An Extractive Text Summarization approach for Analyzing Educational Institution's Review and Feedback Data

Jai Prakash Verma
Assistant Professor
Institute of Technology,
Nirma University, Ahmedabad

Atul Patel, PhD
Professor & Dean
CMPICA
CHARUSAT University, Changa, Nadiad

ABSTRACT
Big Data analytics helps the enterprises and institutions to understand and identify the usability of large amount of data generated by their routine operations. All most third forth part of these types of data is semi-structured text data. Many types of actionable insights can be found from these type of semi-structured text data that can help strategic management for making right decision. In this paper we are proposing a recommendation system model for understanding and finding actionable insight from the large amount of text data generated for an educational institution. Here we are discussing different type of data generation sources of this types of data as well as data cleaning processes required. The wordcloud for an educational institution is published that help strategic management for understanding the sentiment of different stack holders mainly students. Herewith we are identifying different types of findings from these sets of words that helps for betterment in the functionaries of an Educational Institution.

Keywords
Big Data, Big Data Analytics, Extractive text summarization, Educational Word Cloud, Text Analytics, Hadoop Framework Applications.

1. INTRODUCTION
Uses of internet for social connectivity and communication generating large amount of text. People are writing text for their feeling, thought, idea and reactions for an event or particular. This types of text streaming raises the problem of storage, retrieval, and analysis, which is consider under Big Data problem [1, 2, 3]. Text summarization [4, 7] can help for automatic sensing of these large amount of text. In text summarization computer programs can generate a summary of large text document that provide the sense and understanding of the document in a small set of text. There are two approaches for text summarization abstractive text summarization and extractive text summarization. Abstractive text summarization is useful for the applications like sensing small text like books chapter, research papers, answers typed by students etc. This approach consist of the understanding or sensing of the large text and re-writing in fewer words that may or may not part of the original text. Extractive text summarization is useful to summarize a large amount of text like social media text, discussion forums, text documents for a specific domains etc. This approach is based on statistical and linguistic analysis of key words, word frequency etc. that are extracted sentence, paragraph, and document wise from the large amount of text dataset. In Big Data Analytics extractive text summarization approach is more useful then abstractive approach [9, 10].

There are two types of application where extractive text summarization approach may applied, first is generic summarization based application and second is query-based summarization applications. In generic summarization based application large amount of text is converted in a small set of words based on some statistical methods. In query-based text summarization applications text are summarized based on domain specific queries. Text are selected based on some specifics business needs for an enterprise or an institution [11, 12, 13].

In normal circumstances, academic institutes usually focusing on structured feedback of students; ignoring other stakeholders like alumni, parents, organization where they work etc. may not be appropriate at all for betterment. As we know that during chatting / sharing with friends, people usually express their own feelings / correct opinion. So, in place of structured approach, it is better to have unstructured information too. This is why our attention towards Big data because this types of unstructured datasets are generated every year in a large volume. In our previously published paper [15] we a preliminary analysis have been done in this area but as per big data concern more analysis is required. In this paper we are analyzing the sentiment of different stockholders of an educational institution based on feedback and reviews. Here we are extracting text data from different types of feedback systems and text written on websites like different types of online blogs, Facebook, twitter etc.

In section- II we are defining the Big Data and Big Data Analytics. And also discussing different issues for Big Data Analytics for an educational institution. In Section-III we are describing the extractive text summarization approach for Big Data. In section-IV we are telling about the hardware and software setup selected for experimental analysis. In section-V we are proposing a model of recommendation system based on extractive text summarization approach. In section –VI we are presenting experimental analysis steps that is executed for proposed system.

2. BIG DATA AND BIG DATA ANALYTICS
Today's era is for analyzing large dataset and finding knowledge form it for strategic to make right decision. But due to localization of data that knowledge may not effective or some time leads a wrong decision. To overcome this localization issue the concept of Big Data arises [5, 6, 8]. Big data covers large volume of data containing of variety of data types and generating in a large velocity. These three V's are defining this type of data for all the domain to consider for discovering knowledge for making strategic decision to improve business processes and enhance profit. Big data analytics uncover hidden pattern, unknown correlation, market

trends, customer preferences and behavior, and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. Domain specific Big Data analytics add one more 'V' in these three V's of Big Data Analytics dimensions. This 'V' for Value dimension- the final actionable insight and functional knowledge. Educational data domain covers all the types of data generated by the different types of functionaries require to run the institution. Big data analytics for educational data domain help to improve the satisfaction level of different stack holders like student, teacher, parents, governing bodies and society [6,8].

