Tweet Summarization using NLP and Sentiment Analysis

Prof. Omprakash Yadav*, Abhay Shinde1, Omkar Palkar2

**Assistant Professor, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India*

1,2B.E student, Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India

---***---

Abstract - *During recent years, socially generated content has become pervasive on the World Wide Web. The enormous amount of content generated on Twitter that allows a huge number of users to contribute frequent short messages. It consists of small messages which are regarding some events happening in world or formally posting relating to themselves. Most of these messages are a reaction describing same events resulting in redundancy of tweets. The algorithm used takes a trending phrase or any phrase specified by a user, collects a large number of posts containing the phrase, and provides an automatically created summary of the posts related to the term. We get a global view regarding the messages in terms of short summaries relating trending terms during the course of a period of time such as an hour or a day.*

***KeyWords* : Tweets, fitness value, pbest, gbest, *stemming, stopwords, PSO Algorithm*.**

1. INTRODUCTION

Data mining is known as process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems occurs in Data mining. It is an important process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the info mining process is to extract information from a knowledge set and transform it into a clear structure for further use.

1.1 Flow of the Project

1. Retrieval of tweets
The tweets are extracted from Twitter Account

2. Pre-Processing
Pre-processing describes any type of processing performed on tweets to prepare it for another processing procedure.

3. Segmentation
The goal of tweet segmentation is to split the tweet into a sequence of semantically
meaningful unit or any other types of phrases which are more often used together. For tweet segmentation, HybridSeg framework is proposed.

4. Clustering of similar tweets
Clustering also called grouping multiple objects in a way that objects in the same group are more similar to each other than to those in other group (clusters). In this phase, similar tweets are clustered using Particle Swarm Optimization algorithm.

2. Input and Output

Real time data in the form of tweets using the source obtained from the Twitter API.

URLs of the users in the fetched data.

The summarized tweets obtained will be displayed for the user. The most favourite/liked/popular tweets related to the search parameters. The number of Re-tweets available.

3.1 Use-Case Diagram

The use case diagram deals with the front-end working of the system. It depicts how system appears to a user. Here user is the actor that deals with the system.

User will firstly login into the system after which she/he will browse tweets and select tweet for summarization.

