Classification and Analysis of Online News Articles using NDTV

Namita Arun Amdalli[1] Santhosh Kumar K. L.[2] Jharna Majumdar[3]
[1]PG Student [2]Assistant Professor [3]Dean R&D, Professor & Head
[1,2,3]Department of Computer Science And Engineering
[1,2,3]Nitte Meenakshi Institute of Technology, Bengaluru, India

Abstract— News content is one of the most important factors that have influence on various sections. With the increase in the number of news it has got difficult for users to access news of their interest which makes it a necessity to categories news so that it could be easily accessed. By using data mining techniques, the news articles are categorized for easier navigation among articles. This will help users to access the news of their interest in real-time without wasting any time. Text Summarization reduces the effort and time consumption of the user instead of reading whole news article.

Key words: Data Mining, Extraction, Clustering, Classification, Text Summarization, Expectation Maximization, Random Fores

I. INTRODUCTION

Application of data mining system to determine patterns from the WWW is called web mining. As the term says, this is data grouped by mining the web. It utilizes the programmed tools to expose and extract data from servers, web accounts, and it certificates establishments to get to from organized and unstructured information from browser actions, server logs, website and link structure, page content and different sources.

Web mining makes practice of the data mining techniques discover and extract information from Web documents and services. Mining mainly consist below three widely used classes are,

- Web browser action tracking, Web activity, from server logs.
- Web graph, links among pages, people and different data.
- Data find in webpage and also in document are specified in Web content.

Data Mining is also known as Knowledge Discovery in Data (KDD). Most of the data or information on WWW is published as HTML pages and with the development of the web; the number of the pages are continuously increasing. In order to utilize web information for a better purpose, the technologies that can be used for retrieval of information from the web, categorization of web pages and so on. There are different types of news channels which are available online like NDTV, BBC, Times of India, The Hindu etc. which provides news bulletin and lot of information.

Categorization refers to grouping that allows easier navigation among articles. Internet news needs to be divided into categories. This will help users to access the news of their interest in real-time without wasting any time. When it comes to news it is much difficult to classify as news are continuously appearing that need to be processed and those news could be never-seen-before and could fall in a new category.

Due to the fact that there exists difficulties of data excess, fetch to sound and presently developed summaries is necessary. Text summarization can be said as the finest exciting job in information recovery. Lessening of data benefits a user to discover essential information speedily without having to waste time and energy in understanding the entire text collection.

II. LITERATURE SURVEY

In [1], focus on news categorization of the news documents. Consistently, tremendous measure of Bangla news articles are created by many different websites and the rate is also increasing. There are many forms of the online newspaper. Electronic edition is the one form of the printed newspaper. Online publication is related to daily publication; categorization is not in order in respect of content as well as in respect of layout. News website is the second form of the online newspaper, it enables user to browse in the menu which are ordered in topic groups and subgroups. Users read the news via connected to internet. Many readers are keen in reading the news from variety of sources and sites. Most of the time readers read the news of the categories of their interest. Hence the users have to go through all the sites to seek news articles of their importance. For Instance a user who has interest in sports needs to undergo all of the websites to get the information so rather a user would select a system that can collect news articles from diverse bases anytime and anywhere and also from mobile reading device The algorithm which is used here for the text categorization job is the Naive Bayes classifier .The basis for it is bayes theorem. The hint behind this is to make usage of the joint probabilities of words and categories to make estimation the probabilities of categories given a text document.

In [2], focus on cluster mining of documents. Text data is available wherever on the Web, in different forms like enterprise information systems, digital documents and in personal files. Analysis and handling the text data is very important as the number of the text data is increasing exponentially. The technology developed to handle the growing volumes of the text data is text mining. One of the data mining techniques called clustering is normally in use for creating clusters from large amount of unstructured data sources which is the non-numerical data. This technique is used in the data mining problems such as to build relations from a complex dataset, to find associations between the objects and to make generalizations. A data mining technique called text clustering can be used along with ontology to group the similar digital data which enhances the clustering process. Data Mining will be used to extract the similar news articles from the web and cluster them on the basis of concept weight and similarity measures. Combining text clustering and web ontology is an emerging area of research. Ontology is a domain which acts as a knowledge base and can be used to calculate the semantics between the concepts or terms and can be used with clustering.

