ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

An Improved Unsupervised Machine Learning Technique for Tweet Summarization

Maddi Sri V S Suneeta

Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam,

Andhra Pradesh, India

ABSTRACT: People run-out of time and even then, they find a small time to know what's happening all over the world. So, in such hefty schedule people can't go through all the tweets which users post all day long, and the solution for this is summarizing the tweets and getting the abstract of a particular news of event a person is searching for. Basically, summarizations are of two types Single Document Summarization and Multiple Document Summarization. Tweet summarization is a multi-document summarization as micro-blogging environment which is taken into account is a word restricted document where a person is limited to a certain number of words. Therefore, to summarize many tweets to get a perfect information about a topic we use Multiple Document Summarization. Here, key features like the temporal features of the tweets are included to Extractive Summarization of the tweets using an improved version of LexRank an Unsupervised Machine learning technique.

KEYWORDS*:* Tweets, micro-blogging, Single Document Summarization, Multiple Document Summarization,Tweet summarization, Extractive summarization, LexRank.

I. INTRODUCTION

Twitter is a social media networking site where people post messages, videos and pictures publicly every day. Twitter is a microblogging site where the posts were earlier limited to 140 characters later on it was increased, rather doubled to 280 characters for all the languages except for three languages. As the posts in this social media networking site are very short, like the short and sweet chirp of a bird they are called as tweets. Tweets are by default in the public mode i.e. everyone having an account on twitter can see the posts, if the user restricts the visibility then only the tweets are visible to their followers.

People will be posting a lot of information on twitter; it is secondary that if the information is useful to the user or not. Everyday there may be more than 500 million tweets posted on twitter. This short message service microblogging website has become a large source for breaking news where people post may interesting information about everything happening all over the world.

On the other side users not only post but they also follow and read the other people's posts. People have their own interests in reading about any particular topic. If the user tries to search matter about a particular topic, this would be like information overflow as there will be many tweets running down about a topic where people tweet about that particular topic weeks together. Also, people retweet and this can cause data redundancy. Searching for a perfect information about a particular topic would make a huge loss of time. This can be done by getting summary of all the tweets related to the particular topic which will give a brief information about the topic the person is interested in.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Here, summary of tweets means, getting the perfect matter on a topic which gives good idea about the topic the user is interested in. This should be generated automatically by using a system. This system is called a summarizer, a summarizer summarizes documents or sentences by using some methods or algorithms which are relevant according to the requirements of the user. So, instead of searching or reading all the tweets related to a topic to get correct information about it this summarizer gives sentences which gives strong information about the topic. Summarization according to the input type is basically of two types:

1. Single Document Summarization
2. Multiple Document Summarization

Single Document Summarization is summarizing a large document into a small and informative text, i.e. the input will be only one document and the output will be a few lines about the document. *Multiple Document Summarization* is getting a perfect information from many sentences or documents as an input, i.e. input is multiple document. Twitter is a microblogging site, where short messages called tweets are posted and summarizing the tweets comes under multiple document summarization.

Based on the type of the output of the summary, summarization is of two types:

1. Abstractive
2. Extractive

Abstractive Summarization is summarizing the document by changing the sentences by giving not intentionally but like how a human summarizes by forming his own sentences. *Extractive summarization* is done by selecting some important sentences from the given document itself. As abstractive summarization is more difficult to be presented, most of the summarizing models use extractive method to summarize the document.

In this paper the methods used are Extractive summarization and Multiple document summarization.

II. RELATED WORK

Text Summarization:

Summarization in general is the process of filtering important information from a source of a huge text or a document to make the information legible and intelligible. This is done by us i.e. humans always as we have the power to understand and put the idea what we have understood in our own words. This saves time as we need not read the whole document to get the information, we get clear idea about the information about the document. Text summarization is summarizing a text document which contains paras of document. The document may be Single Document

Page rank:

Page ranking is ranking a web page how important is that page. How frequently the page is been searched by the users. Websites with more links from other sites are mostly considered important web pages using the page rank algorithm.

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

d-damping factor, N-No. of nodes, v-vertex, u-adjacent side.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Text rank:

This is a machine learning technique where the article will be concatenated and then the this will be split into single sentences. Then the words will be embedded and represented as vectors. The similarities between the vectors which are calculated will be stored in the similarity matrix and this will be then converted into a graph, where the vertices represent the sentences and the edges represent the similarity score of the sentences connected with the edge. The top most ranked sentences will be the summary of the document as an extracted summary.

