Comparative Study of Sentiment Analysis of Type 1 and Type 2 Diabetic Patients Using Apache Flume and Hive

Harbhajan Singh
Research Scholar, Department of CSA, Sant Baba Bhag Singh University, Jalandhar, India.
E-mail: hsrandhawakca95@gmail.com

Vijay Dhir
Director, R&D, Sant Baba Bhag Singh University, Jalandhar, India.
E-mail: research.sbbsu@gmail.com

Received September 22, 2021; Accepted December 18, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19266

Abstract

The Hadoop ecosystem platform has been used to perform sentiment analysis of the opinion regarding type 1 and type 2 diabetes. This research work concentrates on exploring the public opinion about assumptions, thinking and behavior towards the condition of type 1 and type 2 diabetes. Tweets from twitter have been taken as dataset for sentiment analysis for the research. After conducting a set of repetitive experiments, it has been concluded that people feel more negative about type 2 diabetes. They are more pessimistic about the diet and treatment of type 2 diabetes. Present paper is divided mainly into four sections. The first section, gives an overview of the sentiment analysis and diabetes. The proposed work and methodology have been included in the second section. The third section contains the whole experiment work to support the results. In the fourth section, results regarding opinions about type 1 and type 2 diabetes have been illustrated using bar charts obtained from the experiments followed by concluding remarks, future scope and significance of the research.

Keywords

AFINN Dictionary, Apache Flume, Diabetes, Hadoop, Hive, Sentiment Analysis, Twitter.

Introduction

Sentiment Analysis

Sentiments highlight insight, thinking, feeling and emotions of people towards an entity, an activity and any other situation in real life. Sentiment analysis consists of techniques to extract and specify the opinions and thinking from a set of source files. The key sources are tweets, chat messages, blogs, reviews, comments, judgments, conclusions etc. We
basically use tweets as a source for drawing positive, negative and neutral attitude of a person towards an object. An important application of sentiment analysis is to provide a decision support system for an organization about their products and services. Reading and extracting useful information and opinions from a large set of source files manually seems very tedious and time-consuming. The sentiment analysis provides set of well-established natural language processing and data mining algorithmic approaches to automate the whole process of sentiments finding and providing desired output in efficient ways. Sentiment analysis can be performed at different levels that include document, sentence and aspect level. A document level approach is adopted when Multiple lines or paragraphs are provided from a file or document. The second level takes only single sentence as source and works upon it to find any opinion and its nature. In the third level that is the aspect level, the entity is to be recognized and followed by the classification of features of that entity.

Diabetes

Diabetes is a complication and malfunctioning of body that is related to a hormone called insulin. Insulin distributes and consumes glucose in body for proper functioning of all other body organs. If insulin is not produced in required amount by an organ called pancreas, then the level of glucose in blood increases. This situation is simply termed as blood sugar. When this situation repeats for a longer period and glucose level does not decrease, the situation becomes chronic and it is christened as diabetes in medical science terminology. Diabetes can be classified as type 1 diabetes and type 2 diabetes. The symptoms, a diabetic patient deals with include sudden weight loss, fatigue, frequent urination and thirst etc. It is required to check blood glucose level regularly to assess the diabetes in any person. Deranged values of glucose level in blood confirm the diabetes in a person. A special test AIC is conducted to assess average blood sugar level of past 2-3 months. Patient has to consult endocrinologist for suitable treatment of any type of diabetes. Patient has to consume continued medicines and follow right diet plan prescribed by endocrinologist. Diabetes may also lead to other complications such as heart-failure, kidney failure, if not treated well. Regular monitoring of blood glucose level is prescribed to prevent any emergency situation. Type 2 diabetes is found to be more dangerous than type 1 diabetes.

Public Attitude towards Type 2 Diabetes

Type 2 diabetes makes the human body resistant to insulin. Insulin is produced however human body cannot make full use of it. Due to this fact, type 2 diabetes can be considered
as riskier than that of type 1. In type 1, it is possible to compensate for insulin from external sources such as injection, insulin pump, tablets etc. On the other hand, in case of type 2 diabetes, body cells do not perform the work to utilize the insulin present in body. The worst part in type 2 is that insulin is present, but the body cannot use it in proper manner which results excessive need of insulin. So, at that time medical treatment sometimes also does not appear to be fruitful.