In [3], focus is on the techniques used in Geographic information system (GIS) system. A structure which is skilled of pick up, storage, manipulate, studying, and envisioning all

kinds of geographic information is called Geographic information system (GIS). The GIS acronym similarly means geographic data science or geospatial data studies which speak of the educational discipline or profession at work by global information systems. This agrees users to form communicating questions, examines spatial data, oversee information in plans, offer spatial functions, actions and show the outcomes of each and every operations in an operational manner. An iterative technique known as Expectation-Maximization is utilized for discovering most extreme probability estimate of the factors in numerical prototypes, wherever the information is inadequate or else dependent on unnoticed unseen variables. This algorithm primarily gives arbitrary data to every parameter. Then this one will alternate among two stages, termed first step E is termed expectation step and second step M is termed as maximization step.

- First step E will compute a module intended for the expectation of the log-likelihood calculated by the recent estimation which are in use for the factors.
- Second step M this one again re-estimates factors maximize the expected log-likelihood initiated in first step E.

We now iterate both steps i.e. first step E and second step M up to the likelihood joins. While make use in clustering, this algorithm discovers clusters via defining factors of a numerical dispersal ideal which fits data set.

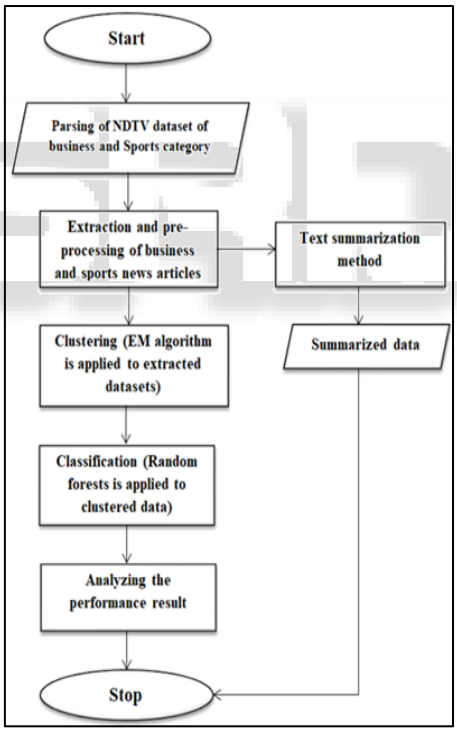
The text file gathering is transformed towards text file word matrix in document clustering. Matrix records could be binary, term frequencies value which shows the existence of word in text file.

In [4], aims at providing a short summary for comment on social network sites. Reducing text or any content to one-third or one-quarter its original size, clearly indicating its meaning, and retaining main thoughts expressed is known as summarization. The purpose is to concisely present the key points of any content in order to provide proper context for user. This summarization is helpful in various kinds of writing and at diverse points in the writing practice. Summarization features such as it provides environment for a paper's theory, compose literature analyses, and interpret a appendix. Benefits of summarization is it lets the reader to contextualize what people are saying, which is very vital in case of huge amount of social media contents generated everyday. It also aids the user to gain a better sense of what exactly the information or content is conveying. Summarization is usually produced by two key types of practices, called extraction and abstraction. Extractive summary is finding relevant sentences that belong to the summary. Abstractive summarization is finding or paraphrasing sections of the content to be summarized. Extractive summarization mines significant data, like sentences, by the input subjects and "puts them together" to formulate precise. Though summaries created in extractive summarization method might lack in consistency, but still extractive methods remain now-a-days as they are of little price and easy for applying to common areas.

In [5] focus is on review of algorithm of classification and regression. In recent few years there is much concern in "ensemble learning" which means approaches that create several classifiers then combine their outcomes. Weighted vote is engaged for calculation. The following trees does not dependant on previous trees in bagging— every tree is individually built by means of bootstrap. Modest mainstream poll is engaged in extrapolation. In the standard trees, every node has been divided by the finest split amongst all variables. In random forest, every node has been divided by the finest amongst a subsection of predictors arbitrarily selected by that node.

III. METHODOLOGY

The news articles are extracted from NDTV website for business and sports category. EM algorithm is used for clustering the extracted data, it forms various clusters for business category like (Sensex news, Banking news, Economy news and Other news) and for sports category like (Cricket news, Tennis news, Football news and other news). It is much difficult to classify as news are continuously appearing that need to be processed and those news could be never-seen-before and could fall in a new category. Random Forest algorithm is applied to clustered data, which analyses and classifies the keywords into either business or sports category. Instead of reading the full news articles text summarization makes short summaries of the extracted news articles, which saves time and effort of the news reader.



