proved to solve this chaos at some level [5] [6].

Sentiment analysis process can be performed at document, sentence and aspect level. At document level multiple lines or paragraphs provided by a single opinion holder are grouped under one entity called document. This document is then processed against sentiment analysis procedures to anticipate the negative, positive or neutral attitude of contributor. The next approach which is sentence level works only on a single sentence. A single sentence may contain both subjective and/or objective information. Subjective information may contain negative, positive or no opinion word [7]. However, objective information just contains facts that doesn't play any vital role to access the sentiments. At the sentence level, subjectivity is to be found to judge the polarity of opinions. The above two levels only state negative or positive attitudes toward an entity. Feature of an entity cannot be determined based on the opinions found in the data. The third level that is aspect level helps to find sentiments about given aspect or feature of a target product or service. At the aspect level, firstly the entity is to be recognized and detected followed by the classification of features of the entity. N-gram modelling techniques can be implemented to classify the sentiments [8]. So, with the tremendous growth in field of data sciences, a number of advanced methods are being implemented to perform sentiment analysis to get accurate results [9].

1.2 Diabetes

Diabetes is related to special hormone called insulin. Insulin is required for distributing and consuming glucose in body for correct functioning of all other body parts such kidneys, heart etc. This insulin is produced by an organ called pancreas. The production level of glucose depends on
what and how much one eats in one's diet plan. Whenever pancreas does not work properly due to sickness or other reasons, it does not produce enough insulin which is necessary to consume glucose in blood. At that time, level of glucose in blood raises abnormally. This unprocessed glucose level crosses its prescribed range, is called diabetes.
Some people call it raising blood sugar in simple ways. The classification on types of diabetes is Type 1, Type 2 and Gestational diabetes. Type 1 diabetes is also named as juvenile onset diabetes. This diabetes is mostly found in young people and children. The insulin is not produced by pancreas in appropriate amount and some time it doesn't produce insulin anymore. This condition can occurs due to autoimmune condition of body. The autoimmune sometimes kills the beta cells in pancreas responsible for producing insulin. Lack of insulin in body raises glucose level due to which a person needs doses of insulin in any form on daily basis. In type 2 diabetes, human body becomes insulin resistant. It means body system doesn't behave properly. This is also called adult onset diabetes. The functioning of insulin which takes glucose from blood into our body cells get affected. In this type of diabetes, the major concern is functioning of insulin system than production of insulin in body. This type of diabetes is usually found in middle and upper age group. Healthy diet plan, physical activity and regular intake of insulin may help to stabilize the blood glucose level in Type 2 diabetes. The gestational diabetes is normally found in pregnant women (during pregnancy) and it automatically disappears after the birth of baby. When a person feels some of the symptoms such as weight loss, fatigue and frequent urination that relates to diabetes, then a person may be suggested to check blood sugar level. Deranged values of glucose level in blood confirm the diabetes in the person. A special test called A1C test is usually prescribed to assess average blood sugar level for 2-3 months. The symptoms of both types of diabetes are almost identical like unexpected weight loss with fatigue, intense thirst and hunger, frequent urination, foot infection, skin and eyes problems etc.
Patient must consult endocrinologist for better treatment of either type of diabetes. Patient has to take regular medicines prescribed by specialist. Diabetes may lead to heart-failure, kidney failure and other complications in body if not treated well. The normal range of blood glucose level is 70 to 130. When level increases to 180 or above, then patient comes under the category of
diabetic patient and he needs to take care. Regular checking of blood glucose level is prescribed to prevent any emergency situation [10].

1.3 Precautionary Measures of Type1 and Type 2 Diabetes.

Prevention is better than cure in all types of diseases. The precautionary measures suggested by different health care institutions in their tweets for the patients of both types of diabetes to reduce the symptoms of diabetes are as follows:

- Regular check-up of blood sugar level and taking appropriate steps accordingly.
- Follow healthy diet plan and increase physical activities. Lower the intake of foods that contain sugar directly or indirectly.
- Regular workout is a key to avoid high glucose level in blood.
- Controlling body weight also helps to maintain required blood sugar level.
- Taking enough sleep and avoiding stress can also help to decrease symptoms of diabetes.
- Always control over blood pressure, cholesterol and cognitive issues.

- Try to prevent skin, foot infection.
- Regular check-up of eyesight because diabetes also affects eyes.

1.4 Diet and lifestyle for Type 1 and Type2 Diabetic patients.

Diet plays very important role to maintain prescribed range of glucose level in blood. Diet for Type 1 and Type 2 diabetic persons is as follows:

- Lessen the intake of carbohydrates, fried and saturated fatty acids.
- The healthy diet for diabetic patients includes fresh fruits, whole grains, high fibres and non-starchy vegetables.
- Take omega-3 fatty acids which are found in flax seeds. Omega 3 fatty acids also improve cardiac health. Cardiac health should be maintained in diabetes to avoid stroke.
- It is also advised to eat into intervals rather than eating too much.

Lifestyle of both types of diabetic patients usually includes weight management, regular workout and limiting intake of sugar. Physical activities also improve blood circulation in diabetic patients. Regular exercises also help to maintain health of other body organs such as heart, lungs,
liver and kidneys. A healthy lifestyle also follows regular walking of 30 minutes post meal. It boosts up pancreatic functioning of body.

1.5 Treatment of Type 1 and Type 2 Diabetes.

In type 1 diabetes, immune system attacks the pancreatic beta cells. Due to this, insulin production in body gets stopped. So, the basic treatment of type 1 diabetes needs regular intake of insulin in body. Various methods for inserting insulin in body are injection, insulin pump, insulin spray etc. Due to advancements in medical science, person can go for other treatments such as islets transplantation, stem cells and gene therapy depending on the severity of diabetes. The treatment of type 2 diabetes highly depends on one's lifestyle. As it does not need any particular treatment if patient follows healthy lifestyle. But, in some cases, medication with supplements of insulin is also adopted by some patients [11].

Diabetes is actually a malfunctioning in body that can be controlled but cannot be cured completely. All precautionary measures can be taken as lifetime treatment for diabetes.

2. Proposed Work

In order to perform Sentiment Analysis on a large dataset, we use Apache Flume, which is a powerful tool that can be used to extract tweets from Twitter.com by configuring the twitter.conf file.

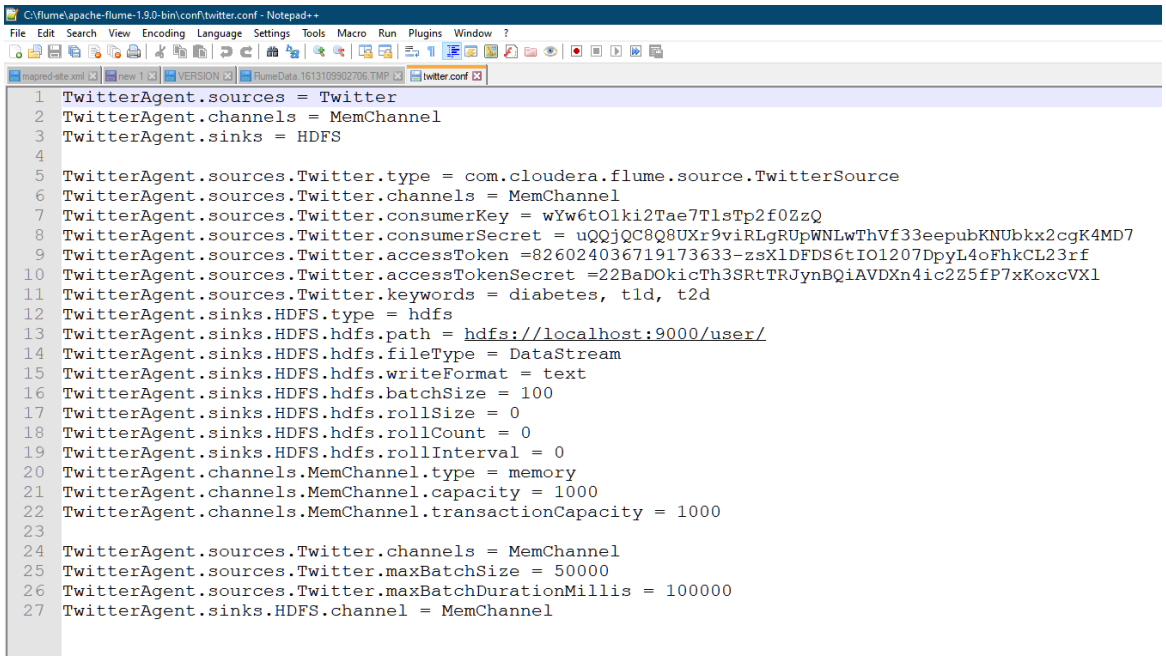


Figure 1: twitter.conf file

We apply for Access token and Consumer key at developer.twitter.com and provide those keys in the configuration file. Apache Hive is used to move the extracted tweets from a text file into a table and then we perform Sentiment Analysis using HiveQL.

