Twitter Sentiment Analysis In Diabetes Domain Using Apache Flume And Hive

Harbhajan Singh¹, Vijay Dhir²
¹Research Scholar, Dept. of CSA, Sant Baba Bhag Singh University, Jalandhar, 144030
²Director, 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 health 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.

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.
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';
```

![Figure 8: Creating a table to store downloaded file using Hive QL](image)

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:

```
select * from diabetes_data_twitter;
```

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

![Figure 9: Table showing tweet ID and text](image)
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;
```

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.
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.
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
```
Sentiment Analysis of extracted tweets:
Now, the final score of all the tweets is generated using the following command:

```sql
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.
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
9. Deva, R., & Kulshreshtha, G. Social Media based Sentimental Analysis using Hive and Flume.