IJETIE International Journal of Emerging Technology and Innovative Engineering
Volume I, Issue 3, March 2015
ISSN: 2394 - 6598
www.ijetie.org

Indian Political Data Analysis Using Rapid Miner

Dr. Siddhartha Ghosh
HOD, CSE, Keshav Memorial Institute of Technology, Hyderabad.
siddhartha@kmit.in

Jagadeeswari Chittiboina
MTech, CSE, Keshav Memorial Institute of Technology, Hyderabad.
jagadeeswaree@gmail.com

Shireen Fatima
MTech, CSE, Keshav Memorial Institute of Technology, Hyderabad.
c.shireenfatima@gmail.com

ABSTRACT

This thesis investigates the manifestos published during 16th General elections April/May 2014 of India by different political parties and correlates them with the best manifesto which promised for an overall growth of the nation.

As in India five years ago, there was a breaking news that it is a nation headed towards Superpower status. This promise has been belied. Growth has decreased from sub-10% to sub-5%. Corruption has eaten every vein of the nation. A nation which cannot provide basic healthcare, housing and food to its people cannot claim to be a "developed" nation.

With such quotes in its front page, a manifesto published in one of the leading news papers of India grabbed the attention of many citizens and parties of India. This manifesto when set as an agenda for the new government, growth rate is considered to increase.

Text mining and information extraction methods can tap immense amounts of valuable textual information available online. These methods when applied to political data can give beneficial results which can be helpful for analyzing political parties agendas.

Although there is hardly any field where Data mining is not applied, expecting a diamond to come out of the mine, in this paper, we as beginners have used the developed system to measure the performance of the prototype which is compared by taking into account 3 manifestos published by different National Parties and a manifesto that a leading news paper of India has published for the parties to study seriously.

Steps taken towards Data mining research work using Rapid miner in politics really gave us an opportunity to understand Data Mining and Indian politics in a better shape.

Keywords-: Data mining, Text Mining, Rapid Miner, text processing, Vector space model, cosine similarity.

1. INTRODUCTION

In this paper we have tried to apply data mining in politics because of the fact that Politicians woo the voters with so many promises which in most cases do not add to the growth rate of India. Hence this paper is an attempt to measure the distance/difference of the manifestos of the parties against the standard agenda which is said to improve the growth rate.

As on 12 March 2014, following is the count of political parties in India which are registered with the Election Commission of India [11]

Total Registered Parties	**: 1616**
National Parties	6
State Parties	47
Unrecognized Parties	1563

Each party on an average comes up with almost 100 promises. So on the whole 1616 * 100 = 161600 promises are given to the nation. Mining such huge data and bringing out the hidden information is the target. As a beginner effort, 30% of the work is done and presented in this paper how we can bring out the similarity or difference between the different agendas.

News articles related to political parties has always been a buzz in Indian democracy, a democracy which is considered to be the world's largest democracy.

"India is a rich country inhabited by poor people." i.e., it is a nation rich in culture as well as corruption, politics as well as poor, wealth also but not stored or utilized for the benefit of itself. It is a nation though rich in wealth **but**

- Cannot feed its people with full meal a day (the basic need of a nation to survive).
- Cannot give clothing to its people
- Cannot give houses to its people

Basically the three above needs are stated as "Roti (food) Kapda (clothing) aur (and) Makaan (Shelter)" locally and also these form the basic needs of every citizen in any country.

Heterogeneity in the Indian population causes division between different sections of people based on region, religion, caste, language and race. This is the cause for the rise of political parties with agendas catering to one or mix of these groups."Elections renew democracies" and only during elections each party either comes up with 1 or mix of the above needs as well as few other promises in their agenda. The other promises include:

- Women-safety
- Anti-corruption
- Inflation control
- Economic issues like poverty, unemployment, development, education, etc.
- Law and order issues like Terrorism, Naxalism, Religious violence, etc.

This agenda of such promises and views is synonymously called as MANIFESTO.

Every 5 years, elections are held in India to bring in a government which promises to bring overall growth of the nation and for a common or poor man it is the promise which quenches his thirst of "Roti (Food) Kapda (Clothing) aur (and) Makaan (Shelter)". 2014 is the year of elections again in India. After a long time, parties claiming their manifestos to be the best came into news grabbing the pulse. It has been a growing interest to everybody to know which type of manifesto can bring in growth of the nation in all perspectives. In this burning situation, one of the newspapers of India has published a manifesto which caters to the best needs of the nation to grow as whole and it also highlights the loop holes of the previous governments which needs correction. This agenda or manifesto in our thesis is set as a standard or benchmarking manifesto.