3. EXTRACTIVE TEXT SUMMARIZATION BASED RECOMMENDATION SYSTEM

Extractive text summarization is an automated system that extract text form the entire collection of data called Big Data without modifying the meaning and reference of data. As per the characteristics of Big Data, extractive text summarization is most suited approach for summarizing text and generating knowledge. This types of actionable insight and functional knowledge can be used for recommendation of a service or product [2, 4, 5].

4. HADDOP FRAMEWORK

Hadoop is a Java based open source programming framework that allows storage and analysis of large volumes of structured, semi-structured, and unstructured data. Hadoop enables processing of Big Data in a distributed computing environment built from commodity hardware. Hadoop is part of the Apache project funded by the Apache Software Foundation [6]. For experimental analysis single node Hadoop file system is implemented with Ubuntu operating system. Here we are using Java word count program on MapReduce frame work for counting word frequencies. We are using Python programming packages for data preprocessing steps for clean and improve quality of text data. Spark framework on Hadoop data storage is used for finding frequent word sets to analyze sentiment of reviews and feedback [16].

5. PROPOSED MODEL

As per Figure 1, we are proposing a recommendation system model for finding actionable insight from the text data extracted from different web pages and from different review or feedback systems. Following are the steps for the proposed model. Data Extraction or Collection (Big Data): Large amount of text data available on the web about the functionaries of an educational institution. Because of data size, different data types and data generation speed these huge amount of text data come under a Big Data problem. Dataset Selection: In this step we select the data that can be used for finding knowledge from available large dataset. Data Preprocessing: In this step we clean the selected dataset that can be used for finding valuable or actionable insights. Methodology: In this step we are selecting a data analysis technique or algorithm for analyzing this selected and cleaned dataset for finding actionable insight. Also identify some visualization technique. Findings and Discussion (Recommendation): In this step we find knowledge and actionable insight that can be used to help in correct decision making.

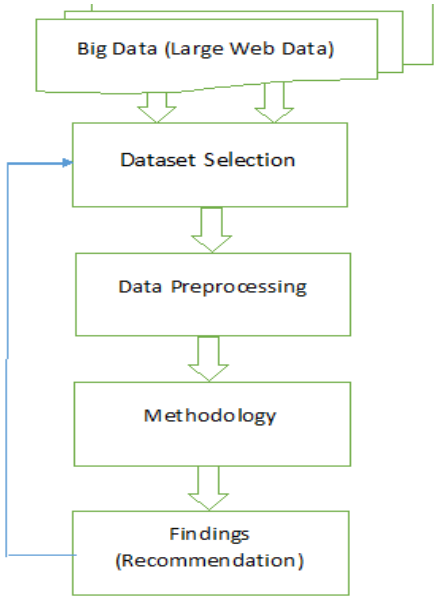

Figure1: Process Flow for Proposed Model

6. EXPRIMANTAL CASE STUDY

For the experimental analysis we are analyzing the text data collected from feedback system and different web pages like Facebook, Twitter, and social media blogs. Online feedback system is designed for generating student feedback and review about the different functionaries of an institution. This system provide the facility to the students to fill a review/feedback form in which student can write their comments and views in the form of text. In this textbox student or reviewer can write whatever the feel and realize about the institution with their hidden identity. There are many issues raise about the language, grammatical and spelling mistakes, spam reviews, and relevancy of the text. As well as some domain specific issues like some of the text have different meaning. For example meaning of "placement" in educational text is job and training opportunity not a location. In following data analysis experiment we are trying to resolve all these types of issues.

6.1 Dataset selection

For the experimental study data extracted from different social media webpages like Facebook, twitter, and discussion forums. As well as these types of text generated by different types of feedback system for different stack holders of the educational institution.

6.2 Characteristics of Dataset

Hadoop file system [6] was implemented for storage and management of this large text dataset. MapReduce programming with Hadoop framework was used for counting the words and word frequencies. Table - 1 gives information about the selected text data. Here we are collecting last years data from different feedback systems and some text extracted from institutes Facebook and Twitter web pages. Python script is used for extracting text from these web pages.