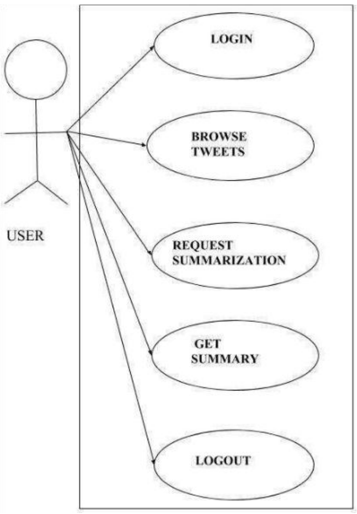

Fig -1: Use-Case Diagram

3.2 Class Diagram

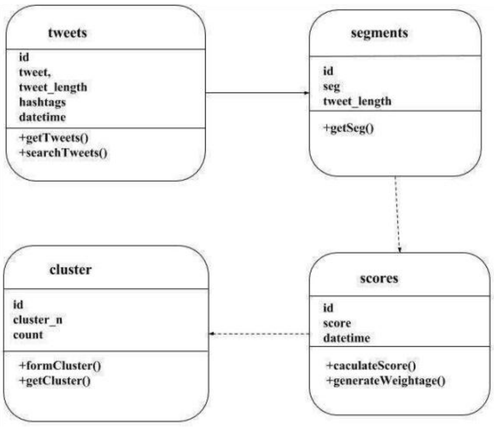

Fig -2: Class Diagram

3.3 Sequence Diagram

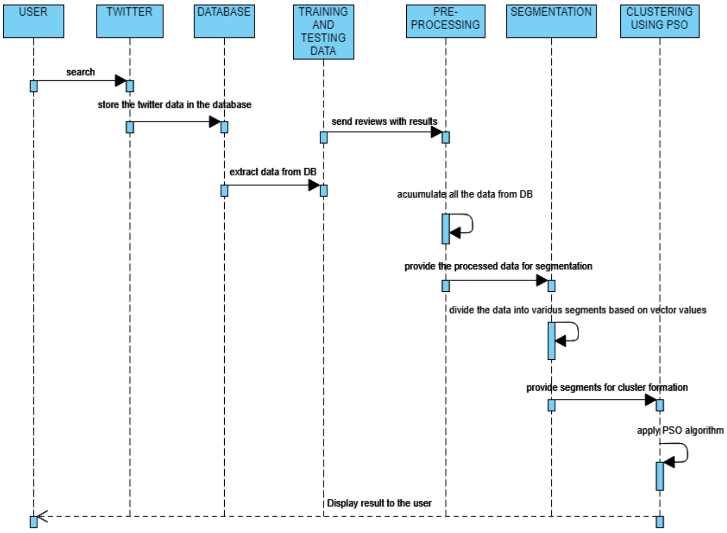
USER
TWITTER
DATABASE
TRAINING AND TESTING DATA
PRE-PROCESSING
SEGMENTATION
CLUSTERING USING PSO
search
store the twitter data in the database
send reviews with results
extract data from DB
acuumulate all the data from DB
provide the processed data for segmentation
divide the data into various segments based on vector values
provide segments for cluster formation
apply PSO algorithm
Display result to the user

Fig -3: Sequence Diagram

3.4 ER Diagram

Requested Summary of Tweet is given as output.

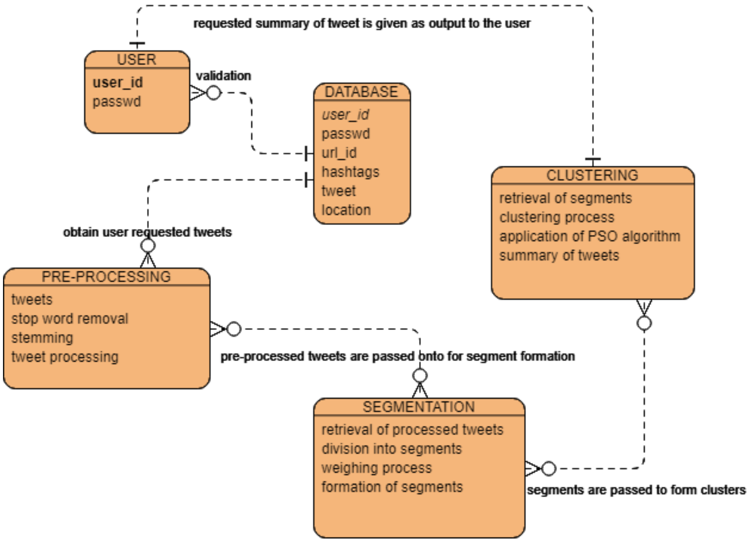

Fig -4: ER Diagram

4. Selecting the Best Algorithm

4.1 Existing Solution

Particle Swarm Optimization-Inspired by the flocking and schooling patterns of birds and fish, Particle Swarm Optimization (PSO) was invented by Russell Eberhart and James Kennedy in 1995. The computer software simulations of birds flocking around food sources, and then later realized how well their algorithms worked on optimization problems was originally develoved by these two scientists.

4.1.1 Particle Swarm Optimization

It might sound complicated, but it's really a very simple algorithm. A group of variables have their values adjusted closer to the member whose value is closest to the target at any given moment, over a number of iterations. Consider, flock of birds circling round an area and the birds may smell a hidden source of food. The one who is closest to the food chirps the loudest and the other birds swing around in his direction. If any of the other circling birds comes closer to the target than the first, it chirps louder and the others veer over toward him. This tightening pattern continues until one of the birds happens upon the food. It's an algorithm that's simple and easy to implement.

The algorithm keeps track of three global variables:

- Target value or condition
- Global best (gBest) is the value that indicates that which particle's data is presently nearest to target.
- Stopping value is indicated when algorithm should stop if target is not found.