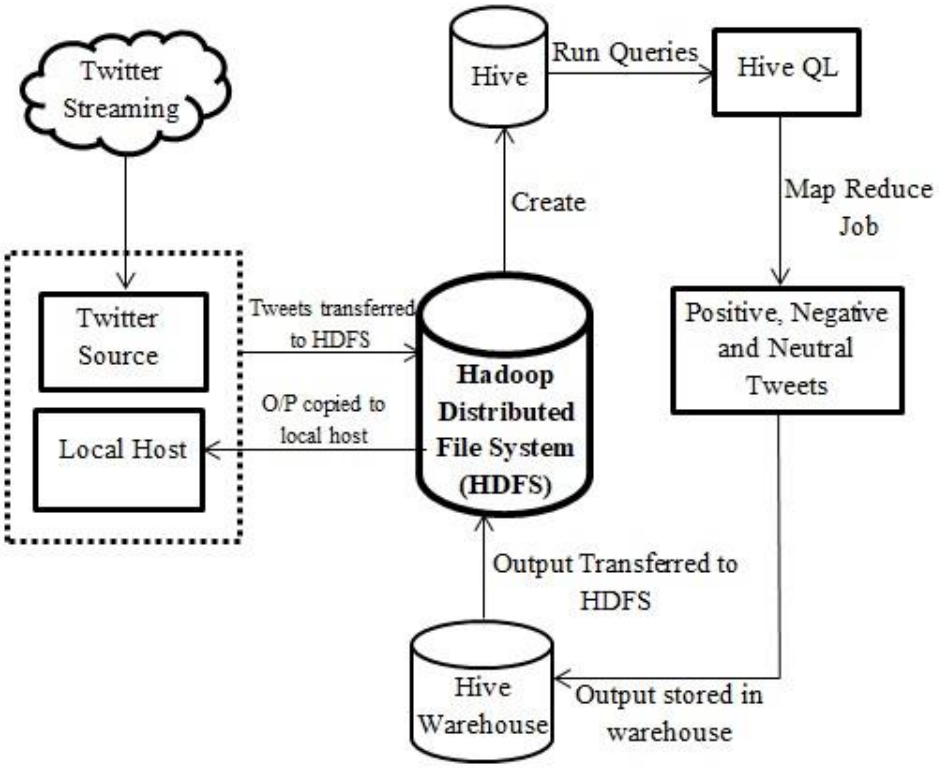


Figure 2: Twitter Sentiment Analysis Architecture

3. Methodology

The described methodology is used to perform Sentiment Analysis:

1. Initially, we turn on **namenode, datanode, yarn node manager** and **resource manager**.
2. For extracting tweets from twitter.com, the keywords **diabetes, t1d, t2d** are used in **twitter.conf** file.
3. Tweets are fetched using Apache Flume which is unstructured data. This format of data is known as JSON format.
4. The file containing tweets is downloaded using Web HDFS and stored in a table using Apache Hive.
5. By comparing the words in the tweets with AFINN dictionary, the score of individual words mentioned in the tweets is calculated based on a scale from -5 to +5.
6. From the score of individual words in a tweet, the final sentiment score of the complete tweet is generated.

Starting the Hadoop Ecosystem:
Hadoop Ecosystem is started using the following command:

After the execution of this command, namenode, datanode, yarn resource manager and yarn nodemanager get started as shown in figure 3.

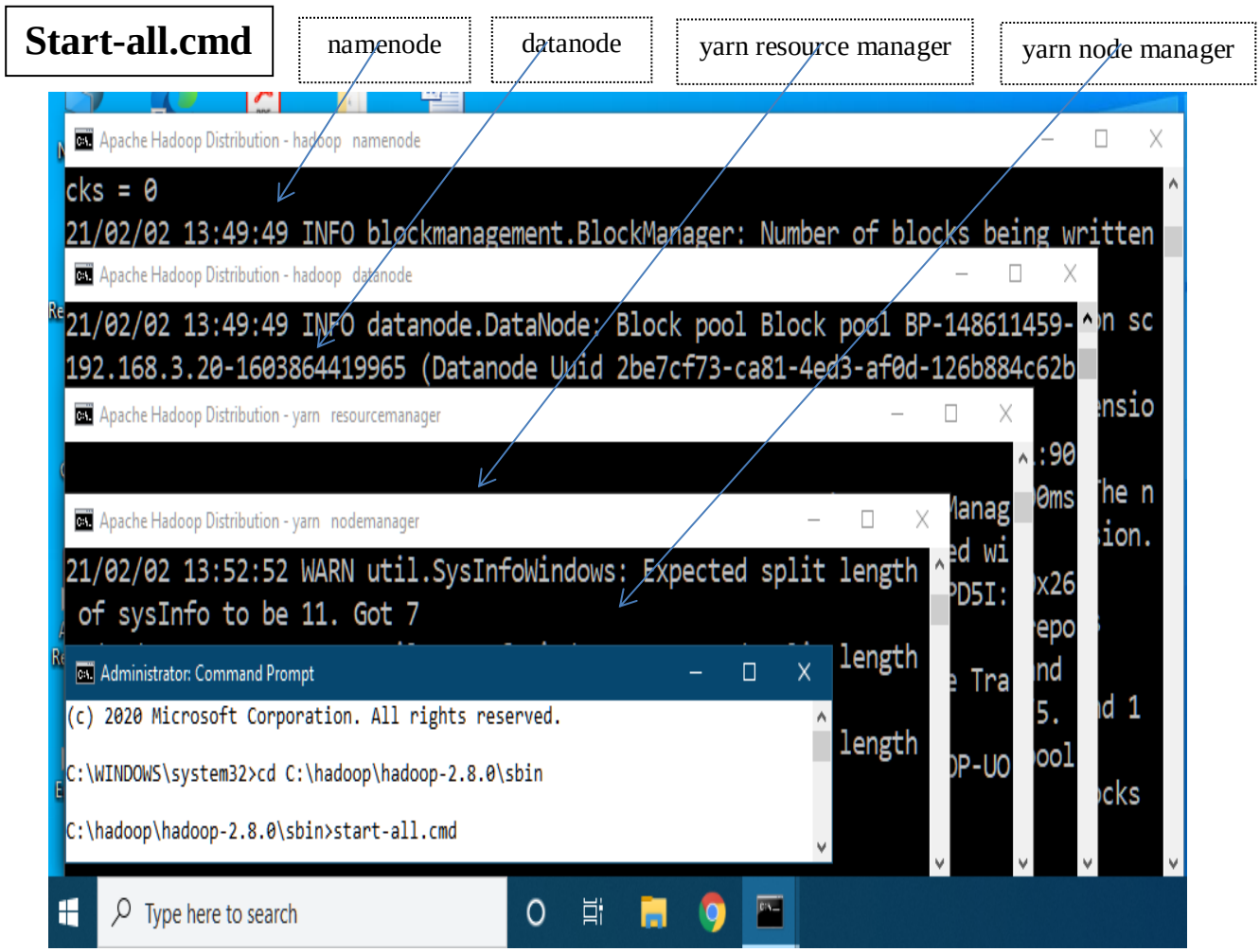


Figure 3: Hadoop Ecosystem

Extracting Twitter Data:
The tweets are extracted from **Twitter.com** using Apache Flume by executing the following command:

Figure 4 shows the extraction of tweets using Apache Flume.

Flume–ng agent –conf ./conf/ -f conf/twitter.conf –property flume.root.logger = DEBUG,console –n TwitterAgent