Fig, 1: Methodology of Proposed System

A. Objective

The objective of the proposed work is to cluster news articles into business and sports categories respectively. The clustered data is further classified into business or sports keywords. The news articles are summarized into shorter text.

1) Extraction

Online connection is established to the webpage of NDTV news categories for business and sports categories using jsoup. The links are as follows: businessURL = "http://profit.ndtv.com/news/latest/page-", sportsURL = "https://sports.ndtv.com/?pfrom=home-sports";

- Input: URL link of NDTV news articles of business and sports category Initially new connection is created to the webpage of NDTV news categories for business and sports using jsoup. Then the html document is fetched, parsed, and finds data within it (screen scraping). The class name of the element is specified to get the required HTML document.
- Output: Extracted data is obtained and stored in CSV file.

2) Clustering

Clustering is the method of grouping the information into numerous clusters. Grouping is accomplished by finding similar characteristics in the actual information.

a) EM Algorithm:

EM algorithm is a recursive way of discovering maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models

This algorithm primarily gives arbitrary data to every parameter. Then this one will alternate among two stages, termed first step E is termed expectation step and second step M is termed as maximization step.

- First step E will compute a module intended for the expectation of the log-likelihood calculated by the recent estimation which are in use for the factors.
- Second step M this one again re-estimates factors maximize the expected log-likelihood initiated in first step E.

Given the statistical ideal makes two sets, one set M which will have observed records, another set N which will have unobserved records also array of unidentified factors, along through a likelihood function $P(\theta{:}M,N) = q(M,N|\theta)$ The "maximum likelihood estimation (MLE)" of unidentified factors is determined by the "marginal likelihood" of observed records:

$$P(\theta; M) = q(M|\theta) = \int q(M, N|\theta) dN$$

This algorithm iterates both stages i.e. first step E and second step M up to the likelihood joins.

First step E: Estimate the expected value of the log likelihood function, with reference to uncertain dissemination of N when M is known beneath the present estimate of the factors: θ^t

$$Qi\left(\theta\middle|\theta^{(ti)}\right) = Ei_{N|M,\theta^{(t)}}[\log P(\theta;M,N)]$$

Second step M: Determine the factor which will maximize the measure:

$$\theta^{(t+1)} = \text{argmax}\ Qi\left(\theta\middle|\theta^{(ti)}\right)$$

- Input: News article dataset of NDTV business and sports categories
- Output: Four clusters related to business category (Sensex news, Banking news, Economy news and Other news) are obtained.

Four clusters related to sports category (Cricket news, Tennis news, Football news and Other news) are obtained.

- *Steps: E-M involves calculating two steps.*
- Here expectation step aims at finding values of similarity of objects between news articles and dictionary words.
- The next step maximizes the likelihood of the existing similarities.

3) Classification

Classification is done for the data mining that assigns objects to the target categories. Clustered data is classified as either business or sports words with the help of dictionary words. Random Forest algorithm.

a) Random Forest algorithm

Random forest is a decision tree, it is mostly used in classification and regression problems**.** This will prepare model which will make predictions of the value of a targeted variable through learning simple decision rules concluded by data structures. In classification problems, it comprises of an arbitrary number of random trees which can be used to find out the last outcome.

Following is algorithm for random forest (together for classification also for regression)

1) From the original data take out i_{tree} instances of bootstrap.
2) For every samples of bootstrap, develop an unpruned classification or regression tree, by subsequent modifications: on every node, instead of selecting the greatest split amongst all predictors, arbitrarily model j_{try} of predictors then pick finest fragment amongst those literals.

The term Bagging is nothing but distinct example of random forest when the condition j_{try}= q, q is predictors's count.

3) Predicting tree through combining various predictions of i_{tree}. Error percentage estimation could be found, on basis of the training data, by the succeeding steps:

a) Predicting information which is not present in bootstrap sample, "out-of-bag" or else OOB, by the tree grownup through the bootstrap example on every repetition of bootstrap.
b) Sum OOB predictions. Determine error percentage, then say it as OOB estimation of error percentage.

- Input: The clustered data in CSV format
- Output: Analyses and classifies the keywords into business and sports category.
- Steps:

1) Firstly, it usages the Bagging (Bootstrap Aggregating) algorithm for generating arbitrary examples. Consider a dataset S1 with x rows and y columns, it will generate a new dataset S2 through sampling x circumstances randomly by replacing from original data. Nearly 1/3 of the rows from S1 are left out, well-known as Out of Bag(OOB) samples.
2) Then, the model trains on S2. OOB sample is in use for concluding unbiased approximation of the error.
3) Out of z columns, Z <<z columns are chosen at every node in the dataset randomly.
4) Numerous trees are grown-up and the final prediction is achieved by averaging or voting.
5) It compares the dictionary words with the dataset of clustered data and then classifies the keywords into their either business or sports category.

