Methodologies used on News Articles: A Survey

Ashwin Kabra[1]

[1]*Student at Department of Computer Science and IT, Veermata Jijabai Technological Institute, Mumbai, India..*

Abstract - *News articles are the basic structured as well as unstructured and text formatted data, which lead it to become the point of interest to the researchers. Various researches like classification of news articles according to domain depending on region or the area of interest of people using various machine learning techniques, sentiment analysis over the news articles using the emotional dictionary, text summarization on the news data so as to reduce the bulkiness of data, Special Character recognition using the Natural Language Processing(NLP), Graphical representation of News articles using the visualization terms etc. Below paper gives a survey on various techniques used to perform such tasks on the News Articles and its data part.*

Key Words: Data Mining, Text Mining, Convolutional Neural Network(CNN), Machine Learning, Classification, Clustering, Sentiment Analysis, Natural Language Processing(NLP), Support Vector Machine(SVM), Naive-Bayes Algorithm.

1. INTRODUCTION

News Data is one of the structured as well as unstructured formatted data, which carries attributes like source, date, location, author, headline text, detailed data. To extract features from the textual data, now a day we prefer the Natural Language Processing approach, which deals with all of the language and text related aspects[7][8][9].

Natural Language Processing(NLP) is the technique to perform various lingual operations over the textual formatted data. NLP performs various operations like stop words removal, word count generations, features of textual data which are nothing but the words which are going to be used further for the task of performing various classification and other machine learning tasks[5].

Machine Learning is the new trending phenomenon now a day, as everyone needs somewhat type of intelligence and automation in their respective aspects, so lots of research is going on in this domain[4]. Machine Learning algorithms and techniques are mainly classified into two main classes i.e. Supervised and Unsupervised Techniques. Supervised techniques are the techniques which are dependent on class variables i.e. the input variables are relates only to the defined output variables e.g. Classification, whereas in unsupervised learning we have only input variables, we do not have pre-defined class for them e.g. Clustering[4].

On the News data Different researcher have performed different-different researches such as:

1) Classification

2) Clustering

3) Sentiment Analysis

4) Data Visualization

5) Text Summarization

The rest of the paper carries the methodologies performed by researchers considering their respective aspects and conclusion dependent on it.

2. Methodologies Used:

First part of any of the approach is gathering and pre-processing of data [7][8][9]. Before Performing any of the methodology the researches have performed somewhat type of Natural Language Processing operations so as to get the required features which are nothing but the words as the news data is a textual formatted data [9].

1.1 Classification:

Classification is the task of mapping the input the to the respective output variables. The classification is divided into two types i.e. binary classification (e.g. Yes/No) and multi-class classification. Various techniques and algorithms are used for the task of classification e.g. Decision Tree, C 4.5, J 48, Random Forest, Neural Networks, etc[1].

Different methodologies have been used for the purpose of classification on the news data, Machine Learning Techniques such as Naive Bayes, Support Vector Machine, Multi-layer Perceptron, Multi-Nomial Naive Bayes, Multi-layer Neural Network, CNN Networks, Classification Tree algorithms such as Decision Trees, Random Forest, Deep Learning approach for the classification purpose has been also used for the news articles according to the domain and particular region [7][8][9].

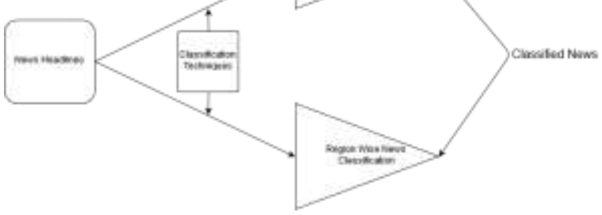
News Headlines
Classification Techniques
Domain Wise News Classification
Region Wise News Classification
Classified News

Fig -1: Classification of News Articles

1.2 Clustering

Clustering is the task of merging up the similar type of input elements into groups and different type into another group. In clustering approach, the main goal is to reduce the intra-cluster distance and increase the inter cluster distance. Various algorithms are been used for the task of clustering such as K-means, K-medoid, Hierarchical Clustering, etc. [3].

Clustering approach for the news headlines generates the groups which were having the same domain or the same region from which that news came. One of the researcher used MCL clustering approach for the clustering approach [10]. Another researcher used the graph based sequencing clustering approach for the purpose of clustering and then using those clusters he created the news retrieval system [11].

