Sentiment Analysis of people perception on Wasa Water condition using machine learning approach.

Abstract—Water is an important element for human [6]. Pure water is for the quality life in smart cities [7]. Recently, In Dhaka WASA drinking water quality has been degraded because of numerous styles of pollution caused by disposal of human wastes, industrial wastes, automobile wastes and lack of proper materials. In this paper our aim is to predict public sentiments about Dhaka WASA drinking water. For classification we use logistic regression machine learning algorithm. The algorithm finds the optimal values from the coefficients.

Keywords –sentiment analysis, logistic regression, Dhaka wasa drinking water

I. Introductions

Cause of the social networking, people will share their view via social media sites as Facebook and other Social sites [8]. The first piped potable system in national capital was established in 1874 by Khwaja Abdul Ghani, the leader that dominated national capitalbelowBrits colonial authorities. The system was fed by a water treatment plant in Chadnighat close to the bank of the stream Buriganga[6]. When independence from Brits in 1948 the Department of Public Health Engineering of the Pakistani government was accountable of potable offer still as hygienically sewers and storm-water voidance. National capital WASA (Water offer& Sewerage Authority) was established in 1963[6]. In 1989, the storm-water system of national capital town was bimanual over to DWASA. In 1990, the spot was extended to incorporate Narayanganj town[6]. Within the early Nineties the globe Bank had aforementioned it'd solely offer a loan for installation in national capital if the utility would enter into a public-private Partnership with a global water works. once this was rejected, it asked that revenue request and assortment ought to be outsourced to a personal company for a minimum of one spot on a pilot basis, which DWASA ought to be reworked into a commercially orientating utility.[4] The outsourcing in one spot was tired 1997, however the pilot program wasn't deemed productive and was stopped. DWASA's activities are reorganized by the national capital WASA Act, 1996 that reworked DWASA into a service-oriented industrial organization.

Sentiment analysis is that the automatic[10] extraction of opinions, emotions, and sentiments from texts. Sentiments, opinions, and emotions square measure subjective impressions and not facts, that square measure objective or neutral. Much not research work on sentiment analysis for classification of Dhaka wasa water.

In this paper we have extracted sentiments or opinions of people from news portal, web scraping, twitter using twitter advanced search and social platform and then identified the overall polarity of texts as positive, negative[12].For classification we use Logistic Regression. For preprocessing data we use regression and then we transform data into BOW (Bag-of-Words) model then transform BOW model into TF-IDF model. After we use logistic regression algorithm to create the classifier. Logistic regression is a type of learning algorithm. It learns from a training dataset, the pattern of the data and applies the learned logics on new data for prediction.

Table 1 some of people opinion about Dhaka WASA water.

1.	Dhaka wasa start a service .Which provide emergency water supply .If any reason water is not supply by wasa .They provide this service. It's really great service. Thanks wasa.
2.	For last 20 days water supply is not enough. More than 2 places pile line are broke down but still there is no one to fix it.
3.	The Dhaka City Corporations and Dhaka Wasa have to do their development works in coordination with each other. To overcome the problems.
4.	Because of undrinkable water we have to buy water. It is impossible to continue it.

Table: 1

II. PREVIOUS WORKS

Our work is inspired from some other work mostly related and a few of them for our information gaining. Mostly [1] they are try to find out detection and biochemical characterization of the Isolates. They are trying to find out microbiological quality of Dhaka WASA drinking water. But our aim to find out the public sentiment about Dhaka Wasa water. We use logistic regression for classification. The sentiment analysis task is mainly a binary classification problem to predict whether a given sentence is positive of negative.

III. METHODOLOGY

Our initial work is to preprocessing data. We use python and Natural Language Toolkitc (NLTK[1]).For vectorization and classification we import sklearn[2].

A. Dataset

We collect data from news portal, web scraping, twitter using twitter advanced search and social platform. At the beginning of our research work we try our best to collect data for preparing our main data sets. We use twitter advanced search rather than Twitter API because using twitter advanced search

we collect data based on several parameter. We also collected data manually from several Microblog Posts. We collect 2234 data for our dataset.

```
#Importing BeautifulSoup
from bs4 import BeautifulSoup as soup  # HTML data structure
from urllib.request import urlopen as uReq  # Web client

# URl to web scrap from.
page_url = "http://www.dhakatribune.com/bangladesh/dhaka/2019/05/07/
dwasa-must-take-responsibility-for-its-polluted-water-supply"
.......
more
......
```