4) Dictionary words

For the ease of classification, 969 business keywords and 842 sports keywords are stored in two text files. Example: Business keywords are {account, commodity, stock, tax,…} and Sports keywords are {champion, finish, final, test, score,….}

5) Text Summarization

– Input: Extracted CSV files Extractive method is used for summarizing the text. It's a procedure of mining or gathering vital sentence or phrases from original documents and it presents that data as a short summary without changing original text.
– Output: Summarized text is obtained and stored in CSV file]
– Example: Sania Mirza reached her first Grand Slam semifinal season with Shuai Peng. It is best performance of Sania at the Majors as she fell in the third rounds at the French Open. She had started the season by teaming up with Barbora Strycova.
– Steps followed are:

1) Each word in the sentence is converted to lowercase:
sania mirza reached her first grand slam semifinal season with shuai peng. it is best performance of sania at the majors as she fell in the third rounds at the french open. she had started the season by teaming up with barbora strycova.
2) Frequency of each word is calculated:
sania -2, mirza-1, reached-1, her-1, first-1, grand-1, slam-1, semifinal-1, season-2, with-2, shuai-1, peng-1,it -1,is-1,best-1,performance-1, of-1, at-2,the-4, majors-1, as-1,she-2,fell-1,in-1,third-1,rounds-1,french-1,open-1,had-1, started-1, by-1, teaming-1, up-1, barbora-1, strycova-1
3) The sentence which has maximum number of words will be retained by considering the count of repeated word as one. In this case second sentence has maximum words and it has repeated words (the,at):
It is best performance of Sania at the Majors as she fell in the third rounds at the French Open.

IV. EXPERIMENTAL RESULTS

The extracted data of business webpage is stored in CSV file format. It contains five attributes (title, source, time, link, content) and 79 instances. The extracted data of sports webpage is stored in CSV file format. It contains four attributes (sport, title, link, content) and 22 instances. EM algorithm is applied on the extracted CSV files. It forms four clusters in each category. In business category the clusters are: Sensex, Banking, Economy and others. In sports category the clusters formed are: Cricket, Tennis, Football and others. On the clustered data Random forest algorithm is applied, it compares the dictionary words with the dataset of clustered data and then classifies the keywords into their specific category.

The graphical representation of business clusters and sports clusters are shown below in pie chart in Fig 2 and Fig 3 respectively:

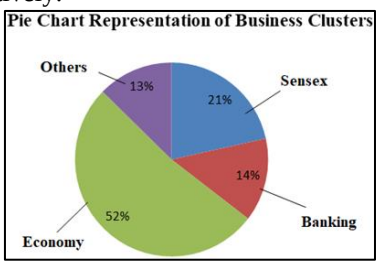


Fig. 2: Visualization of Business Clusters

In business dataset there are 79 instances, the instances are clustered into four different categories. Table 1 shows the number of instances and its result in percentage in each cluster of business category.

Clusters	Sensex	Banking	Economy	Others
Instances	17	11	41	10
Percentage Results	21%	14%	52%	13%

Table 1: Result of Business Clusters

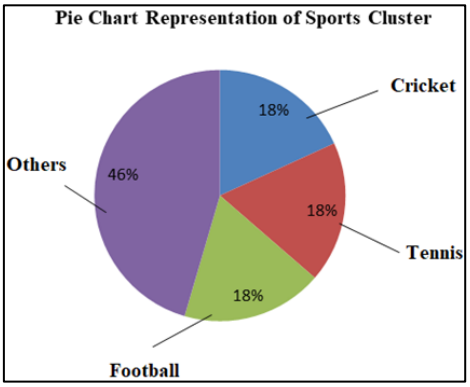


Fig. 3: Visualization of Sports Clusters

In sports dataset there are 22 instances, the instances are clustered into four different categories. Table 2 shows the number of instances and its result in percentage in each cluster of sports category.