In the long run, either type of diabetes can cause serious health issues, if not controlled. A number of complications such as damaging of both small and large blood vessels, high risk of cardiovascular disease, blindness and kidney problems can occur. The poor lifestyle, unhealthy diet and lethargic exercise habits can cause the condition of type 2 diabetes. It produces issues with the mechanism of the body that extracts glucose from blood and put it into cells for energy transformation. The chronic, undiagnosed and especially untreated diabetes of type 2 may lead to the condition of Hyperosmolar Hyperglycaemic Syndrome (HHS). In this situation, the level of glucose in blood rises instantly which results in dehydration and confusion in patient. Patient may lose his/her consciousness and need immediate medical treatment.

As it has been observed that after a few years with type 2 diabetes, it becomes a chronic condition in patients, the patient’s feelings seem to be very unenthusiastic and negative about the life. The tweets and responses, collected through twitter from the people suffering from type 1 and type 2 diabetes indicate that people have very pessimistic and negative thoughts about recovery and reversing of type 2 diabetes. They become hopeless about the recovery and the control of this type of diabetes. On the other hand, different scenario has been revealed about type 1 diabetes by the data from the same source. The perception of the people with type1 diabetes is positive to handle their condition of diabetes. According to opinions of people, it is found that type 1 diabetes can be managed and cured at highest possible level by adopting stable and healthy diet, better work out strategies and regular check-up of blood sugar level by taking other preventive measures that decrease the symptoms of type 1 diabetes. People feel optimized about their recovery and stronger health by organizing their diet, workout and stress management plan.

**Proposed Work**

In order to compare the sentiments related to type 1 and type 2 Diabetes, Sentiment Analysis is performed separately for both categories of Diabetes. The dataset containing type 1 diabetes tweets is extracted using Apache Flume. The same procedure is repeated for extracting tweets related to type 2 diabetes. The tweets are loaded into a table using
Apache Hive. HiveQL is used to generate the sentiment score of each tweet by comparing the words of tweets with the AFINN Dictionary. This process is executed separately for type 1 and type 2 diabetes tweets. Finally, the sentiment score for both types is exported to an excel file and the sentiments are compared using bar chart and pie chart. This experiment is repeated five times in order to achieve maximum accuracy in the final sentiment comparison for type 1 and type 2 diabetes. The average of the sentiment scores of five experiments is considered as the final result for the research.

Methodology

The following methodology is used to perform the experiment:

1. Initially, the Hadoop Ecosystem is turned on.
2. The tweets containing the keywords ‘type 1 diabetes’ and ‘type 2 diabetes’ are extracted separately using Apache Flume by specifying the keywords in ‘twitter.conf’ file.
3. The extracted tweets are downloaded using WebHDFS.
4. Then the tweets are loaded into a table using ApacheHive.
5. By using the HiveQL commands, the sentiment score of each word is compared with the words in the AFINN Dictionary.
6. The final sentiment score of tweets is calculated by adding the score of individual words.
7. The same steps for type 1 diabetes and type 2 diabetes are performed and sentiment scores of the tweets are stored in separate files.
8. The process to find sentiment score of tweets related to type 1 diabetes and type 2 Diabetes is repeated five times and sentiment scores for both the types of diabetes are stored in separate files.
9. The final results pertaining to the average of the sentiment scores obtained from five experiments are calculated and illustrated visually with the help of a pie chart and a bar chart.

Experiment for Type 1 Diabetes

Creating a Table Using ApacheHive

An empty table is created using the following command where the tweet Id and text of tweet for type1 diabetes are stored.

```
CREATE EXTERNAL TABLE type1_tweets (id BIGINT, text STRING) ROW
```
FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user';

The following figure shows the tweets stored in the table type1_tweets. This table contains the id and text of each tweet extracted.