Figure: 1

Web Scrapping

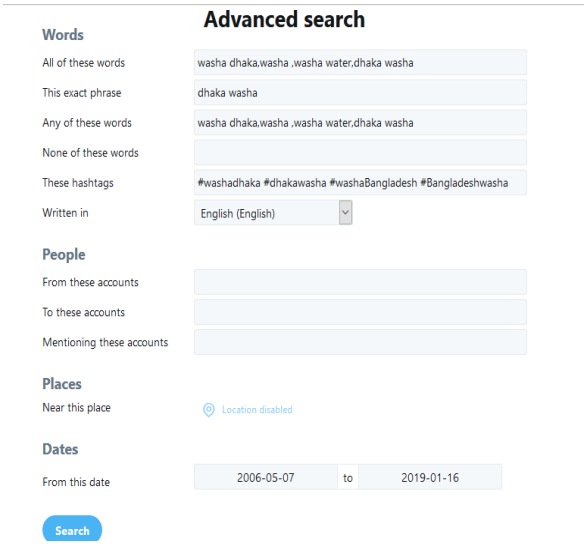


Figure: 2

B. Preprocessing

From raw data which we collected through news portal, web scraping, twitter advanced search, social platform and manually from several Micro blog Posts are contain unnecessary things .So, we need to preprocess those data. We preprocess

Tokenization:

Tokenization is the way toward isolating a grouping of strings into people, for example, words, catchphrases, expressions, images and different components known as tokens. Tokens can be singular words, states or even entire sentences. Quite the procedure of tokenization, a few characters like accentuation marks are disposed of. The tokens function as the contribution for various procedures like parsing and content mining [3].

In a sentence can be containing stop words like 'the', 'is', 'are'. Stop words can be filtered from the sentence to be processed. Actually there is no universal list of stop words in natural language processing. So we generally ignore these words to enhance the accuracy of our analysis. In different format there are different stop words depending on the country, language etc. Example in English format there are several stop words [3].

Part-Of-Speech Tagger (POS Tagger):

A Part-Of-Speech Tagger (POS Tagger) is a process of assigning one of the parts of speech to the given word. It is generally referred to as POS tagging. Parts of speech generally contain nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories .Computational applications use more fine-grained POS tags like 'noun-plural'. Parts of Speech tagger is a program that does this job [3].

Tagging. In raw data contain some special characters (! @#! #$ %^), digits (0-9), parts of through Regular Expression, tokenization, normalization, stemming, Part of Speech (POS)

C. Feature Extraction

Bag of Words:

Bag of Words: Bag of the word is a method of extracting options by representing simplified text or knowledge, utilized in language process and information retrieval [3]. During this model, a text or a document is pictured because the bag of its words. So, merely bag of words in sentiment analysis is making a listing of helpful words. We've used a bag of words approach to extract our feature sets. Once the preprocessed dataset, we tend to used pos tagging to separate completely different elements of speech and from that, we choose nouns and adjectives and use those to form a bag of words. Then we convert BOW model into IF-IDF model. Then we use Logistic regression is a prediction.

IV. LOGICTRIC REGRESSION

The sentiment analysis task is mainly a binary classification problem to predict whether a given sentence is positive of negative. In our demonstrations we denote '0' as negative and '1' as positive.

Logistic Regression – The points concept

- Each sentence is mapped to a point.
- If the point is greater than 0.5 then positive else negative.

Logistic is a learning algorithm is a specific type of algorithm whose performance increases with time. Logistic regression is a type of learning algorithm. It learns from a training dataset, the

pattern of the data and applies the learned logics on new data for prediction.

Logistic Regression – Linear Equation

Consider the equation: $y = a + bx_1 + cx_2 + \cdots + dx_n$

a, b, c, d = coefficients

X_1 , X_2 , …, X_n = independent variables

y = dependent variable

Logistic Regression – The point's concept

Words/Documents	Today	Dhaka	wasa	water	quality	is	very	good	Points
Document 1	0	0.07	0.17	0.17	0	0.07	0.17	0	0.62
Document 2	0	0	0.07	0.07	0.07	0	0	0	0.41
Document 3	0	0.05	0.05	0.05	0	0	0	0	0.72

Logistic Regression – Predicting Sentiment

If y >= 0.5 → Positive sentiment

If y < 0.5 → Negative sentiment

Logistic Regression – Value range

For some values of the dependent variables, the value of y can be > 1 or < 0.

For that, we need some way to restrict the value of y within the range 0 and 1.

Logistic Regression – Value range

Assume n = 2000

$$\text{For } y > 0, \quad y = e^{(a+bx_1+cx_2+\cdots+dx_{2000})}$$

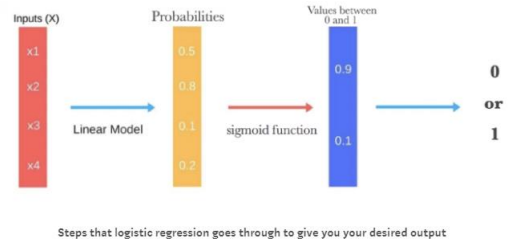
$$\text{For } y < 1, \quad y = \frac{e^{(a+bx_1+cx_2+\cdots+dx_{2000})}}{e^{(a+bx_1+cx_2+\cdots+dx_{2000})} + 1}$$

Figure: 4

$$\ln\left(\frac{y}{y-1}\right) = a + bx_1 + cx_2 + \cdots + dx_{2000}$$

Figure: 5

Steps that logistic regression goes through to produce desire output.