Off late, many national and regional parties came up with their manifestos too. This paper considers 3 manifestos of 3 national parties and text mining is done to see how best they are close to the standard ones.

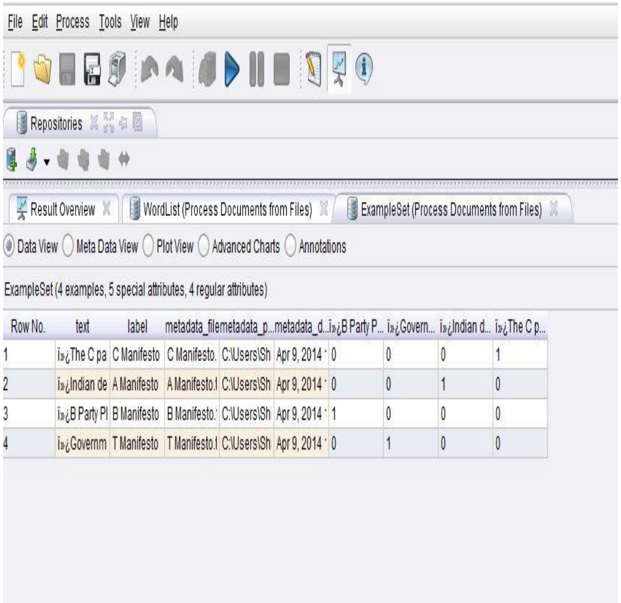
Text mining (Konchady, 2006) (Text mining: Wikipedia) refers to the process of deriving high quality information from text. High quality in text mining refers to some combination of relevance, novelty and interestingness.

RapidMiner, world's leading open source software serves the software solutions for data mining and extracting hidden information out of the bulk content. As we will see in the following, processes are created from many nested operators and finally represent the required information as graphs by using **Vector space model** and **Cosine similarity**. Each process in turn serves as a step to refine the bulk data and to mine, up to a level required to produce relevant hidden information. RapidMiner offers the analyst the possibility of defining break points and of therefore inspecting virtually every intermediate result. Successful operator combinations can be pooled into building blocks and are therefore available again in later processes.

2. METHODOLOGY :

The proposed system consists of 4 components:

- First component:



ExampleSet (4 examples, 5 special attributes, 4 regular attributes)

Row No.	text	label	metadata_file	metadata_p...	metadata_d...	ï»¿B Party P...	ï»¿Govern...	ï»¿Indian d...	ï»¿The C p...
1	ï»¿The C pa	C Manifesto	C Manifesto.	C:\Users\Sh	Apr 9, 2014 ·	0	0	0	1
2	ï»¿Indian de	A Manifesto	A Manifesto.I	C:\Users\Sh	Apr 9, 2014 ·	0	0	1	0
3	ï»¿B Party Pl	B Manifesto	B Manifesto:	C:\Users\Sh	Apr 9, 2014 ·	1	0	0	0
4	ï»¿Governm	T Manifesto	T Manifesto.I	C:\Users\Sh	Apr 9, 2014 ·	0	1	0	0

Figure 1: The manifestos gathered in Rapid miner

Gathers the 4 manifestos out of which one serves as the benchmarking manifesto published by a newspaper [1], the rest are the manifestos published recently by 3 renowned National parties claiming to be the

- Second Component:

Text mining usually involves the process of structuring the bulk input which is in the form of text and deriving patterns/features from the structured data.

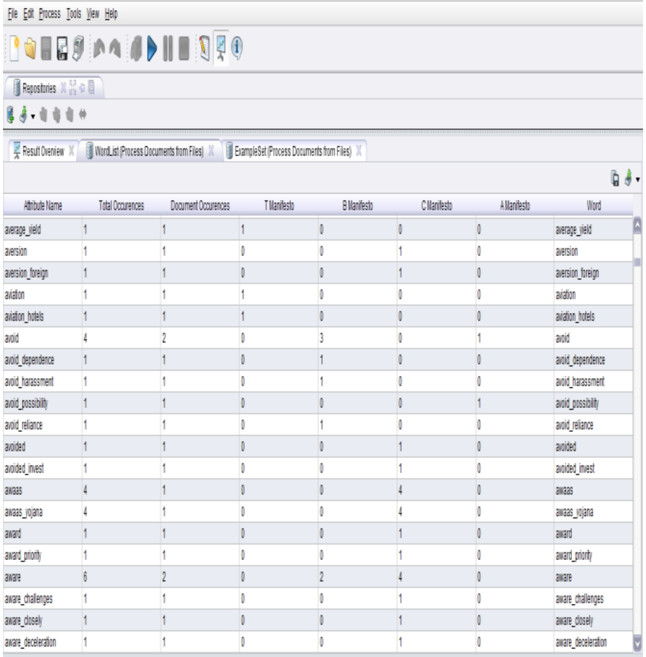
The relevant features of this thesis include the burning issues of the agendas taken into consideration. Deriving these features is done by following steps:

- Linguistic **Tokenizer** by words: Linguistic tokenizer by words is the process of splitting the text stream into words or phrases or other meaningful elements called tokens. For example, the tokenizing method used in this thesis would tokenize the following text string:

"India is a rich country inhabited by poor people".

Splitting this sentence on blank spaces would result in the following tokens

['India', 'is', 'a', 'rich', 'country', 'inhabited', 'by', 'poor', 'people'].


File Edit Process Tools View Help
Repositories
Result Overview
WordList (Process Documents from Files)
ExampleSet (Process Documents from Files)

Attribute Name	Total Occurences	Document Occurences	T Manifesto	B Manifesto	C Manifesto	A Manifesto	Word
average_yield	1	1	1	0	0	0	average_yield
aversion	1	1	0	0	1	0	aversion
aversion_foreign	1	1	0	0	1	0	aversion_foreign
aviation	1	1	1	0	0	0	aviation
aviation_hotels	1	1	1	0	0	0	aviation_hotels
avoid	4	2	0	3	0	1	avoid
avoid_dependence	1	1	0	1	0	0	avoid_dependence
avoid_harassment	1	1	0	1	0	0	avoid_harassment
avoid_possibility	1	1	0	0	0	1	avoid_possibility
avoid_reliance	1	1	0	1	0	0	avoid_reliance
avoided	1	1	0	0	1	0	avoided
avoided_invest	1	1	0	0	1	0	avoided_invest
awaas	4	1	0	0	4	0	awaas
awaas_yojana	4	1	0	0	4	0	awaas_yojana
award	1	1	0	0	1	0	award
award_priority	1	1	0	0	1	0	award_priority
aware	6	2	0	2	4	0	aware
aware_challenges	1	1	0	0	1	0	aware_challenges
aware_closely	1	1	0	0	1	0	aware_closely
aware_deceleration	1	1	0	0	1	0	aware_deceleration

Figure 2: Shows the tokenized text

- **Filtering** by stop words by length and by English words: Stop words are high frequency words that does not carry any significant information on their own. These words are often removed at the preprocessing stage to reduce the number of features. This thesis considers the words with 5-25 characters.
- **Stemming**: In linguistic morphology, stemming is the reduction of a word from its inflected form to its root, stem or base form. For example the word, 'walks' and another word "walking" will both have their word reduced to its root which is for both of them "walk".

Stemmer, stemming, stemmed -> stem.
Fishing, fishes, fished, fisher -> fish.

- **Transforming** all the text into lower case or upper case.
- **N-gram**: Feature sets made from unigrams are made of all the selected single words that are left after the documents pre processing steps.

- Third component:

Creation of vectors for each attribute/word in the documents with vector values ranging from 0-1.

Value 0 represents that the feature/attribute is missing in one or more manifestos

Value 1 represents that the feature is present in all the manifestos considered.

The following features are observed in all the 4 manifestos to be common issues:

- Women safety
- Health assurance
- Sports investment
- Inflation control
- Corruption- check
- Employment opportunities
- Education programs

The above issues are considered for a comparison on a graph as follows:

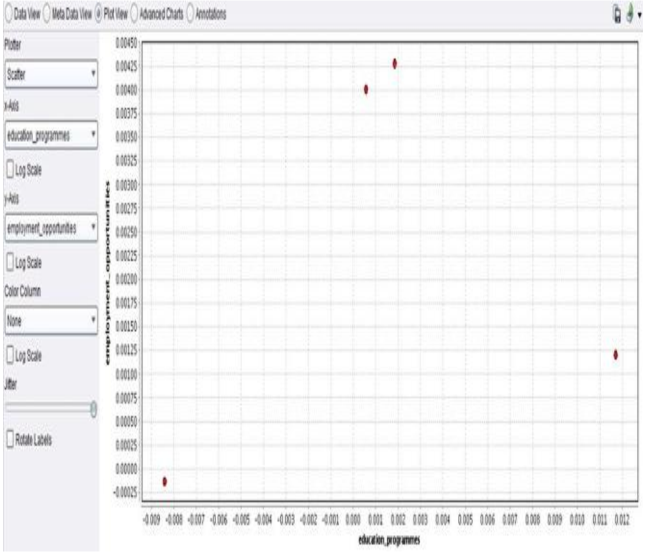


Figure 3: Education Programs Vs. Employment Opportunities

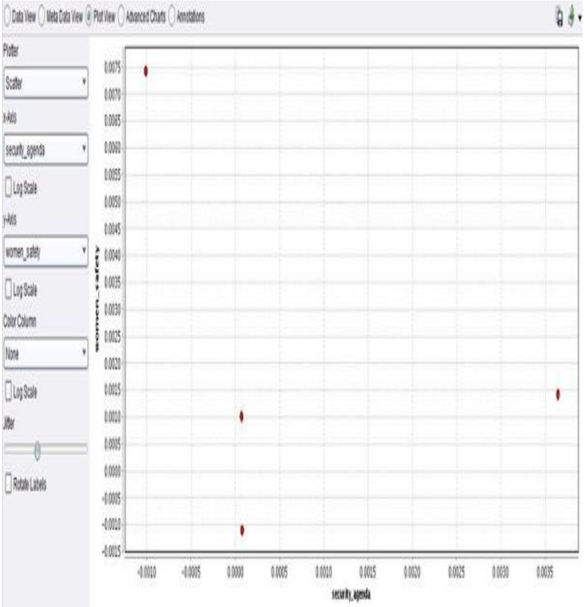
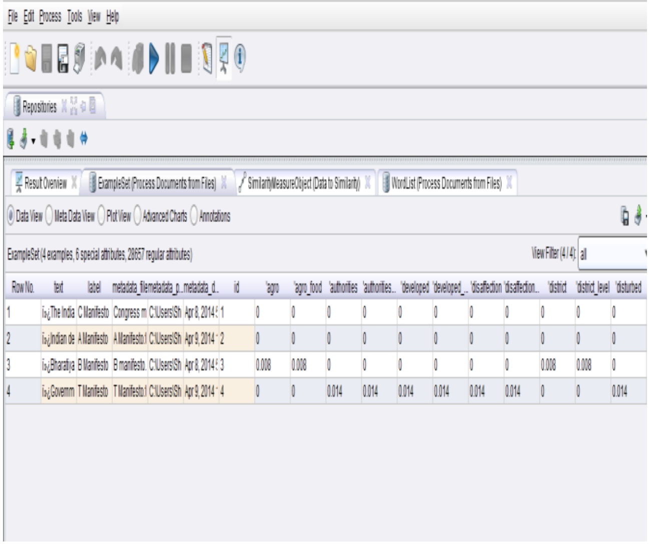


Figure 4: Women safety Vs. Security Agenda

- Fourth Component:

A Vector Space model is created for all the attributes in the example set segregated by the third component.

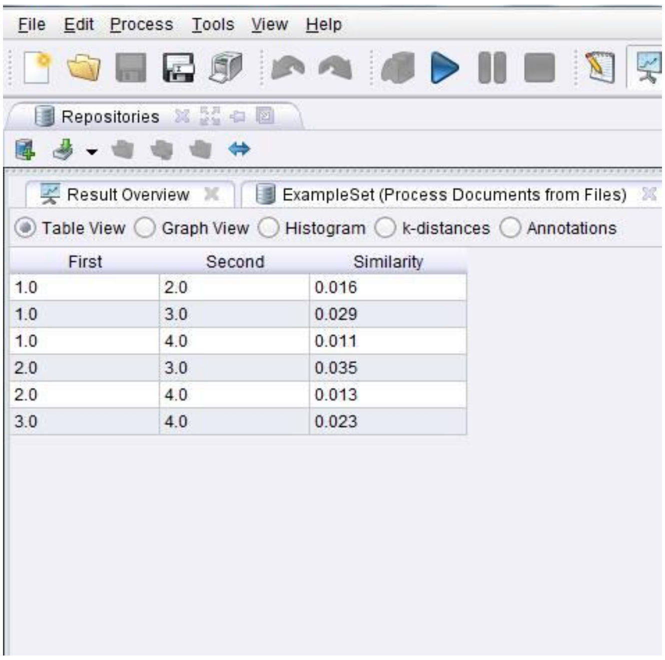

File Edit Process Tools View Help
Repositories
Result Overview
ExampleSet (Process Documents from Files)
SimilarityMeasureObject (Data to Similarity)
WordList (Process Documents from Files)
Data View
Meta Data View
Plot View
Advanced Charts
Annotations
View Filter (4 / 4): all