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

$S(V_i)$ - Score of a vertex V, E – set of edges, d – damping factor which is between 0 to 1 and usually set to 0.85.

LexRank:

Lexrank is an unsupervised and extractive summarization method which takes input of sentences and used graph-based centrality and compares the sentences and produces two sentences as the output. This is an extractive method of summarization, where it all deals with importance of sentences. Sentences with great importance will be the output of the algorithm.

$$\mathrm{idf}_i = \log\left(\frac{N}{n_i}\right)$$ idf - inverse document frequency

$$idf-modified-cosine(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum x_i \in x (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum y_i \in y (tf_{y_i,y} idf_{y_i})^2}}$$

w - word in two sentences x, y.

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{\text{idf-modified-cosine}(u,v)}{\sum_{z \in adj[v]} \text{idf-modified-cosine}(z,v)} p(v)$$

$p(u)$-LexRank

III. UNSUPERVISED LEARNING TECHNIQUES

Machine Learning: Deep learning is a subset of Artificial Neural Network and in turn Artificial Neural Network is a subset of Machine Learning and Machine Learning is in turn a subset of Artificial Intelligence. *Artificial Intelligence* is adding Human Intelligence to machines mainly computers where the system is able to learn, give reasons and correcting the mistakes made by itself. *Machine Learning* being a part of Artificial Intelligence provides the machines the ability to learn, give reasons and make corrections by itself without explicitly programming the system. Machine Learning allows the system to learn and correct itself by learning from its past experiences like how a human learns.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Types of machine learning algorithms:

1. *Supervised learning:* This is a Machine Learning technique where the machine learns under the guidance and the training data set is labelled. We have the input and output explicitly defined.
2. *Unsupervised Learning:* Here the machine will not have any labelled training data set and no explicitly defined input and output which is the exact opposite to supervised learning.
3. *Reinforcement Learning:* In this type of machine learning the learning all deals with rewards, where the machine gives feedback after every action performed by the artificial intelligent system. The data set is not trained but the system decides which method or which process, or which path is the best to follow.

Unsupervised Learning:

Unsupervised Machine learning is making the machine learn like how a human learns by his past experiences. The data set is not labelled. So, the machine checks for some sort of patterns to learn what kind of data is that. There are many unsupervised machine learning algorithms like clustering, anomaly detection, associative clustering. These are some of the unsupervised machine learning algorithms used in general. As, mentioned in the related work, text summarization, page ranking, text ranking and lexranking are the methods used for summarization purpose. In this paper we are using LexRank algorithm to summarize the tweets which are tweeted by a particular person at a particular time period on a particular topic.

IV. ABOUT THE DATASET

The datasets for this paper are collected form twitter API. The tweets made on a particular topic by the same person on a particular time period are identified and collected. The data sets are taken in csv format. In this way ten data sets are collected and each dataset is then pre-processed and summarized.

V. EXPERIMENT AND RESULTS

Proposed Method

This paper provides the improved LexRank is used to summarize the tweets. In general, the timestamp is not considered, here we consider the time of the tweet. This makes it easy to summarize the tweets. The tweets produced at a particular time period by the same person, will most probably be on a same topic. So, it is difficult to read many tweets on the same topic to know about what the tweets are. Using LexRank these tweets are summarized and this summary gives about what the person is talking about in those tweets.

Procedure

1. Firstly, the tweets are collected from twitter, by identifying the time period of the post.
2. These tweets are saved in a csv file and then pre-processed. In pre-processing the punctuations are removed, tokenization is done, then stopwords are removed, stemming and lemmatization is done.
3. The number of tweets and number of words in tweets are calculated.
4. Tweets are summarized to get the main idea what the person is talking about in that time period.
5. Also, we get the degree centrality scores with the summary.
6. The precision, recall and f-score values are calculated and graphs are produced.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

Result

LexRank gives the results which are produced using the time of the tweet. The precision is higher than the recall while using lexrank. In this paper we have used ten datasets and each one is pre-processed by removing punctuations, tokenizing, removing stopwords, stemming and lemmatizing the tweets. The tweets are then summarized and precision and recall values are produced using ROUGE scores and least common subsequence, by considering and comparing the system generated summary and manually generated summary. Then the f_scores are also calculated and the graphs are generated for every dataset. As there are ten datasets each dataset produces different precision, recall and f_score values. Using these values, we produce bar graphs for each output. So, we will have ten graphs comparing the precision, recall and f measure values. Graphs produced are shown below.