Clusters	Cricket	Tennis	Football	Others
Instances	4	4	4	10
Percentage Results	18%	18%	18%	46%

Table 2: Result of Sports Clusters

The graphical representation of Classification of keywords is shown in pie chart in Fig 4:

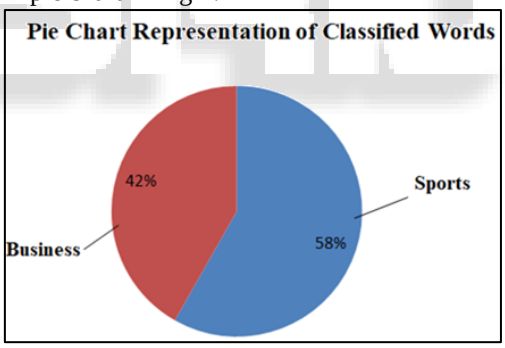


Fig 4: Visualization of News Classification

Table 3 shows the number of keywords matched and its result in percentage of each category:

Classification	Business Keywords	Sports Keywords
Keywords Matched	231	322
Percentage Results	42%	58%

Table 3: Result of News Classification

Word Cloud class shows the prominence of the individually classified word by highlighting words in diverse font styles. Fig 5 and Fig 6 shows the word cloud display of business and sports keywords.

Fig. 5: Word Cloud display of Business Keywords

V. CONCLUSION AND FUTURE WORK

Classification and Analysis of online news article is achieved using keyword comparison and Random forest algorithm. The NDTV dataset is base for classifying online news articles. Large number of news article is collected of two different categories. These news articles are processed, extracted and stored in database for keyword matching. The online news classification tool will classify the news article into specific news categories such as business and sports. The Random forest algorithm is used for evaluating and analysis of classifying news articles. Text summarization method condenses the news articles, it saves time and effort needed to read the full news articles.

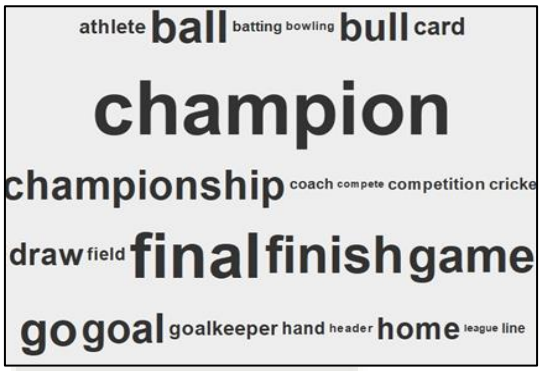


Fig. 6: Word Cloud display of Sports Keywords

The classification of news articles takes input only from NDTV news website because of its html source code formatting. To achieve this for different websites such as CNN, Fox News and Times Now etc, require more effort time and huge storage space. This can be achieved in future as extended version this application. The category covered in this projected is restricted up to two categories. In future at least ten plus categories can be covered. Overall the future work depends upon requirement of end user.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Prof N.R. Shetty, Advisor and Dr. H.C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out the research at NMIT.

The authors extend their thanks to Vision Group on Science and Technology(VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

REFERENCES

[1] Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das, "Bangla News Classification using Naive Bayes classifier", 16th International Conference Computer and Information Technology, March 2014
[2] Twinkle Svadas1, Jasmin Jha2, "Document Cluster Mining on Text Documents", International Journal of Computer Science and Mobile Computing, June 2015
[3] Chandan K. Reddy, Hsiao-Dong Chiang, Bala Rajaratnam, "Stability Region based Expectation Maximization for Model-based Clustering", Sixth International Conference on Data Mining , 2006
[4] Tayfun DOGDAS, Selim AKYOKUS, "Document Clustering using GIS Visualizing and EM Clustering Method", Innovations in Intelligent Systems and Applications, June 2013
[5] Ms. Pooja S. Choudhari, Prof. S. S. Nandgaonkar, "A Survey On Short Text Summarization Of Comment Streams On Social Network Sites", International Journal of Advanced Research in Computer Engineering & Technology, November 2015
[6] Deepali K. Gaikwad and C. Namrata Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering, March 2016
[7] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R Foundation for Statistical Computing, December 2002
[8] Yu Zhang, Mengdong Chen and Lianzhong Liu, "A Review on Text Mining", 6th IEEE International Conference on Software Engineering and Service Science, 2015
[9] R.C.Saritha, Annarao Kulkarni, "Text Clustering Using Mixture Models", Publications of Problems & Application in Engineering Research, 2013
[10] https://www.theguardian.com/business/glossary-business-terms-a-z-jargon
[11] http://www.enchantedlearning.com/wordlist/sports.shtml