Table 1: Characteristics of Dataset

Dataset File Name	Feedback.txt	Reviews.txt
File Description	Test dataset generated from different feedback systems	Test data extracted from institute web pages like Facebook and twitter
Size	342KB	

Number of Word	15059
Number of lines	4038
Word frequencies file Name	Feedback_Review_Word_Frequency.txt
Total Dataset	TotalWord.txt

6.3 Data Preprocessing Steps

There is many possibility that this types of data are not cleaned. For correcting spelling mistakes and removing irrelevant text form the selected text dataset, following data preprocessing steps are executed using python scripting language.

a. Remove Punctuations from the dataset:

```
>>> import string
>>> string.punctuation
   '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
>>> newString = myString.translate(string)
```

b. Converting all number in lowercase:

```
fh = open('TotalWord2.txt','w')
with open('TotalWords.txt', 'r') as fileinput:
    for line in fileinput:
        line = line.lower()
        fh.write(line)
fh.close()
```

c. Remove Numbers from the dataset:

```
with open('TotalWord2.txt', 'r') as f1,
open('TotalWord3.txt', 'w') as f2:
    f2.write("".join([c for c in f1.read() if not c.isdigit()]))
```

d. List of published stop word with NLTK package of python package:

```
>>> import nltk
>>> from nltk.corpus import stopwords
>>> print set(stopwords.words('english'))
```

```
set([u'all', u'just', u'being', u'over', u'both', u'through',
u'yourselves', u'its', u'before', u'herself', u'had', u'should',
u'to', u'only', u'under', u'ours', u'has', u'do', u'them', u'his',
u'very', u'they', u'not', u'during', u'now', u'him', u'nor',
u'did', u'this', u'she', u'each', u'further', u'where', u'few',
u'because', u'doing', u'some', u'are', u'our', u'ourselves',
u'out', u'what', u'for', u'while', u'does', u'above', u'between',
u't', u'be', u'we', u'who', u'were', u'here', u'hers', u'by', u'on',
u'about', u'of', u'against', u's', u'or', u'own', u'into',
u'yourself', u'down', u'your', u'from', u'her', u'their',
u'there', u'been', u'whom', u'too', u'themselves', u'was',
u'until', u'more', u'himself', u'that', u'but', u'don', u'with',
u'than', u'those', u'he', u'me', u'myself', u'these', u'up',
u'will', u'below', u'can', u'theirs', u'my', u'and', u'then', u'is',
u'am', u'it', u'an', u'as', u'itself', u'at', u'have', u'in', u'any',
u'if', u'again', u'no', u'when', u'same', u'how', u'other',
u'which', u'you', u'after', u'most', u'such', u'why', u'a', u'off',
u'i', u'yours', u'so', u'the', u'having', u'once'])
>>>
```

e. Remove stop words from the dataset:

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import codecs
import nltk
stopset = set(stopwords.words('english'))
writeFile = codecs.open("TotalWord4.txt", "w",
encoding='utf-8')
with codecs.open("TotalWord3.txt", "r",
encoding='utf-8') as f:
    line = f.read()
    tokens = nltk.word_tokenize(line)
    tokens = [w for w in tokens if not w in stopset]
    for token in tokens:
        writeFile.write(token)
        writeFile.write('\n')
```

f. Correcting Spelling of word for English language: Bayesian classification technique is implemented in Python for find nearly correct word.

g. Converting all plural word to singular word: Adding following code in the above file

```
from pattern.en import singularize
token = singularize(token)
```

6.4 Findings

Figure-2 shows the WordCloud for student suggestion, feedback and comments about an educational institution.

Figure 2: Word cloud for student feedback

Table -2 shows the top 100 frequent word generally used by student whenever they submit feedback and review about an educational institution.

Table - 2 : Feedback and review word with their frequency

Word	Frequency	Word	Frequency
student	340	study	20
placement	159	increase	20
attendance	107	improvement	20
good	91	fee	19
facility	88	compulsory	19
subject	87	class	19
improve	75	need	19

much	74	reduce	19
institute	73	thing	19
faculty	72	learning	18
Please	65	level	18
Nirma	62	final	18
Work	52	branch	18
Activity	48	thank	17
Industry	47	proper	17
comment	47	teach	17
Give	46	provided	16
project	44	water	16
practical	43	try	16
industrial	43	technical	16
quality	42	curriculum	16
college	40	environment	16
make	39	higher	15
sport	39	course	15
time	37	criterion	15
syllabus	36	day	14
year	36	current	14
research	36	assignment	14
company	33	problem	14
lab	32	made	14
smelter	31	person	14
university	30	related	14
education	30	room	13
focus	30	required	13
academic	29	poor	13
knowledge	29	drinking	13
teaching	29	development	13
exam	28	classroom	12
canteen	27	freedom	12
improved	27	future	12
training	27	also	12
system	26	add	12
experience	26	lack	12
change	26	increased	12
visit	25	elective	12
provide	24	keep	12
department	23	bit	12
camp	22	lot	12
support	22	theory	12
part	21	management	12

Figure 3: Here we used FPGrowth algorithm for finding frequent words used in reviews. FPGrowth algorithm for finding frequent itemset is implemented in Machine Learning library (MLLIB) with Spark. For generating text input file, all the text of reviews and feedback have been transformed in the form of unique occurrence of words generated in table-2. Figure-4 is showing the output of FPGrowth algorithm execution with the minimum support 2 and number of partition