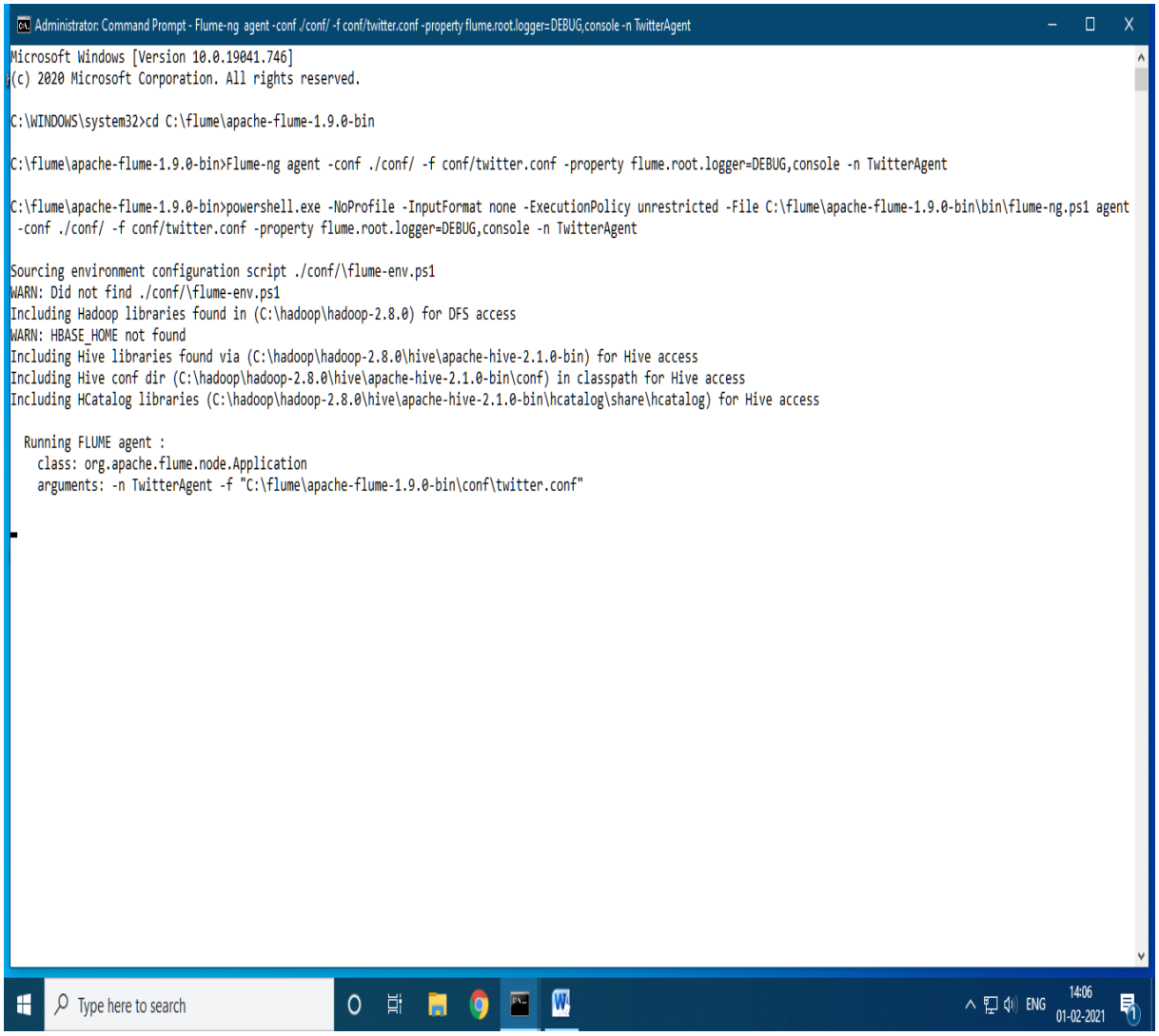


Figure 4: Apache Flume extracting Twitter Data

The tweets are stored in a text file as shown in figure 5 and then this file is moved to HDFS in order to perform Sentiment Analysis using Apache Hive.

Figure 5: File containing tweets downloaded from HDFS

The image shown in figure 6 shows the details of file containing tweets downloaded via Apache Flume.

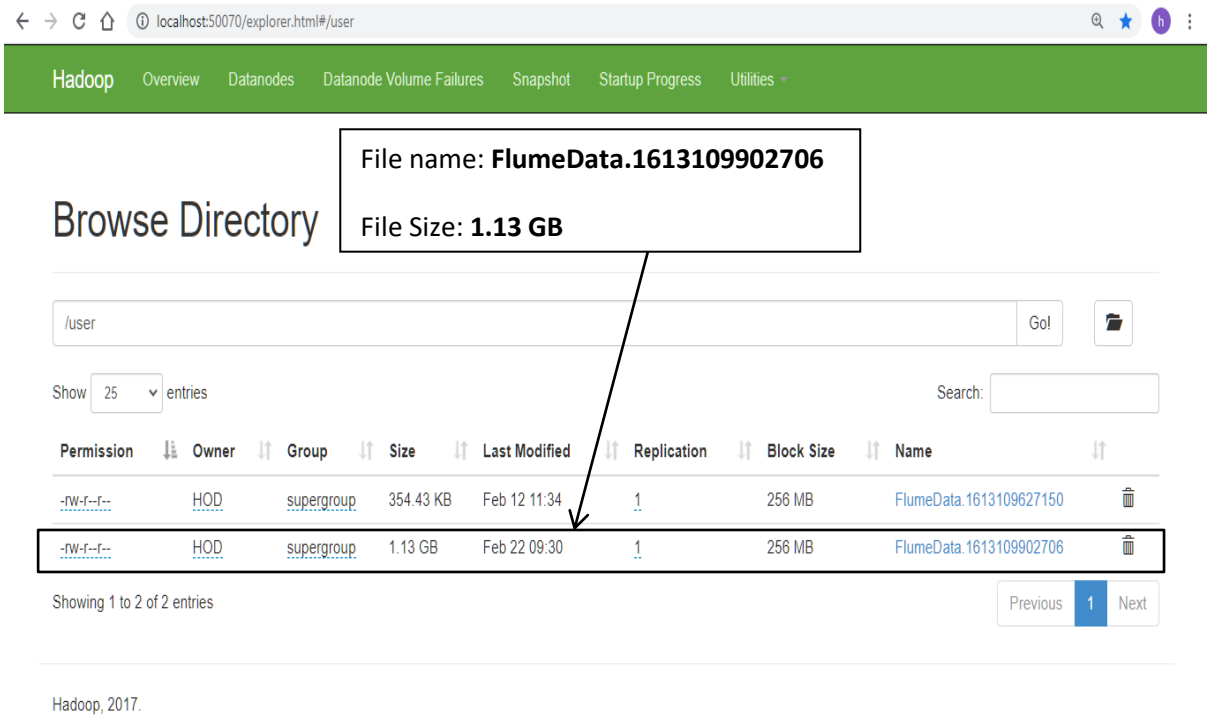


Figure 7 illustrates the properties of the file downloaded from HDFS.

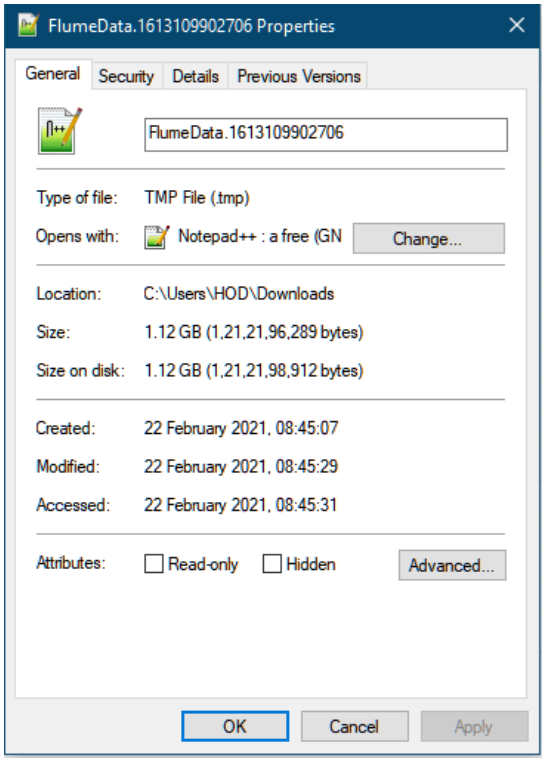


Figure 7: Properties of the file downloaded from HDFS

Creating a table using Apache Hive:
We create an empty table where the tweet Id and text of tweet are stored using the following command as shown in figure 8.

CREATE EXTERNAL TABLE diabetes_data_twitter (id BIGINT, text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user';

```
hive> CREATE EXTERNAL TABLE diabetes_data_twitter (id BIGINT,text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user';
OK
No rows affected (0.237 seconds)
hive>
```

Figure 8: Creating a table to store downloaded file using Hive QL

Moving data into the table:
We move the twitter data into the table **diabetes_data_twitter** using the following command:

LOAD DATA INPATH '/user/flume/tweets/FlumeData.1613109902706' INTO TABLE diabetes_data_twitter;

Displaying the table in Apache Hive:
In order to display the data from the table, the following command is executed:

After the execution of above Hive command, the table showing tweet ID and text is displayed on Hive Terminal as shown in figure 9.