ExampleSet (4 examples, 6 special attributes, 28657 regular attributes)

Row No.	text	label	metadata_file	metadata_p...	metadata_d...	id	'agro	'agro_food	'authorities	'authorities...	'developed	'developed_...	'disaffection	'disaffection...	'district	'district_level	'disturbed
1	ï»¿The India	C Manifesto	Congress m	C:\Users\Sh	Apr 8, 2014	1	0	0	0	0	0	0	0	0	0	0	0
2	ï»¿Indian de	A Manifesto	A Manifesto.t	C:\Users\Sh	Apr 9, 2014	2	0	0	0	0	0	0	0	0	0	0	0
3	ï»¿Bharatiya	B Manifesto	B manifesto.	C:\Users\Sh	Apr 8, 2014	3	0.008	0.008	0	0	0	0	0	0	0.008	0.008	0
4	ï»¿Governm	T Manifesto	T Manifesto.t	C:\Users\Sh	Apr 9, 2014	4	0	0	0.014	0.014	0.014	0.014	0.014	0.014	0	0	0.014

Figure 5: Features represented as Vectors

Next the cosine similarity function is used which is used to measure the similarity between two document vectors. This function is applied to the example set to give the document similarity. The required results are loaded with similarity measure between the manifestos in tabular form. The above graph resembles nodes as follows:
1.0 representing C party's manifesto
2.0 representing A party's manifesto
3.0 representing B party's manifesto
4.0 representing the T manifesto which is published in the newspaper.



First	Second	Similarity
1.0	2.0	0.016
1.0	3.0	0.029
1.0	4.0	0.011
2.0	3.0	0.035
2.0	4.0	0.013
3.0	4.0	0.023

Figure 6: Tabular representation of the similarity measure of 4 document vectors

3. RESULT:

A graph is plotted between different features extracted to examine the similarity measure between different portfolios to the standard portfolio and this shows that B party's manifesto is close to the standard portfolio published in the newspaper. Hence is concluded that if B party wins the election, the growth rate of India would drastically improve.

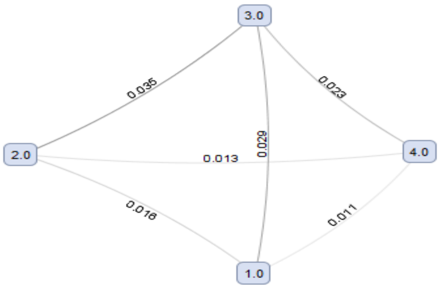


Figure 7: Graphical representation of the similarity measure.

4. Conclusion and Future work:

We described our work on a developed system capable of mining and extracting huge information from political data to measure the performance of similarity of the three national political parties' manifestos with the standard one which is said to increase the growth rate of the country. The results are shown by comparison of different documents of political manifestos using vector space model and shows similarity measure between them. This helps us to extract important features for each document vector and thereby examines the results in a graphical view. Benchmarking few documents with the standard document is done to examine which is the closest match. The completed work leaves many approaches for future work. Most immediate future work will focus on adding more text processing operators and algorithm for mining details of political articles. We believe that the work done in this paper provides interesting challenges to the students and research communities working on political data analysis.

5. REFERENCES

[i]http://timesofindia.indiatimes.com/home/lok-sabha-elections-2014/news/TOI-manifesto-An-agenda-for-the-new-government/articleshow/31973967.cms
[ii] http://bjpelectionmanifesto.com/
[iii] http://inc.in/manifesto/
[iv] http://www.aamaadmiparty.org/manifesto-2014
[v]http://articles.economictimes.indiatimes.com/2014-04-07/news/48939754_1_2014-elections-congress-party-lok-sabha-polls
[vi]http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/rapidminer-5.0-manual-english_v1.0.pdf

[vii] https://rapid-i.com/rapidforum/index.php?topic=1606.0
[viii] J. Han and M. Kamber. Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers, 2000.
[ix] Kim-Georg Aase : Text Mining of News Articles for Stock
Price Predictions
[10] George Forman : Feature selection for Text Classification
[11] http://en.wikipedia.org