Each particle consists of:

- A possible solution in form of data.
- How much the Data can be changed which can be indicated by velocity value.
- Personal best (pBest) indicates nearest the particle's Data has ever come to Target

4.2 Proposed Algorithm

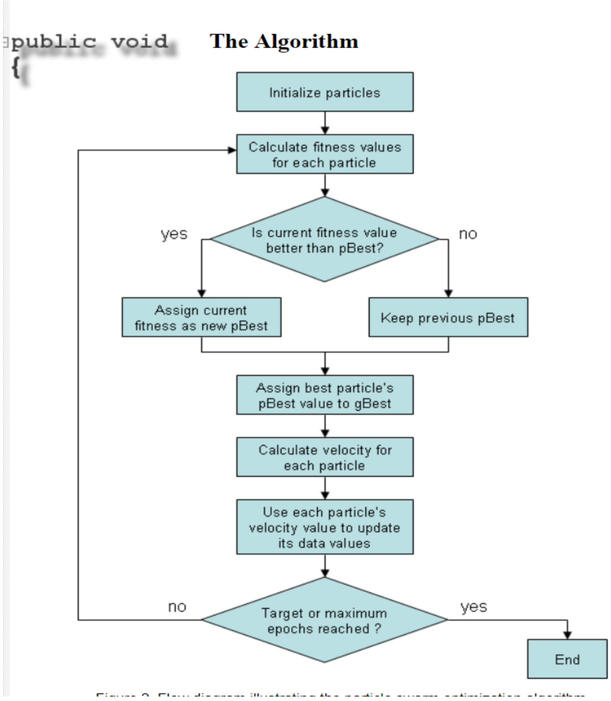
public void
{
The Algorithm
Initialize particles
Calculate fitness values for each particle
Is current fitness value better than pBest?
yes
no
Assign current fitness as new pBest
Keep previous pBest
Assign best particle's pBest value to gBest
Calculate velocity for each particle
Use each particle's velocity value to update its data values
Target or maximum epochs reached ?
no
yes
End

Fig -5: PSO Flowchart

4.3 K-means Algorithm

K-means Clustering is simplest and popular unsupervised machine learning algorithms. Unsupervised algorithms usually tend to make inferences from datasets where they use only input vectors without referring to known, or labelled, outcomes.A collection of data points is known as cluster which is aggregated together because of certain similarities. The number of centroids you need in the dataset, is denoted by which is its target number. Centroid is known as imaginary location which represents the centre of the cluster. Each cluster is allocated a data point by reducing the in-cluster sum of squares.

4.4 Working

The learning data is processed by the K-means algorithm where in process of data mining starts with a first group of centroids which are randomly selected and are used as beginning points for each cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

Halts creating and optimizing clusters when:
- Stabilized centroids are achieved — i.e there is no change in values as the clustering is successful.
- Achieved defined number of iterations.

5. CONCLUSIONS

A branch of data mining that deals with opinions, expressions and decision making is referred as Sentiment Analysis. There are many systems that perform summarization on twitter to gain the different opinions expressed by user via tweets. These systems make use of various clustering algorithms like K-means are used for clustering data and providing summary of a tweet, these algorithms may be easy to implement but they have a less accuracy rate.

The system developed gives user an application that summarizes tweets more accurately so that user gets a proper summary of the tweet user wants. For this purpose, the system makes use Particle Swarm Optimization algorithm for clustering tweets and providing user with a more accurate summary for a user desired tweet. This results in a tweet summarization system with higher accuracy rate as PSO is an optimization algorithm, with accuracy rate higher than that of other systems. Thus, the proposed system provides a new way of summarizing tweets more accurately.

REFERENCES

[1]. S. Dutta, "A graph based clustering technique for tweet summarization. "Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on. IEEE, 2015.

[2].M.Al-Dhelaan and H.Alhawasi. "Graph Summarization for Hashtag Recommendation.", Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on. IEEE, 2015.

[3]. Tae-Yeon Kim, "A tweet summarization method based on a keyword graph." Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication. ACM, 2014.

[**4**]. A. Chellal, B. Mohand and B. Dousset. "Multi-criterion real time tweet summarization based upon adaptive threshold." Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on. IEEE, 2016.

[5]. J. Weng, "Tweet Segmentation and its Application to Named Entity Recognition." (2015): 1-15.

[**6**]. R. P. Narmadha and G. G. Sreeja. "A survey on online tweet segmentation for linguistic features." Computer Communication and Informatics (ICCCI), 2016 International Conference on. IEEE, 2016.