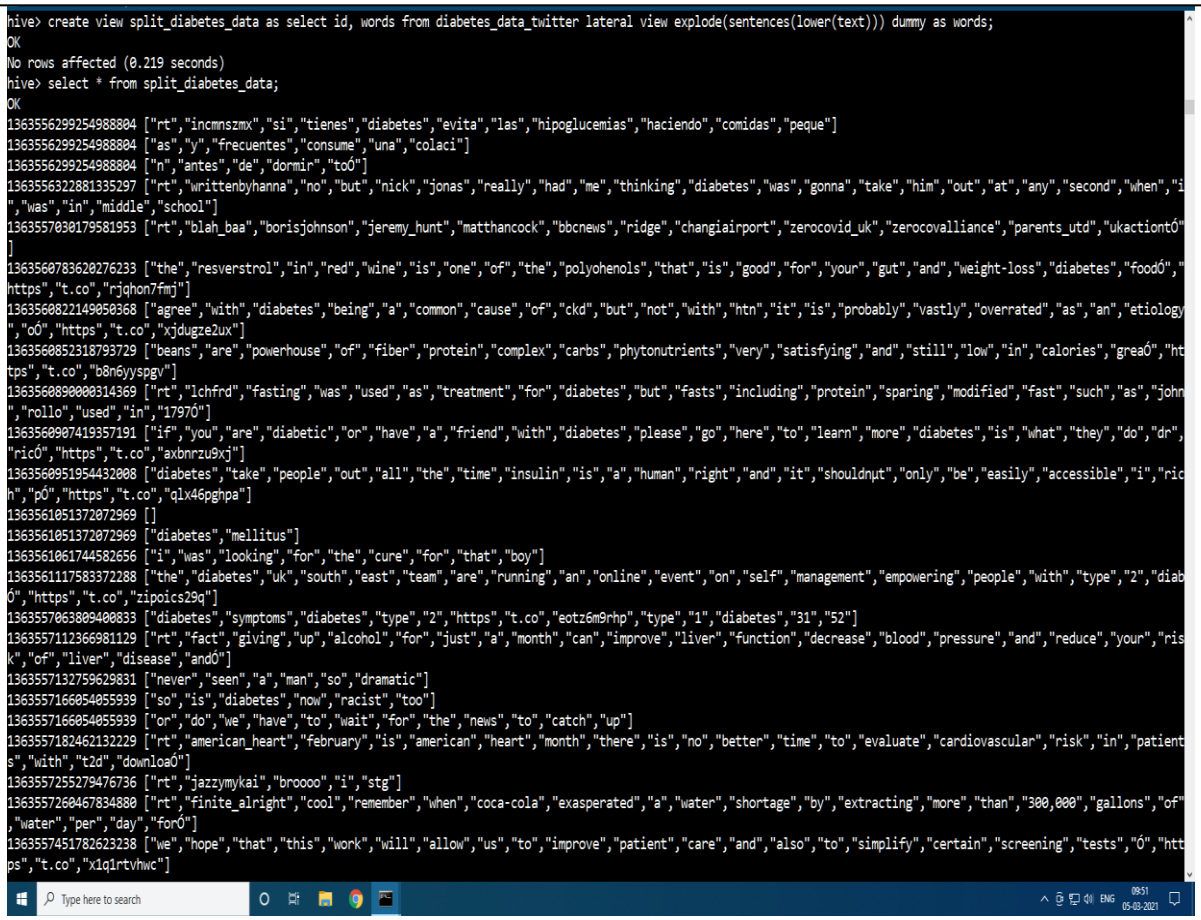
select * from diabetes_data_twitter;

```
Administrator: Command Prompt - hive
hive> select * from diabetes
OK
1363556299254988804  RT @incmnszmx: Si tienes #diabetes evita las hipoglucemias haciendo comidas peque?as y frecuentes, consume una colaci?n antes de dormir. ToÓ
1363556322881335297  RT @WrittenByHanna: No but Nick Jonas really had me thinking diabetes was gonna take him out at any second when I was in middle school ??
1363557030179581953  RT @blah_baa: @BorisJohnson @Jeremy_Hunt @MattHancock @BBCNews #Ridge @ChangiAirport @ZeroCovid_UK @ZeroCovAlliance @Parents_Utd @UKActiontÓ
1363560783620276233  The resverstrol in red wine is one of the polyohenols that is good for your gut and weight-loss#diabetes #foodÓ https://t.co/rJqHoN7fMJ
1363560822149050368  Agree with diabetes being a common cause of CKD but not with HTN...it is probably vastly overrated as an etiology oÓ https://t.co/XjDUGzE2uX
1363560852318793729  Beans are powerhouse of fiber/protein/complex carbs/phytonutrients..very satisfying and still low in calories..greaÓ https://t.co/b8n6YYSpgv
1363557063809400833  #diabetes symptoms diabetes type 2 https://t.co/EoTZ6m9rhP type 1 diabetes 31:52
1363557112366981129  RT @Fact: Giving up alcohol for just a month can improve liver function, decrease blood pressure and reduce your risk of liver disease andÓ
1363557132759629831  never seen a man so dramatic
1363557166054055939  So is diabetes now racist too?  Or do we have to wait for the news to catch up?
1363557182462132229  RT @American_Heart: February is American Heart Month. There is no better time to evaluate Cardiovascular risk in patients with T2D. DownloaÓ
1363557255279476736  RT @Jazzymykai__: Broooo i stg
1363557260467834880  RT @finite_alright: cool. remember when coca-cola exasperated a water shortage by extracting more than 300,000 gallons of water per day forÓ
1363557451782623238  We hope that this work will allow us to improve patient care and also to simplify certain screening tests Ó https://t.co/x1Q1RTVHwc
1363557295456727041  BorisJohnson I would like to see a road map for things such as: Anorexia &amp;amp as well as obesity and diabetesÓ https://t.co/efGlggmtlr
1363556344876462087  Top Selling Diabetes Care:OneTouch, Accu-Chek, FreeStyle and morehttps://t.co/7x7j5nN1Wi #ebaysellerÓ https://t.co/ycGFOjIag8
1363557451782623238  We hope that this work will allow us to improve patient care and also to simplify certain screening tests Ó https://t.co/x1Q1RTVHwc
1363557468991873028  RT @parthaskar: And here is the update from @NHSDigital #GestationalDiabetes #Shielding Thank you from @NHSDiabetesProg to @DHSCgovuk @NiÓ
1363557491347398664  @7VENKMEN omg same here! my grandpa had diabetes, cholesterol and heart problems too?? same here feel free to sendÓ https://t.co/9xPplN1Mx8
1363557612013387776  RT @LEAD_Coalition: Prediabetes linked to worse brain health https://t.co/NYW1de7aoy #diabetes #Alzheimers #dementia @LindaLeeKing @CycliÓ
1363557612013387776  RT @LEAD_Coalition: Prediabetes linked to worse brain health https://t.co/NYW1de7aoy #diabetes #Alzheimers #dementia @LindaLeeKing @CycliÓ
1363557635602198529  Iãd cry my eyes out during his a little bit longer speech ??
1363557643495874563  @irsievan @Rad_Demo Mental psychosis diabetes
1363557666258358275  Not to be an agent of doom but all I see in this video is Diabetes
1363557670876241922  ¶Banting sold the patent rights for insulin to The University of Toronto for $1, claiming that the discovery belongÓ https://t.co/ChmTc5aZVn
1363557708482322434  Cant finish them all in one sitting. Sobra yung sugar baka may diabetes na ko bukas. But I love them all kasi lahatÓ https://t.co/1l1hgub08Y
1363556351725592577  RT @GHargraveWrites: @michelleptweets @WrittenByHanna The fact that she hid her diabetes from Luca in the movie to the point that she almosÓ
1363557761045303297  #diabetes @AmDiabetesAssn @DiabetesUK
1363557791991033858  Hi yãall- keep on telling the hubby to lose a few pounds- doctors say his BMI is dangerously close to being in theÓ https://t.co/onpbMG4zis
1363557805240770560  RT @OlveiraPaco: packaging infames El tipo se ocup? de los programas para pacientes de Diabetes y HIV De la prevenci?n del C?ncer G?nito-MaÓ
1363557843165732877  Another part of a heart healthy lifestyle that can help you manage your diabetes is to focus on what youãre eating.Ó https://t.co/3e1SmhJeFl
1363557871401766912  Nevergetsick World's Greatest Healing Miracle of All Time÷100% Scientifically Proven to Cure Cancer, Diabetes TyÓ https://t.co/mzajYETdqO
1363557880620793862  Diabetes is tough @wedsgp1
1363557914892509186  RT @MarcLobliner: .@CocaCola helps create an obesity and diabetes epidemic in black communities then rallies against being white.Sounds aÓ
1363557930168176640  Poor countries like ours who have richer countries crappy food foisted on us see the highest growth rates of NCDs 1Ó https://t.co/MckoNtvJkK
1363557940972634112  @CancerWarrior8 Oh, I know she is going to start screaming at people in a few hours. The second someone brings up diabetes.
1363557970882289666  Hi Darren, youãre speaking to a young adult who got covid just before the first lockdown and now has diabetes becauÓ https://t.co/Y9noHgp4Lh
1363556434080849927  @robkhenderson All of this coincides with the rise of heart disease, diabetes, and obesity...this sounds like sometÓ https://t.co/sqQnMUjm7U
1363557986812231680  RT @thewayoftheid: Thereãs a character named Suga Feet bc ppl get diabetes every time he hits the floor
1363558012267294721  @WrittenByHanna FR? I was like ¶my baby has diabetes?÷ ??
```

Figure 9: Table showing tweet ID and text

Splitting the tweets into separate words:
In the next step, we split the text into separate words using the following command:

```
create view split_diabetes_data as select id, words from diabetes_data_twitter lateral view
explode(sentences(lower(text))) dummy as words;
```