Steps that logistic regression goes through to give you your desired output

Figure: 6

Graph

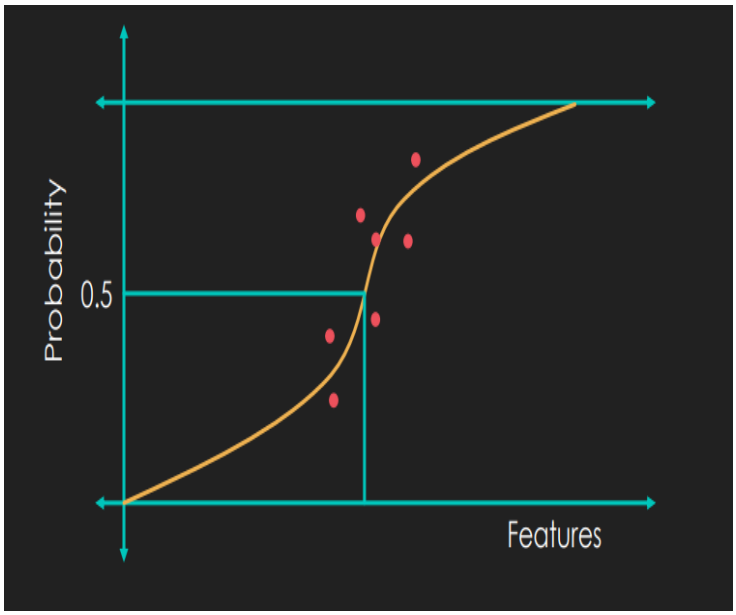


Figure: 7

V. DATA GENERALIZATION

For classify the opinion, we have collected about 2234 comments from different sources using news portal, web scraping, twitter using twitter advanced search and social platform. After collecting we generated dataset by labeling the comments as positive or negative proper way. An example of this file shows Table 2.

Table: 2 Sample of the dataset

Positive	Dhaka wasa start a service .Which provide emergency water supply .If any reason water is not supply by wasa .They provide this service. It's really great service. Thanks wasa.
Negative	Free School Street, Kathal Bagan of Dhaka city last 3

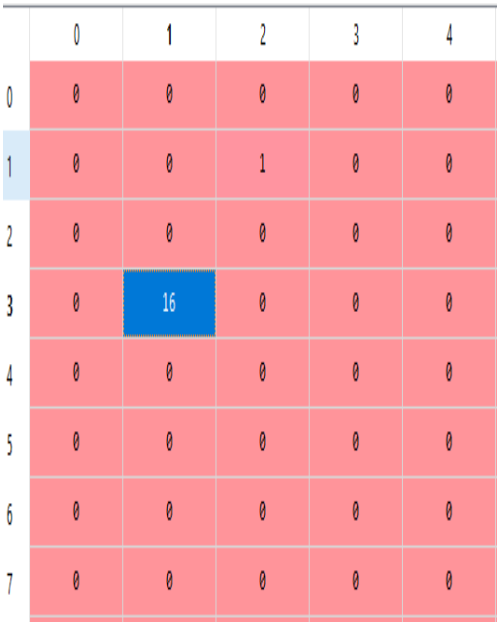
	Days no have water. It's very disgusting time we are passing. Please authority take a step and realize our situation.
Positive	Dhaka wasa water buy new pump .I think government taken a good decision.
Negative	The problem is rooted in Dhaka Wasa distribution system more than the supplied water itself.

Table: 2

VI. EXPERIMENTAL RESULTS AND EVALUATION:

To evaluate the performance our classifier we test our classifier manually and test over 500 comments.

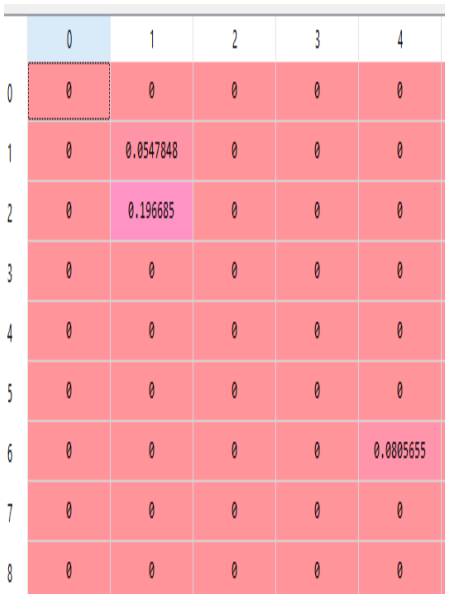
Bag-of-words: Some part of Bag-of-words array