```
from pyspark.mllib.fpm import FPGrowth
from pyspark import SparkContext, SparkConf

conf = SparkConf()
sc = SparkContext(conf=conf)

data = sc.textFile("hdfs://localhost:8020/user/jaiprakash/Review.txt")
transactions = data.map(lambda line: line.strip().split(' '))
model = FPGrowth.train(transactions, minSupport=0.2,
numPartitions=10)
result = model.freqItemsets().collect()

for fi in result:
        print(fi)
```

Figure 3: Python implementation of FPGrowth on Spark Framework with Hadoop

Frequently used words in reviews generated by FPGrowth on Spark framework with Hadoop

```
FreqItemset(items=[u'need'], freq=49)
FreqItemset(items=[u'student'], freq=156)
FreqItemset(items=[u'lab'], freq=47)
FreqItemset(items=[u'place'], freq=84)
FreqItemset(items=[u'place', u'student'], freq=66)
FreqItemset(items=[u'placement'], freq=79)
FreqItemset(items=[u'placement', u'place'], freq=79)
FreqItemset(items=[u'placement', u'place', u'student'], freq=62)
FreqItemset(items=[u'placement', u'student'], freq=62)
FreqItemset(items=[u'institute'], freq=43)
FreqItemset(items=[u'improve'], freq=77)
FreqItemset(items=[u'improve', u'place'], freq=43)
FreqItemset(items=[u'improve', u'student'], freq=63)
FreqItemset(items=[u'good'], freq=68)
FreqItemset(items=[u'good', u'student'], freq=52)
FreqItemset(items=[u'up'], freq=62)
FreqItemset(items=[u'up', u'student'], freq=50)
FreqItemset(items=[u'part'], freq=56)
FreqItemset(items=[u'part', u'student'], freq=48)
FreqItemset(items=[u'give'], freq=53)
FreqItemset(items=[u'give', u'student'], freq=47)
FreqItemset(items=[u'subject'], freq=51)
FreqItemset(items=[u'attendance'], freq=51)
```

Figure 4: Frequent words used in reviews Generated by FPGrowth Algorithm

6.5 Discussions

Figure 2 shows the word cloud for student feedback and reviews about an institute. This types of word cloud is not published so far for an educational institution. The management or governing bodies of an educational institution can use this type of visual word cloud to understand the sentiments of students. Word Cloud: word cloud is an image that contains all the words related to particular domain. In which the size and darkness of each word indicates its frequency or importance.

Herewith we are proposing word cloud for student feedback or review for an educational institution.

Table - 2 provides the list of the top 100 feedback and review related words that were used by students to explain satisfaction ratings along with their total frequency. These words reflect a wide spectrum of aspects related to the educational institute and student satisfaction.

1. As per the table student, placement have high frequent words, so these two are more considerable areas and subjects.
2. Other words like faculty, facilities, attendance also have higher frequent words in the feedback and reviews.

Figure- 4 shows the frequently used word sets which can be used to understand the sentiment of different stack holders. It shows that student and placement are frequent word set. It means placement is highest priority for students.

As per above analysis we have generated year-wise summary of text words. This types of large set of data can be used to analyzing the changes of sentiment with time dimension.

7. CONCLUSION AND FUTURE WORK

Big data analytics become a new research paradigm or challenge for computer automated industry and universities. We have seen very few applications or research for text summarization in the area of educational institutions that explore the capabilities of Big Data Analytics. This study applies text analytics for a large amount text extracted from different webpages and from the different type information system in the form of reviews and feedback given by different stack holders of an educational institution. It assess the quality of these data as well as identify actionable insight that help strategic management of the institution for making right decision.