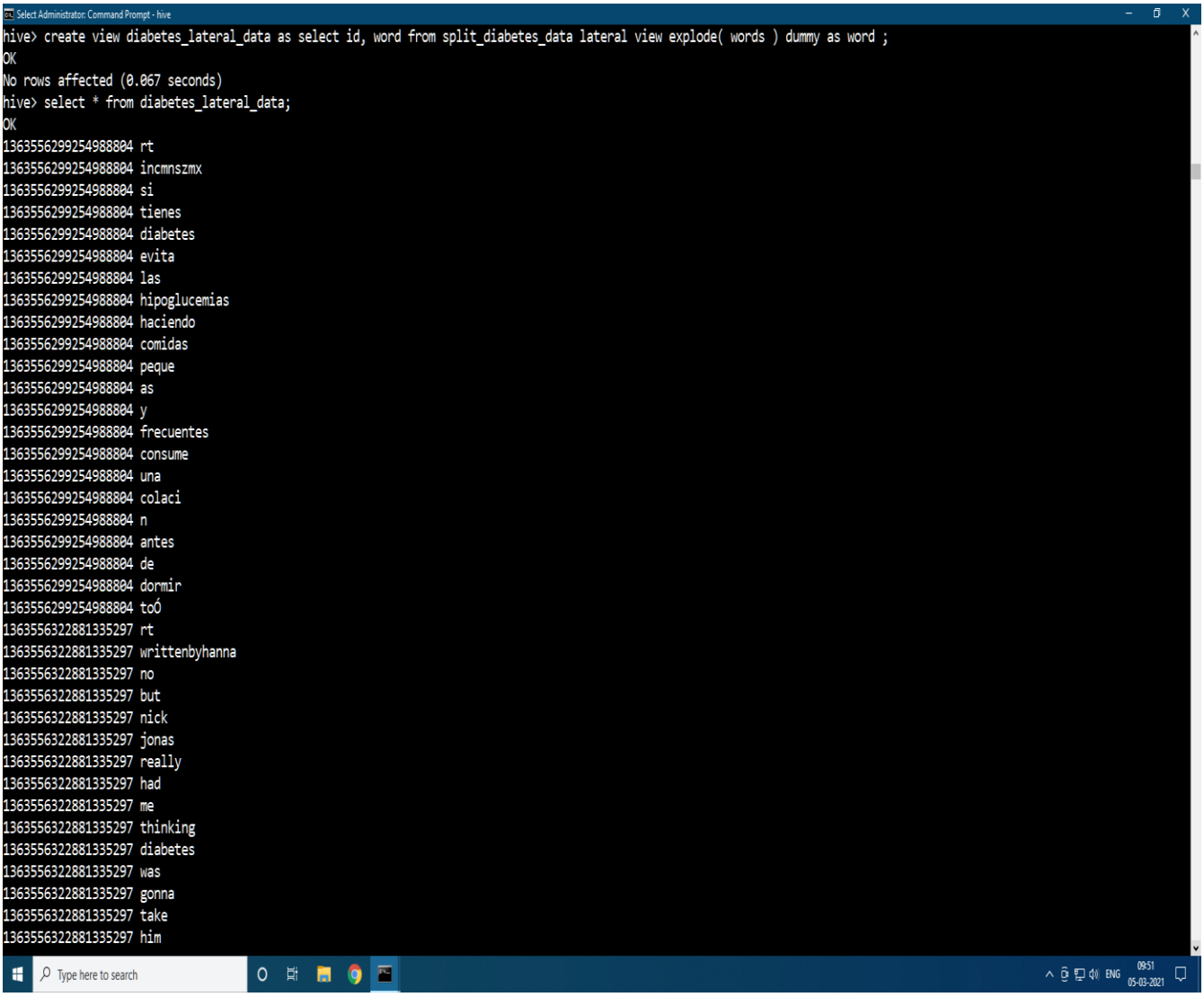
```
hive> create view split_diabetes_data as select id, words from diabetes_data_twitter lateral view explode(sentences(lower(text))) dummy as words;
OK
No rows affected (0.219 seconds)
hive> select * from split_diabetes_data;
OK
1363556299254988804 ["rt","incmnszmx","si","tienes","diabetes","evita","las","hipoglucemias","haciendo","comidas","peque"]
1363556299254988804 ["as","y","frecuentes","consume","una","colaci"]
1363556299254988804 ["n","antes","de","dormir","toÓ"]
1363556322881335297 ["rt","writtenbyhanna","no","but","nick","jonas","really","had","me","thinking","diabetes","was","gonna","take","him","out","at","any","second","when","i
","was","in","middle","school"]
1363557030179581953 ["rt","blah_baa","borisjohnson","jeremy_hunt","matthancock","bbcnews","ridge","changiairport","zerocovid_uk","zerocovalliance","parents_utd","ukactiontÓ"
]
1363560783620276233 ["the","resverstrol","in","red","wine","is","one","of","the","polyohenols","that","is","good","for","your","gut","and","weight-loss","diabetes","foodÓ","
https","t.co","rjqhon7fmj"]
1363560822149050368 ["agree","with","diabetes","being","a","common","cause","of","ckd","but","not","with","htn","it","is","probably","vastly","overrated","as","an","etiology
","oÓ","https","t.co","xjdugze2ux"]
1363560852318793729 ["beans","are","powerhouse","of","fiber","protein","complex","carbs","phytonutrients","very","satisfying","and","still","low","in","calories","greaÓ","ht
tps","t.co","b8n6yyspgv"]
1363560890000314369 ["rt","lchfrd","fasting","was","used","as","treatment","for","diabetes","but","fasts","including","protein","sparing","modified","fast","such","as","john
","rollo","used","in","1797Ó"]
1363560907419357191 ["if","you","are","diabetic","or","have","a","friend","with","diabetes","please","go","here","to","learn","more","diabetes","is","what","they","do","dr",
"ricÓ","https","t.co","axbnrzu9xj"]
1363560951954432008 ["diabetes","take","people","out","all","the","time","insulin","is","a","human","right","and","it","shouldnµt","only","be","easily","accessible","i","ric
h","pÓ","https","t.co","qlx46pghpa"]
1363561051372072969 []
1363561051372072969 ["diabetes","mellitus"]
1363561061744582656 ["i","was","looking","for","the","cure","for","that","boy"]
1363561111583372288 ["the","diabetes","uk","south","east","team","are","running","an","online","event","on","self","management","empowering","people","with","type","2","diab
Ó","https","t.co","zipoics29q"]
1363557063809400833 ["diabetes","symptoms","diabetes","type","2","https","t.co","eotz6m9rhp","type","1","diabetes","31","52"]
1363557112366981129 ["rt","fact","giving","up","alcohol","for","just","a","month","can","improve","liver","function","decrease","blood","pressure","and","reduce","your","ris
k","of","liver","disease","andÓ"]
1363557132759629831 ["never","seen","a","man","so","dramatic"]
1363557166054055939 ["so","is","diabetes","now","racist","too"]
1363557166054055939 ["or","do","we","have","to","wait","for","the","news","to","catch","up"]
1363557182462132229 ["rt","american_heart","february","is","american","heart","month","there","is","no","better","time","to","evaluate","cardiovascular","risk","in","patient
s","with","t2d","downloaÓ"]
1363557255279476736 ["rt","jazzymykai","broooo","i","stg"]
1363557260467834880 ["rt","finite_alright","cool","remember","when","coca-cola","exasperated","a","water","shortage","by","extracting","more","than","300,000","gallons","of"
,"water","per","day","forÓ"]
1363557451782623238 ["we","hope","that","this","work","will","allow","us","to","improve","patient","care","and","also","to","simplify","certain","screening","tests","Ó","htt
ps","t.co","x1q1rtvhwc"]
```

Figure 10: Table showing individual words in the tweets

Lateral view of extracted tweets:
To create the lateral view of the tweets, the following command is executed:

```
create view diabetes_lateral_data as select id, word from split_diabetes_data lateral view explode(
words ) dummy as word ;
```

The lateral view of the tweets is shown in figure 11.



```
hive> create view diabetes_lateral_data as select id, word from split_diabetes_data lateral view explode( words ) dummy as word ;
OK
No rows affected (0.067 seconds)
hive> select * from diabetes_lateral_data;
OK
1363556299254988804 rt
1363556299254988804 incmnszmx
1363556299254988804 si
1363556299254988804 tienes
1363556299254988804 diabetes
1363556299254988804 evita
1363556299254988804 las
1363556299254988804 hipoglucemias
1363556299254988804 haciendo
1363556299254988804 comidas
1363556299254988804 peque
1363556299254988804 as
1363556299254988804 y
1363556299254988804 frecuentes
1363556299254988804 consume
1363556299254988804 una
1363556299254988804 colaci
1363556299254988804 n
1363556299254988804 antes
1363556299254988804 de
1363556299254988804 dormir
1363556299254988804 toÔ
1363556322881335297 rt
1363556322881335297 writtenbyhanna
1363556322881335297 no
1363556322881335297 but
1363556322881335297 nick
1363556322881335297 jonas
1363556322881335297 really
1363556322881335297 had
1363556322881335297 me
1363556322881335297 thinking
1363556322881335297 diabetes
1363556322881335297 was
1363556322881335297 gonna
1363556322881335297 take
1363556322881335297 him
```

Figure 11: Table showing lateral view of individual words in tweets

Loading the AFINN dictionary into HDFS:
Now, we move AFINN Dictionary into HDFS by executing the following Hadoop command:

LOAD DATA INPATH '/user/AFINN.txt' into TABLE dictionary;

In order to view the contents of the dictionary table, we perform the following operation:

select * from dictionary;

The contents of the **dictionary** table are as shown in figure 12.