	0	1	2	3	4
0	0	0	0	0	0
1	0	0	1	0	0
2	0	0	0	0	0
3	0	16	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0

Figure: 8

TF-IDF model: Some part of If-IDF array



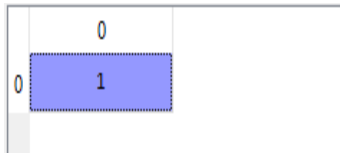
	0	1	2	3	4
0	0	0	0	0	0
1	0	0.0547848	0	0	0
2	0	0.196685	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0.0805655
7	0	0	0	0	0
8	0	0	0	0	0

Figure: 9

Test Case 1: Manually -> Positive

```
# Using our classifier
with open('tfidfmodel.pickle','rb') as f:
    tfidf = pickle.load(f)
with open('classifier.pickle','rb') as f:
    clf = pickle.load(f)
sample = ["Emergency Water service it's really very good"]
sample = tfidf.transform(sample).toarray()
sentiment = clf.predict(sample)
```

Result:



	0
0	1

Figure: 10

We see result is 1.That means Positive.

Test Case 2: Manually -> Negative

```
# Using our classifier
with open('tfidfmodel.pickle','rb') as f:
    tfidf = pickle.load(f)
with open('classifier.pickle','rb') as f:
    clf = pickle.load(f)
sample = ["Dhaka wasa water is not useable.Very bad"]
sample = tfidf.transform(sample).toarray()
sentiment = clf.predict(sample)
```

Result:

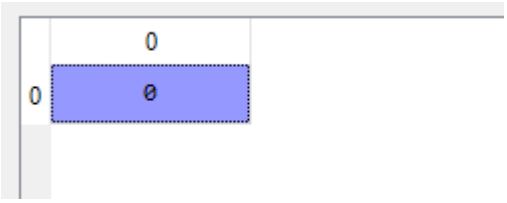


Figure: 11

We see result is 0.That means Negative.

Result given below on 500 comments.

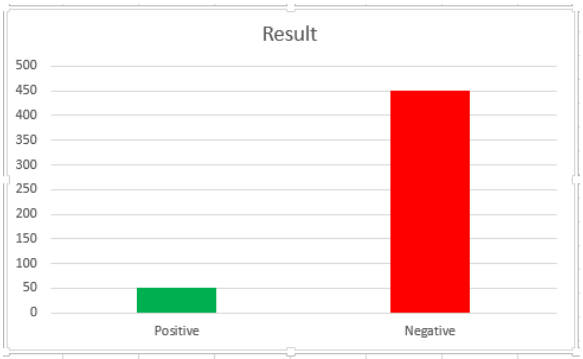


Figure: 12

VII. CONCLUSION AND FUTURE WORK

For future we need some improvement in our research.
First of all, we have a limited amount of data where we
have used 30% of our dataset as test set and found around
60% accuracy. Now our first target is to increase our dataset. We are trying to improve our approach for better
accuracy of our result applying deep learning theory for our existing system. We are trying to using proper natural language processing will highly improve our system. So we think that we will definitely go on to workout with the more accurate result.

VIII. ACKNOWLEDGMENT

Our deep thanks to our supervisor Md.Alomgir Hossain for his guidance, giving flexibility and continuous support throughout the work. We are also thankful to our family, friends for their support and encouragement [3]. Finally we thank our faculties at the Department of Computer Science and Engineering of IUBAT University for their support throughout the research work.

REFERENCES

[1] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, 2018

[2]. K. Mahbub, A. Nahar, M. Ahmed, and A. Chakraborty, "Quality Analysis of Dhaka WASA Drinking Water: Detection and", *JESNR*, vol. 4, no. 2, pp. 41-49, Mar. 2012.

[3] S. Arafin Mahtab, N. Islam and M. Mahfuzur Rahaman, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, 2018.

[4]http://dspace.bracu.ac.bd/xmlui/handle/10361/8246.

[5] M. Ali and A. M. Qamar, "Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan," *Eighth International Conference on Digital Information Management (ICDIM 2013)*, Islamabad, 2013

[6]https://en.wikipedia.org/wiki/Water_management_in_Dhaka

[7] G. Kang, J. Z. Gao and G. Xie, "Data-Driven Water Quality Analysis and Prediction: A Survey," *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, San Francisco, CA, 2017

[8] S. Abu Taher, K. Afsana Akhter and K. M. Azharul Hasan, "N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, 2018

[9] https://www.statisticssolutions.com/what-is-logistic-regression/

[10] https://en.wikipedia.org/wiki/Sentiment_analysis