Our paper is contributing to research in the area of big data analytics in several ways. First, this paper demonstrate the importance of big data analytics in identifying novel patterns of feedbacks, reviews, and comments written by different stack holders of an educational institution. While our experimental study is based on text data generated from different feedback/review systems, from specific website on a specific time period but they reflected the way the text data generated by the all involving people on different time. Second, we are proposing an extractive text summarization based recommendation system and its implementation steps using python with Hadoop framework. Third, in this paper we are proposing a word cloud and frequent word sets for an educational institution that can be used by the strategic management or governing bodies for making right decision for betterment of the system. Fourth, herewith we are proposing the uses of text corpus generated timely as future research in this work, which can be used for finding many types of actionable insight that can be used for better understanding for the different functionaries of an educational institution.

8. REFERENCES

[1] A. Ittoo, et al., Text analytics in industry: Challenges, desiderata and trends, Comput. Industry (2016), http://dx.doi.org/10.1016/j.compind.2015.12.001

[2] Zheng Xiang, Zvi Schwartz, John H. Gerdes Jr., Muzaffer Uysal, "What can big data and text analytics tell us about hotel guestexperience and satisfaction?", International Journal of Hospitality Management 44 (2015) 120–130 , 0278-4319/© 2014 Elsevier

[3] Venkat N. Gudivada, Dhana Rao, Vijay V. Raghavan, "Big Data Driven Natural Language Processing Research and Applications", Chapter 9, Handbook of Statistics, Vol. 33. http://dx.doi.org/10.1016/B978-0-444-63492-4.00009-5 © 2015 Elsevier B.V

[4] Suvarna D. Tembhurnikar, Nitin N. Patil, "Topic Detection using BNgram Method and Sentiment Analysis on Twitter Dataset", 978-1-4673-7231-2/15 ©2015 IEEE

[5] Weiyi Ge, Chang Liu, Shaoqian Zhang, and Xin Xu, "Summarizing Events from Massive News Reports on the Web", 2015 International Conference on Network and Information Systems for Computers, 978-1-4799-1843-0/15 © 2015 IEEE

[6] Jai Prakash Verma, Bankim Patel, and Atul Patel, "Big Data Analysis: Recommendation System with Hadoop Framework", 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 978-1-4799-6023-1/15 © 2015 IEEE

[7] LI Bing, Keith C.C. Chan, "A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data", 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, 978-1-4799-7881-6/14 © 2014 IEEE

[8] Yogesh Sankarasubramaniam, Krishnan Ramanathan, Subhankar Ghosh, "Text summarization using Wikipedia", Information Processing and Management 50 (2014) 443–461, 0306-4573/_ 2014 Elsevier

[9] Amir Gandomi∗, Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management 35 (2015) 137–144, 0268-4012/© 2014 The Authors. Published by Elsevier Ltd

[10] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti a, Rinaldo Lima a, Steven J. Simske b, Luciano Favaro", Assessing sentence scoring techniques for extractive text summarization", Expert Systems with Applications 40 (2013) 5755–5764, _ 2013 Elsevier

[11] Elena Lloret, Manuel Palomar, "Tackling redundancy in text summarization through different levels of language analysis", Computer Standards & Interfaces 35 (2013) 507–518, © 2012 Elsevier

[12] Tushar Ghorpade, Lata Ragha, "Hotel Reviews using NLP and Bayesian Classification", 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

[13] Derek Bridge, Paul Healy, "The GhostWriter-2.0 Case-Based Reasoning system for making content suggestions to the authors of product reviews", Knowledge-Based Systems 29 (2012) 93–103, © 2011 Elsevier

[14] Xintian Yang, Amol Ghoting, and Yiye Ruan, "A Framework for Summarizing and Analyzing Twitter Feeds, KDD'12, August 12–16, 2012, Beijing, China., Copyright 2012 ACM 978-1-4503-1462-6 /12/08"

[15] Jai Prakash Verma, Bankim Patel, and Atul Patel, "Web Mining: Opinion and Feedback Analysis for Educational Institutions", International Journal of Computer Applications (0975 – 8887) Volume 84 – No 6, December 2013

[16] (2016). Frequent Pattern Mining - spark.mllib, [Online]. Available: https://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html