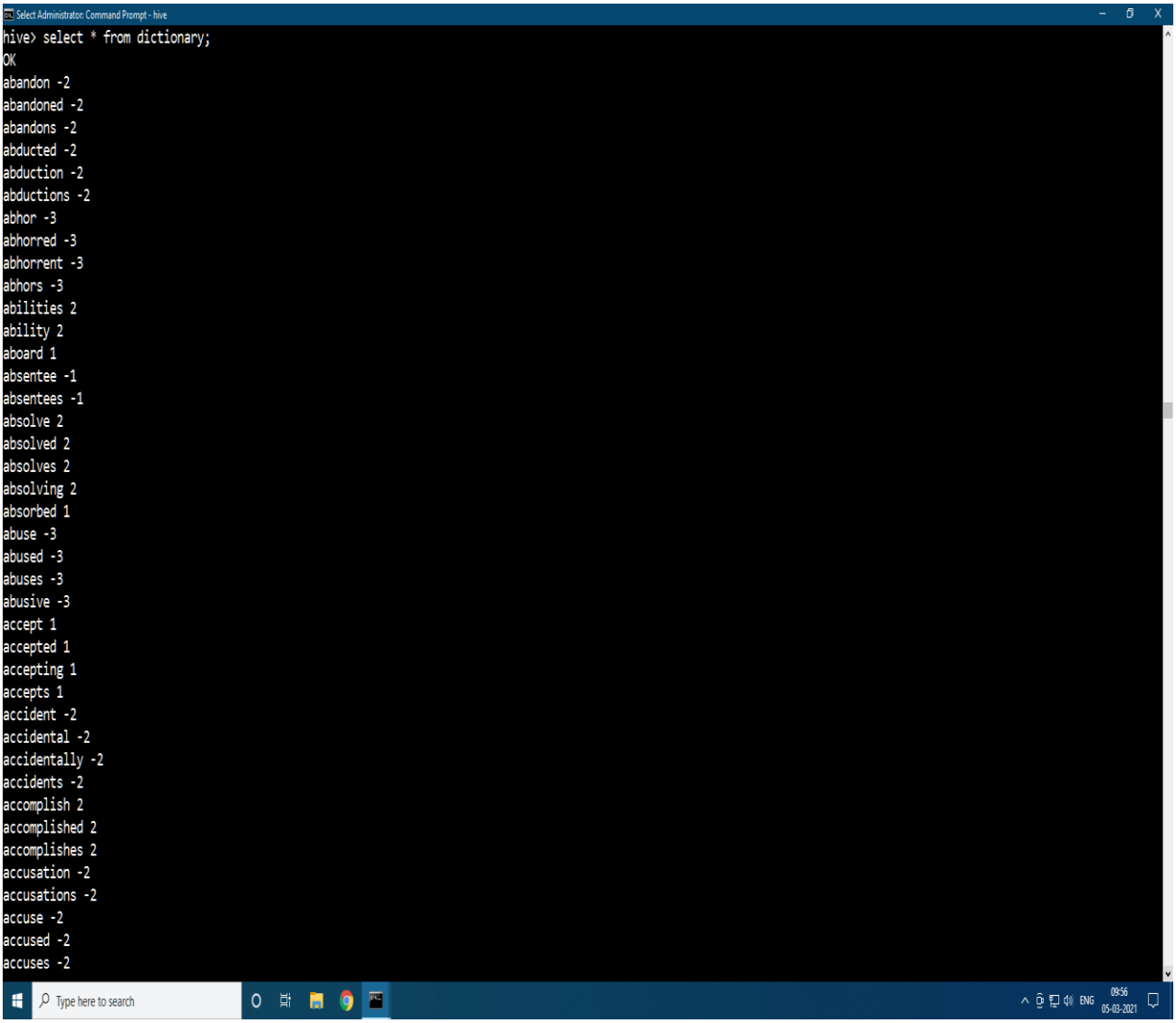
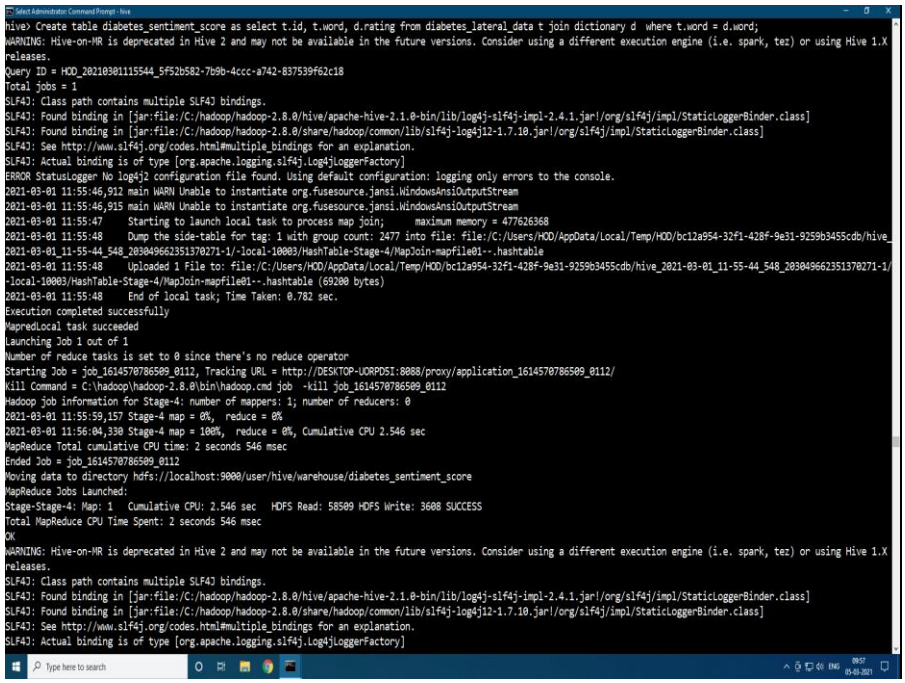


Figure 12: Table showing AFINN dictionary

Generating Sentiment score of individual words:

In the next step, we compare the words in the text with the dictionary in order to find the sentiment score of each word in a tweet. For this operation, the following command is executed:

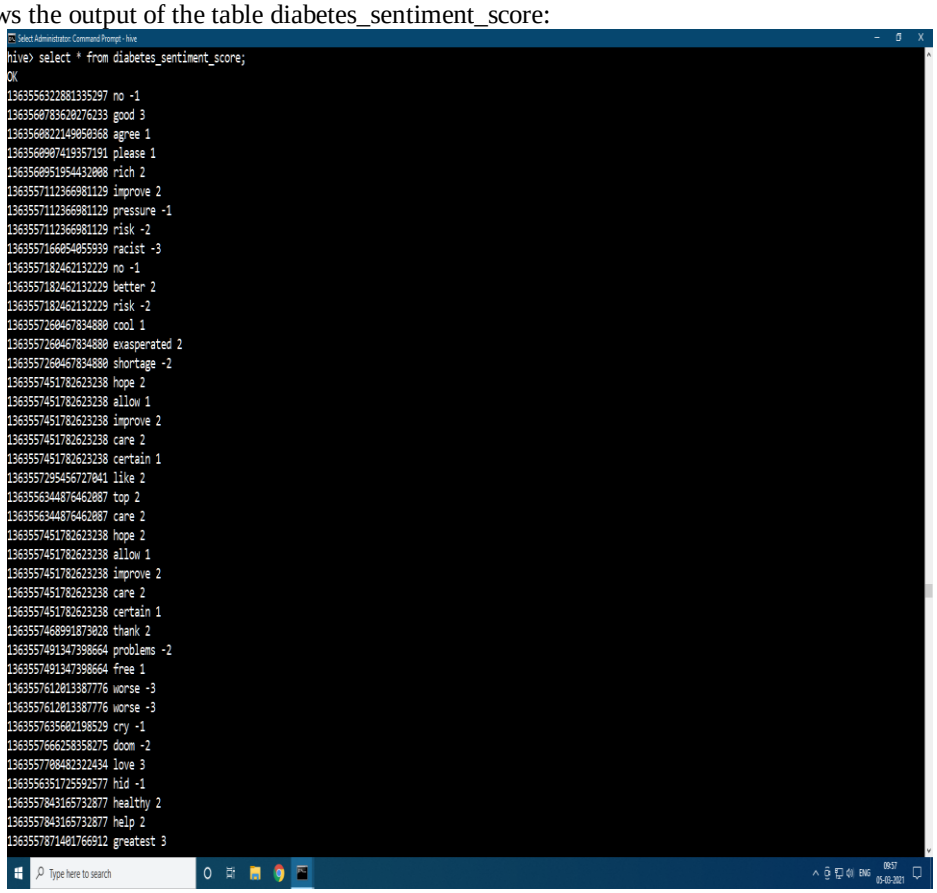
Create table diabetes_sentiment_score as select t.id, t.word, d.rating from diabetes_lateral_data t join dictionary d where t.word = d.word



```
hive> Create table diabetes_sentiment_score as select t.id, t.word, d.rating from diabetes_lateral_data t join dictionary d  where t.word = d.word;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X
releases.
Query ID = HOD_20210301115544_5f52b582-7b9b-4ccc-a742-837539f62c18
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop/hadoop-2.8.0/hive/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop/hadoop-2.8.0/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console.
2021-03-01 11:55:46,912 main WARN Unable to instantiate org.fusesource.jansi.WindowsAnsiOutputStream
2021-03-01 11:55:46,915 main WARN Unable to instantiate org.fusesource.jansi.WindowsAnsiOutputStream
2021-03-01 11:55:47     Starting to launch local task to process map join;      maximum memory = 477626368
2021-03-01 11:55:48     Dump the side-table for tag: 1 with group count: 2477 into file: file:/C:/Users/HOD/AppData/Local/Temp/HOD/bc12a954-32f1-428f-9e31-9259b3455cdb/hive_
2021-03-01_11-55-44_548_2030496623513702 71-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2021-03-01 11:55:48     Uploaded 1 File to: file:/C:/Users/HOD/AppData/Local/Temp/HOD/bc12a954-32f1-428f-9e31-9259b3455cdb/hive_2021-03-01_11-55-44_548_2030496623513702 71-1/
-local-10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable (69200 bytes)
2021-03-01 11:55:48     End of local task; Time Taken: 0.782 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1614570786509_0112, Tracking URL = http://DESKTOP-UORPDSI:8088/proxy/application_1614570786509_0112/
Kill Command = C:\hadoop\hadoop-2.8.0\bin\hadoop.cmd job  -kill job_1614570786509_0112
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2021-03-01 11:55:59,157 Stage-4 map = 0%,  reduce = 0%
2021-03-01 11:56:04,330 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 2.546 sec
MapReduce Total cumulative CPU time: 2 seconds 546 msec
Ended Job = job_1614570786509_0112
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/diabetes_sentiment_score
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1   Cumulative CPU: 2.546 sec   HDFS Read: 58509 HDFS Write: 3608 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 546 msec
OK
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X
releases.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop/hadoop-2.8.0/hive/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop/hadoop-2.8.0/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
```