![Figure 1 Tweets related to Type 1 Diabetes](image)

**Figure 1 Tweets related to Type 1 Diabetes**

### Splitting the Tweets into Separate Words

The tweets in the table type1_tweets are splitted into separate words using the following command.

```
create view split_type1_data as select id, words from type1_tweets lateral view explode(sentences(lower(text))) dummy as words;
```

Tweet id and splitted words are shown in figure 2 below:
The following command is used to create lateral view of the tweets stored in the table split_type1_data.

```
create view type1_lateral_data as select id, word from split_type1_data lateral view explode (words) dummy as word;
```

The lateral view of the tweets is shown in figure 3.

Figure 2 Individual words of Type 1 Diabetes' tweets

Lateral View of Extracted Tweets

The following command is used to create lateral view of the tweets stored in the table split_type1_data.

```
create view type1_lateral_data as select id, word from split_type1_data lateral view explode (words) dummy as word;
```

The lateral view of the tweets is shown in figure 3.
In order to find the sentiment score of each word in the tweets, AFINN dictionary has been used. The following hive command is executed to move AFINN Dictionary into ‘dictionary’ table stored in HDFS.

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

The contents of ‘dictionary’ table as shown in figure 4 are displayed using the following command.

select * from dictionary;
Generating Sentiment Score of Individual Words

To compute sentiment score of individual words in the tweets, the words in the text of tweets are compared with the dictionary by using the following command.

```
Create table type1_sentiment_score as select t.id, t.word, d.rating from type1_lateral_data t join dictionary d where t.word = d.word;
```

The figure 5 shows the execution of map reduce job generated by the above command.
The sentiment score of individual words in the tweets of type 1 diabetes is stored by the previous command in the table type1_sentiment_score. The following command displays the sentiment score of individual words in the tweets of type 1 diabetes as shown in figure 6.

```
Select * from type1_sentiment_score;
```

The sentiment score is obtain by comparing the words in the tweets of type 1 diabetes with the AFINN dictionary.
Sentiment Analysis of Extracted Tweets

The following hive command is executed to obtain the final results (sentiment score).

```
SELECT id, sum(rating), case when sum(rating)>0 then 'POSITIVE' when sum(rating)<0 then 'NEGATIVE' else 'NEUTRAL' end as sentiment from type1_sentiment_score GROUP BY id;
```

The execution of above command and final sentiment score of each tweet related to type1 diabetes is given in figure 7.
The similar procedure is followed in order to perform sentiment analysis for type 2 diabetes. An empty table is created using the following command where the tweet Id and text of tweet for type2 diabetes are stored.

```
CREATE EXTERNAL TABLE type2_tweets (id BIGINT, text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user';
```

The following figure shows the tweets in the table ‘type2_tweets’.

Figure 7 Final Sentiment Score of each tweet related to Type 1 Diabetes

Experiment for Type 2 Diabetes

Creating a Table to Store Tweets Related to Type 2 Diabetes

The similar procedure is followed in order to perform sentiment analysis for type 2 diabetes. An empty table is created using the following command where the tweet Id and text of tweet for type2 diabetes are stored.

```
CREATE EXTERNAL TABLE type2_tweets (id BIGINT, text STRING) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe' LOCATION '/user';
```

The following figure shows the tweets in the table ‘type2_tweets’.
The sentiment score of individual words in the tweets of type 2 diabetes is computed by executing the following command and are stored in the table `type2_lateral_data`.

```
CREATE TABLE type2_sentiment_score AS
SELECT t.id, t.word, d.rating
FROM type2_lateral_data t JOIN dictionary d ON t.word = d.word;
```

The execution of above command is shown in figure 9.
Figure 9 Map Reduce job generating sentiment score of each word in tweets related to Type 2 Diabetes

Displaying Sentiment Score of Words in Type 2 Diabetes Tweets

The command to display sentiment score of individual words in the tweets of type 2 diabetes is as under:

```
Select * from type2_sentiment_score;
```

The sentiment score of individual words in the tweets of type 2 diabetes stored in the table ‘type2_sentiment_score’ is shown in figure 10.
To generate the final score of all the tweets related to type 2 diabetes, the following command is executed:

```
SELECT id, sum(rating), CASE WHEN sum(rating) > 0 THEN 'POSITIVE' WHEN sum(rating) < 0 THEN 'NEGATIVE' ELSE 'NEUTRAL' END AS sentiment
FROM type2_sentiment_score
GROUP BY id;
```

After the execution of above command, the final sentiment score of the tweets related to type 2 diabetes is shown in figure 11.