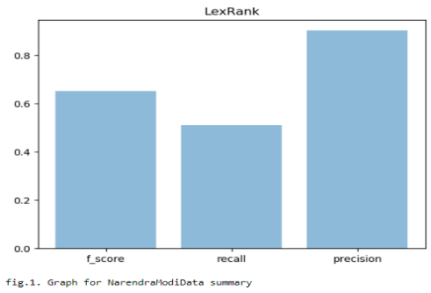

Fig.1. Graph for Narendhra Modi Dataset summary

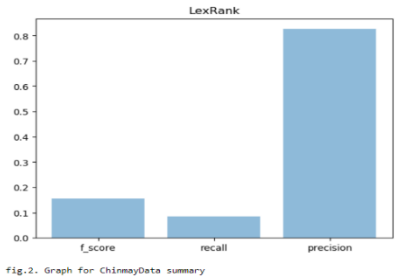

Fig.2. Graph for Chinmayi Dataset summary

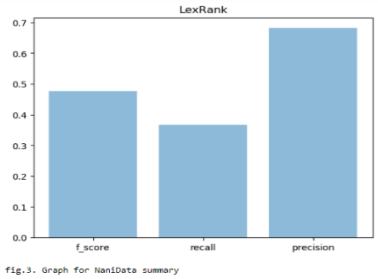

Fig.3. Graph for Nani Dataset summary

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

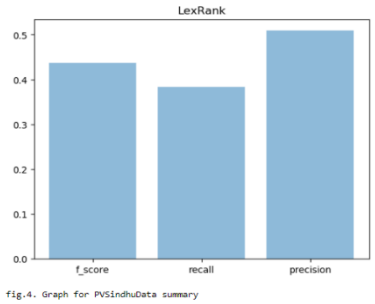

Fig.4. Graph for P V Sindhu Dataset summary

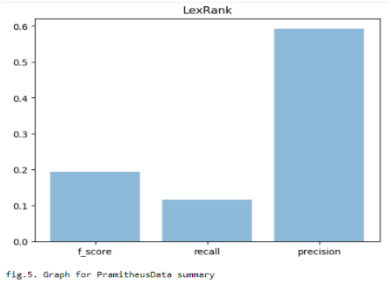

Fig.5. Graph for Pramitheus Dataset summary

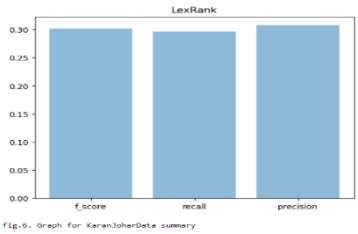

Fig.6. Graph for Karan Johar Dataset summary

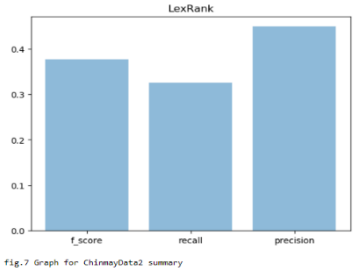

Fig.7. Graph for another Dataset of Chinmayi summary

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

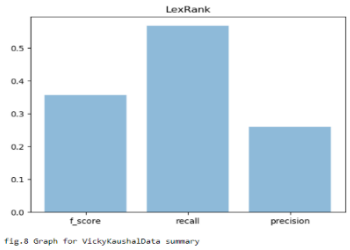

Fig.8. Graph for Vichy Kaushal Dataset summary

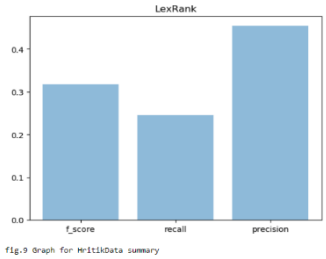

Fig.9. Graph for Hrithik Dataset summary

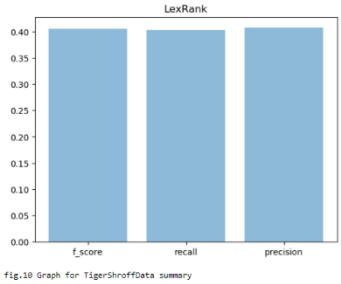

Fig.10. Graph for Tiger Shroff Dataset summary

VI. CONCLUSION

All the graphs show precision is higher compared to the recall while using the LexRank. This method can be improved further, the tweets which are in different languages an also be considered, also the tweets which are on the same topic but the time period also may vary, i.e. the person may tweet on the same topic again later and there may be many tweets on the same topic which are related to each other which are not posted in the same time and by another person.