Figure 13: Map Reduce processing to generate sentiment score of individual words

Figure 14 shows the output of the table diabetes_sentiment_score:



```
hive> select * from diabetes_sentiment_score;
OK
1363556322881335297 no -1
1363560783620276233 good 3
1363560822149050368 agree 1
1363560907419357191 please 1
1363560951954432008 rich 2
1363557112366981129 improve 2
1363557112366981129 pressure -1
1363557112366981129 risk -2
1363557166054055939 racist -3
1363557182462132229 no -1
1363557182462132229 better 2
1363557182462132229 risk -2
1363557260467834880 cool 1
1363557260467834880 exasperated 2
1363557260467834880 shortage -2
1363557451782623238 hope 2
1363557451782623238 allow 1
1363557451782623238 improve 2
1363557451782623238 care 2
1363557451782623238 certain 1
1363557295456727041 like 2
1363556344876462087 top 2
1363556344876462087 care 2
1363557451782623238 hope 2
1363557451782623238 allow 1
1363557451782623238 improve 2
1363557451782623238 care 2
1363557451782623238 certain 1
1363557468991873028 thank 2
1363557491347398664 problems -2
1363557491347398664 free 1
1363557612013387776 worse -3
1363557612013387776 worse -3
1363557635602198529 cry -1
1363557666258358275 doom -2
1363557708482322434 love 3
1363556351725592577 hid -1
1363557843165732877 healthy 2
1363557843165732877 help 2
1363557871401766912 greatest 3
```

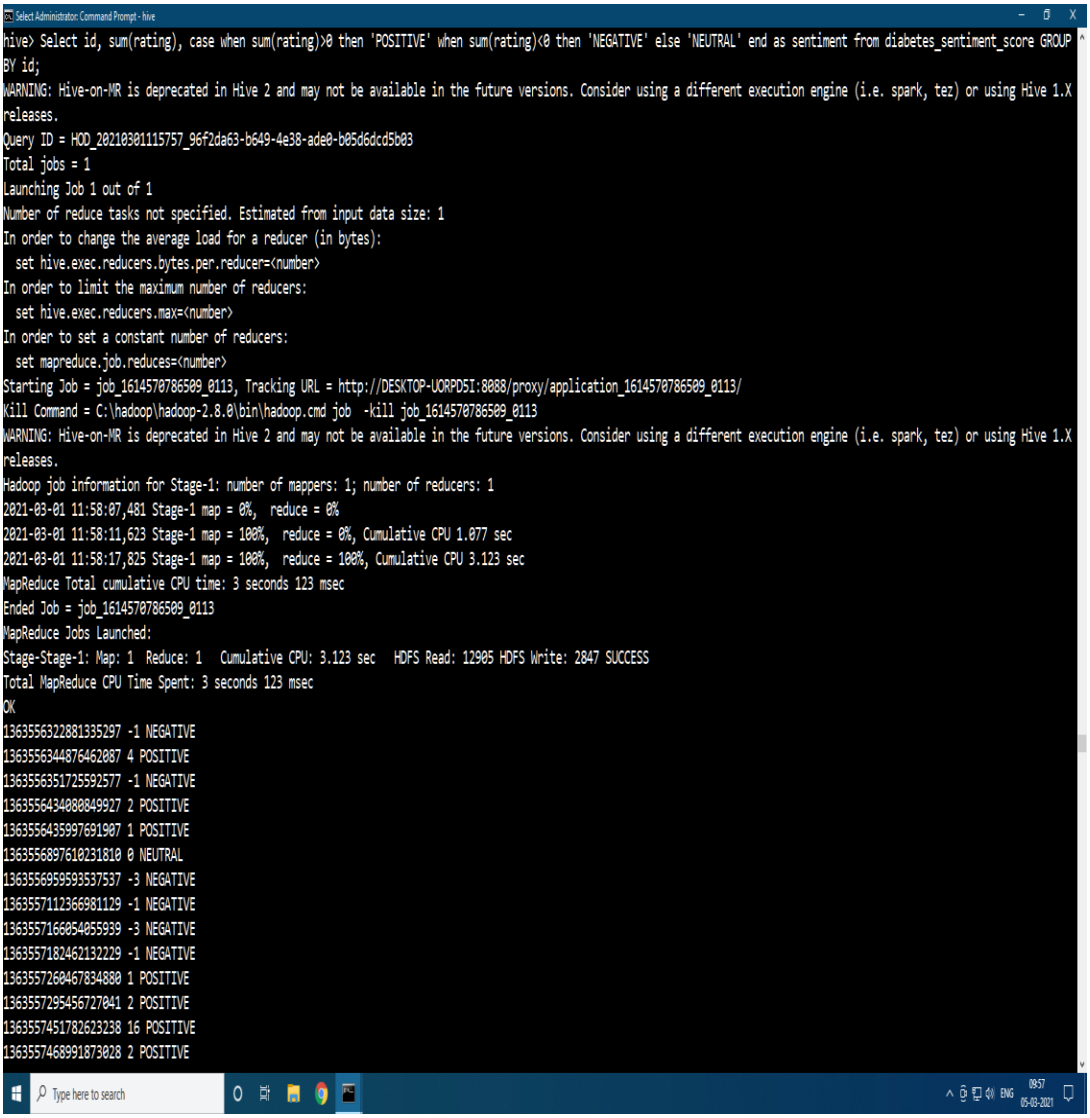
Figure 14: Table showing sentiment score of each word

Sentiment Analysis of extracted tweets:

Now, the final score of all the tweets is generated using the following command:

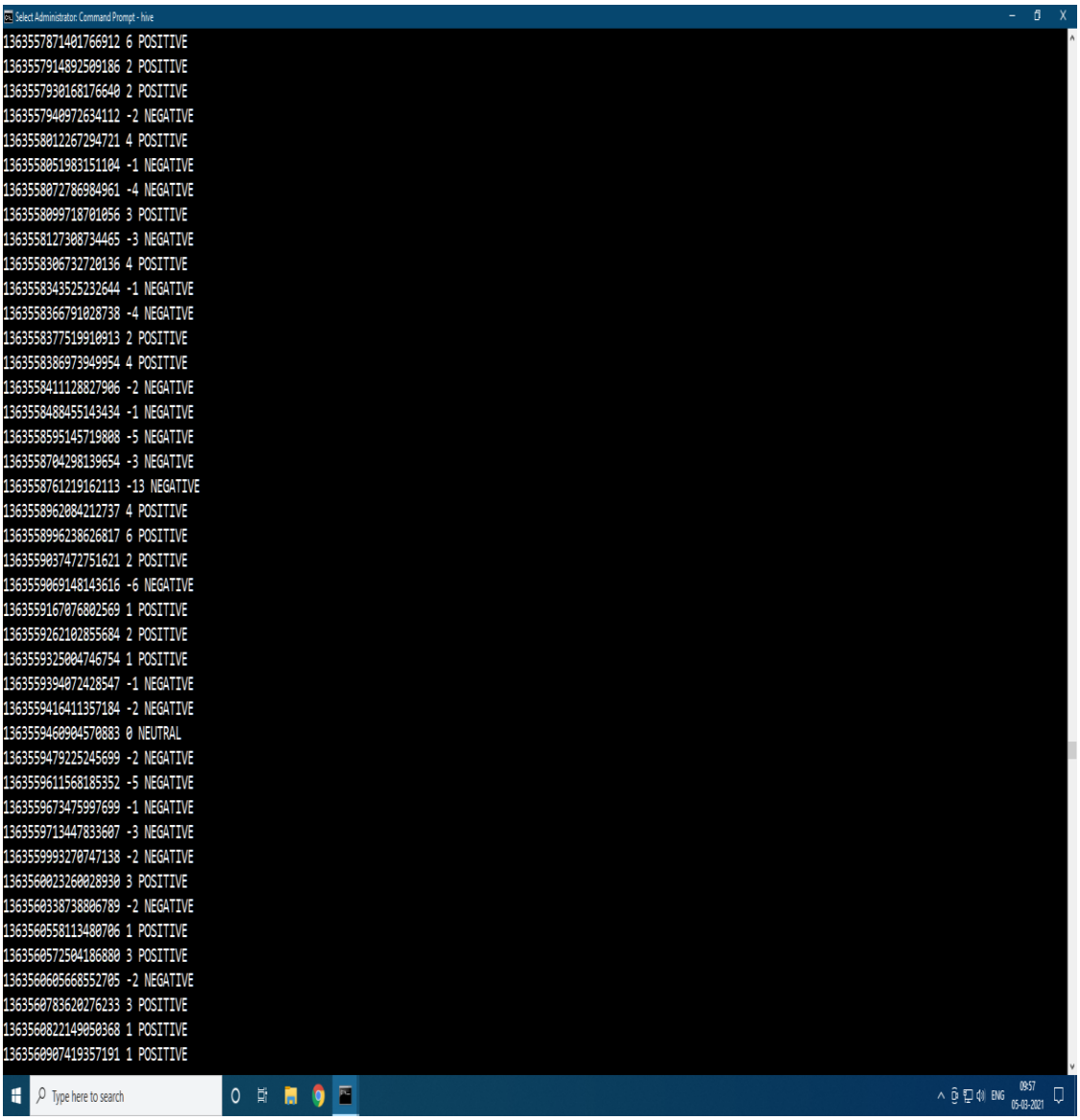
Select id, sum(rating), case when sum(rating)>0 then 'POSITIVE' when sum(rating)<0 then 'NEGATIVE' else 'NEUTRAL' end as sentiment from diabetes_sentiment_score GROUP BY id;

The processing of the above command and sentiment score of each tweet is shown in figure 15.



```
hive> Select id, sum(rating), case when sum(rating)>0 then 'POSITIVE' when sum(rating)<0 then 'NEGATIVE' else 'NEUTRAL' end as sentiment from diabetes_sentiment_score GROUP
BY id;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X
releases.
Query ID = HOD_20210301115757_96f2da63-b649-4e38-ade0-b05d6dcd5b03
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1614570786509_0113, Tracking URL = http://DESKTOP-UORPD5I:8088/proxy/application_1614570786509_0113/
Kill Command = C:\hadoop\hadoop-2.8.0\bin\hadoop.cmd job  -kill job_1614570786509_0113
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X
releases.
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-03-01 11:58:07,481 Stage-1 map = 0%,  reduce = 0%
2021-03-01 11:58:11,623 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.077 sec
2021-03-01 11:58:17,825 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.123 sec
MapReduce Total cumulative CPU time: 3 seconds 123 msec
Ended Job = job_1614570786509_0113
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.123 sec   HDFS Read: 12905 HDFS Write: 2847 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 123 msec
OK
1363556322881335297 -1 NEGATIVE
1363556344876462087 4 POSITIVE
1363556351725592577 -1 NEGATIVE
1363556434080849927 2 POSITIVE
1363556435997691907 1 POSITIVE
1363556897610231810 0 NEUTRAL
1363556959593537537 -3 NEGATIVE
1363557112366981129 -1 NEGATIVE
1363557166054055939 -3 NEGATIVE
1363557182462132229 -1 NEGATIVE
1363557260467834880 1 POSITIVE
1363557295456727041 2 POSITIVE
1363557451782623238 16 POSITIVE
1363557468991873028 2 POSITIVE
```

Figure 15: Map Reduce job generating sentiment score of each tweet



```
1363557871401766912 6 POSITIVE
1363557914892509186 2 POSITIVE
1363557930168176640 2 POSITIVE
1363557940972634112 -2 NEGATIVE
1363558012267294721 4 POSITIVE
1363558051983151104 -1 NEGATIVE
1363558072786984961 -4 NEGATIVE
1363558099718701056 3 POSITIVE
1363558127308734465 -3 NEGATIVE
1363558306732720136 4 POSITIVE
1363558343525232644 -1 NEGATIVE
1363558366791028738 -4 NEGATIVE
1363558377519910913 2 POSITIVE
1363558386973949954 4 POSITIVE
1363558411128827906 -2 NEGATIVE
1363558488455143434 -1 NEGATIVE
1363558595145719808 -5 NEGATIVE
1363558704298139654 -3 NEGATIVE
1363558761219162113 -13 NEGATIVE
1363558962084212737 4 POSITIVE
1363558996238626817 6 POSITIVE
1363559037472751621 2 POSITIVE
1363559069148143616 -6 NEGATIVE
1363559167076802569 1 POSITIVE
1363559262102855684 2 POSITIVE
1363559325004746754 1 POSITIVE
1363559394072428547 -1 NEGATIVE
1363559416411357184 -2 NEGATIVE
1363559460904570883 0 NEUTRAL
1363559479225245699 -2 NEGATIVE
1363559611568185352 -5 NEGATIVE
1363559673475997699 -1 NEGATIVE
1363559713447833607 -3 NEGATIVE
1363559993270747138 -2 NEGATIVE
1363560023260028930 3 POSITIVE
1363560338738806789 -2 NEGATIVE
1363560558113480706 1 POSITIVE
1363560572504186880 3 POSITIVE
1363560605668552705 -2 NEGATIVE
1363560783620276233 3 POSITIVE
1363560822149050368 1 POSITIVE
1363560907419357191 1 POSITIVE
```

Figure 16: Table showing the sentiment score of each tweet

4. Conclusion and Future Scope

In the present study, the sentiments in the tweets related to diabetes domain were analysed using Hadoop Ecosystem. The tweets were compared with the AFINN dictionary and Sentiment Analysis was performed. In the whole process it was found that a large data set of the tweets can be classified, categorized and scaled with assistance of Apache Flume and Apache Hive. However the research was limited to the categorization and scaling of tweets based on their polarity related to diabetes. The comparative study of tweets pertaining to type 1 and type 2 diabetes will be pursued in the future. The relevance of the present study lies in its findings as the recognition of sentiments associated with the disease can be used to change the mind-set of the people suffering from diabetes and their families as well as improve public health strategies.

References

1. Rodrigues, A. P., & Chiplunkar, N. N. (2019). A new big data approach for topic classification and sentiment analysis of Twitter data. Evolutionary Intelligence, 1-11.
2. Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Analyzing social media through big data using infosphere biginsights and apache flume. Procedia computer science, 113 , 280-285.

3. Mishra, R. K., Lata, S., & Kumari, S. (2020). Twitter Sentimental Analytics Using Hive and Flume. In International Conference on Intelligent Computing and Smart Communication 2019 (pp. 159-165). Springer, Singapore.
4. ReddyP, Y. S., & PadmaP, M. Sentiment Analysis of Twitter by using Apache Flume.
5. Rao, N. P., Srinivas, S. N., & Prashanth, C. M. (2015). Real time opinion mining of twitter data. Int J Comput Sci Inf Technol , 6 (3), 2923-2927.
6. Vissamsetti, M. M., Prasanth, Y., & Jacob, T. P. (2020). Twitter Data Analysis for Live Streaming by Using Flume Technology (No. 2915). EasyChair.
7. Kumari, S., Sen, B., Lata, S., & Mishra, M. R. Twitter Sentimental Analytics using Hive and Flume.
8. Karthika, I., Gokulraj, P., & Saravanan, S. (2016). Prediction of sales using Big data analytics. Journal Of Advances In Chemistry, 12 (20).
9. Deva, R., & Kulshreshtha, G. Social Media based Sentimental Analysis using Hive and Flume.
10. Zierath, J. R. (2019). Major advances and discoveries in diabetes-2019 in review. Current diabetes reports, 19(11), 1-9.
11. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., ... & Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The lancet Diabetes & endocrinology, 6(5), 361-369.