**Figure 10** Sentiment score of words in Type 2 Diabetes’ tweets

**Sentiment Analysis of Tweets Related to Type 2 Diabetes**

To generate the final score of all the tweets related to type 2 diabetes, the following command is executed:

```
SELECT id, sum(rating), CASE WHEN sum(rating) > 0 THEN 'POSITIVE' WHEN sum(rating) < 0 THEN 'NEGATIVE' ELSE 'NEUTRAL' END AS sentiment
FROM type2_sentiment_score
GROUP BY id;
```

After the execution of above command, the final sentiment score of the tweets related to type 2 diabetes is shown in figure 11.
Results

The above experiment was repeated five times for both type 1 and type 2 diabetes and the following results in terms of percentage of tweets were obtained:

<table>
<thead>
<tr>
<th>Tweet Category</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
<th>Experiment 4</th>
<th>Experiment 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Tweets (Percentage)</td>
<td>27.6</td>
<td>41.8</td>
<td>30.3</td>
<td>43.5</td>
<td>25.2</td>
</tr>
<tr>
<td>Neutral Tweets (Percentage)</td>
<td>4.1</td>
<td>6.1</td>
<td>5.6</td>
<td>3.8</td>
<td>5.3</td>
</tr>
<tr>
<td>Positive Tweets (Percentage)</td>
<td>68.3</td>
<td>52.1</td>
<td>64.1</td>
<td>53.7</td>
<td>69.5</td>
</tr>
</tbody>
</table>
The average of the five experiments gave the following results in terms of percentage of tweets:

<table>
<thead>
<tr>
<th>Tweets Category/Type of Diabetes</th>
<th>Type1 Diabetes (%)</th>
<th>Type2 Diabetes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Tweets</td>
<td>28.94</td>
<td>49.16</td>
</tr>
<tr>
<td>Neutral Tweets</td>
<td>4.96</td>
<td>4.1</td>
</tr>
<tr>
<td>Positive Tweets</td>
<td>66.1</td>
<td>46.94</td>
</tr>
</tbody>
</table>

Figure 12 Polarity results for experiment 1; Figure 13 Polarity results for experiment 2

Figure 14 Polarity results for experiment 3; Figure 15: Polarity results for experiment 4
Conclusion

The present study mainly focuses on to express the comparison between negative and positive opinions toward both the types of diabetes. Tweets were taken as source data for the experiment. After performing similar experiment multiple times, it has been concluded that people’s behavior toward type 2 diabetes is more pessimistic as compared to their behavior toward type 1 diabetes. On an average, 49.16% people are pessimistic and have negative opinion about type 2 diabetes whereas only 28.94% people show negative behavior toward type 1 diabetes. The percentage of negative sentiments for type 2 diabetes as depicted in the graphs formed from the information received through multiple experiments clearly corroborates the results. On the basis of results obtained from repetitive experiments, it has been observed that people give up the hope to reverse the condition of type 2 diabetes. On the other hand, people are very much optimistic towards the condition of type 1 diabetes. Patients suffering from type 2 diabetes have to be more careful regarding their condition, life style and diet plan.
Significance and Future Scope

The present study aims to create an awareness about the diabetes related issues and attitudes among the people. It comes out with the result that diabetes type 2 patients require encouragement to make their lives pleasant which would help them to create a positive thinking about the cure, as strong desire to live will provide them the strength to fight with the disease. The study will help the future researchers to evaluate the feelings and assumptions of the diabetic patients which may vary with the time. The researchers
wish to develop a novel technique for performing sentiment analysis on twitter data in diabetes domain in future which would function more accurately to produce better results.

References


Pushp, D.A. Real-time Sentiment Analysis of Twitter Data using Flume & Hive on Hadoop.