REFERENCES

1. See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/308765988 Modified LexRank for Tweet Summarization Article in International Journal of Rough Sets and Data Analysis · October 2016 DOI: 10.4018/IJRSDA.2016100106
2. www.ijraset.com Volume 4 Issue IX, September 2016 IC Value: 13.98 ISSN: 2321-9653 International Journal for Research in Applied Science & Engineering Technology (IJRASET) ©IJRASET: All Rights are Reserved 390 An Novel Summarization and Timeline Generation Process of Evolutionary Tweet Streams Vedavyas Poola1, A. Narayanarao2, M. Srinivasulu31M.Tech, 2Associate Professor, 3Assistant Professor Shree Institute of Technical Education,Tirupati

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 10, October 2019

3. Araki, T., Ikeda, N., Dey, N., Chakraborty, S., Saba, L., Kumar, D., & Laird, J. R. et al. (2015). A comparative approach of four different image registration techniques for quantitative assessment of coronary artery calcium lesions using intravascular ultrasound. Computer Methods and Programs in Biomedicine, 118(2), 158–172.doi:10.1016/j.cmpb.2014.11.006 PMID:25523233
4. Azar, A. T., & Hassanien, A. E. (2015). Hybrid TRS-PSO clustering approach for Web2. 0 social tagging system.International Journal of Rough Sets and Data Analysis, 2(1), 22–37.
5. Automatic Summarization Of Tweet From Social Media Using Tweet Classification and Clustering(Based On GPU).ThoratPrashant,IchamVivek,Modi Malay,UmrekarPandurang Student, Computer Department, Met's BkcIoeNashik, Maharashtra, India
6. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015On Summarization and Timeline Generation for Evolutionary Tweet StreamsZhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra
7. 2016 IEEE/WIC/ACM International Conference on Web IntelligenceMulti-criterion real time tweet summarization based upon adaptive thresholdAbdelhamid Chellal IRIT Laboratory University of Toulouse III Toulouse, France Email: abdelhamid.chellal@irit.fr Mohand Boughanem IRIT Laboratory University of Toulouse III Toulouse, France Email: mohand.boughanem@irit.fr Bernard Dousset IRIT Laboratory University of Toulouse III Toulouse, FranceEmail: bernard.dousset@irit.fr
8. Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features Mr. G. S. Mane Research Scholar, Computer Science & Engineering Department Walchand Institute of Technology, Solapur Mrs. A. R. Kulkarni Assistant Professor, Computer Science & Engineering Department,Walchand Institute of Technology, Solapur
9. J Inf Process Syst, Vol.14, No.1, pp.79~100, February 2018 ISSN 1976-913X (Print)https://doi.org/10.3745/JIPS.02.0079 ISSN 2092-805X (Electronic)A Survey on Automatic Twitter Event Summarization Dwijen Rudrapal*, Amitava Das**, and Baby Bhattacharya***
10. Available online at www.sciencedirect.com 1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of the Organizing Committee of SLTU 2016 doi: 10.1016/j.procs.2016.04.054 ScienceDirect 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016 9-12 May 2016, Yogyakarta, Indonesia Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement AlgorithmMeganingrum Arista Jiwanggi*, Mirna Adriani
11. © The British Computer Society 2013. All rights reserved. For Permissions, please email: journals.permissions@oup.com Advance Access publication on 8 October 2013 doi:10.1093/comjnl/bxt109 Summarization of Twitter Microblogs Beaux P. Sharifi1, David I. Inouye2 and Jugal K. Kalita1,∗ 1Department of Computer Science, University of Colorado, Colorado Springs, CO 80918, USA 2Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA ∗Corresponding author: jkalita@uccs.edu
12. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 3, MARCH 2013 Automatic Twitter Topic Summarization With Speech Acts Renxian Zhang, Wenjie Li, Dehong Gao, and You Ouyang
13. Faculty of Computer Science, Universitas Indonesia, Depok, West Java, IndonesiaSee discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/228942438 Automatic Summarization of Twitter Topics Article · January 2010