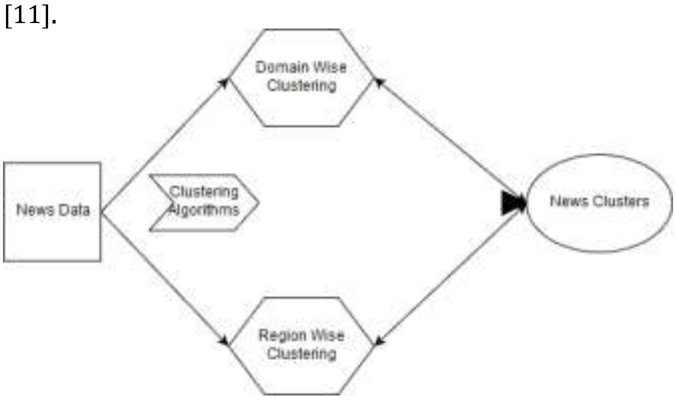

Fig -2: Clustering Approach for News Articles

1.3 Sentiment Analysis

Sentiment analysis is a task of generating opinion. Sentiment analysis takes various features as input and generates the opinion regarding various attributes and aspects e.g.{good-bad, yes- no} using the defined keyword or emotions defined [2]. Various methods such as the neural networks ,Support vector machine(SVM), Naive Bayes Algorithm are being used for the purpose of sentiment analysis [12][15][16].

sentiment analysis has been used for multiple purposes on news data, such as finding whether news should be read by kids or not [15], giving opinion on NEWS data using various emotional dictionary words will show what to read or what to not. One of the researcher have used the news data so as to find-out the customer behaviour and what he buys most using the sentiment analysis [14]. One of the researcher identified the various articles which were somewhat criminal or which were Terrorist activity related [16]. One of the researchers have used sentiment analysis for the measurement of credidibilty of news articles, so that a person will decide whether to read that news or not [22].

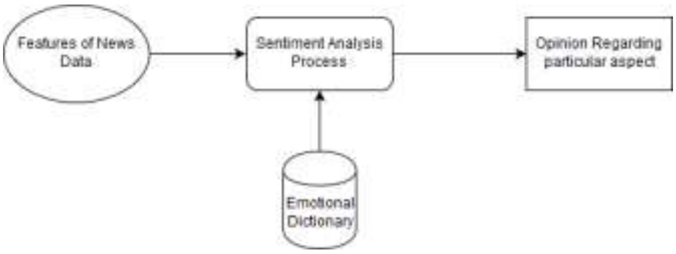
Features of News Data
Sentiment Analysis Process
Opinion Regarding particular aspect
Emotional Dictionary

Fig -2: Sentiment Analysis for News Articles

1.4 Data Visualization

Data Visualization is the technique which is used for representing the data and required aspects in the diagrammatic and graphical format. Data Visualization gives the animated view of any situation or data-set by which person makes his decision and performs various imagery operation [6].

Data visualization on news articles gives the animated or the graphical representation of the news data [21]. One of the researchers used the graphical representation of news data so as to predict the future relationship among the past data [17].

1.5 Text Summarization

Text summarization is the process of reducing the size of textual information using the NLP and various machine learning techniques. Word stemming concept is mainly used for the task of text summarization [22].

As the News data contains bulky articles, some of the readers don't want to read those bulky articles so the text summarization helps them to understand the articles in brief manner. One of the author have used the process as they first converted the data into tokens, then removed the stop words and performed the word stemming on that. They converted all information letters into small letters and represented them using the mathematical count using the matrix representation and calculated the TF-IDF count and generated the summarized information [22].

3. CONCLUSIONS

News articles are the textual formatted data, which can be used as structured as well as in unstructured manner. Various researchers have contributed lots of researches in respect to news articles such as classification, clustering, sentiment analysis, text summarization, visualization. Each and every researcher used various algorithms which were either of machine learning or of their respective approach for the respective task and they got their results accordingly. For the task of calculation and accuracy measurement and calculating false rate they have used the Score matrices, TF-IDF count, Confusion matrices as per their requirement and need.

The future scope of this survey is to gain more and more information and knowledge regarding researches which were conducting on the news data, and using this approach and studying those approaches so as to understand news data and all intelligence related aspect to it in depth

REFERENCES

[1] Seema Sharma, Jitendra Agrawal, Shikha Agrawal, Sanjeev Sharma "Machine Learning Techniques For Data Mining: A survey" IEEE 2013.

[2] Zohreh Madhoushi, Abdul Razak Hamdan, Suhaila Zainudin "Sentiment Analysis Techniques in Recent Works" Science and Information Conference 2015, IEEE.

[3] Nisha, Puneet Jai Kaur "A Survey Of Clustering Techniques and Algorithms " 2015 IEEE.

[4] The Duy Bui, Duy Khuong Nguyen, Tien Dat Ngo " Supervising an Unsupervised Neural Network " 2009 First Asian Conference on Intelligent Information and Database System, 2009 IEEE.

[5] Pranav Kaushik, Akanksha Rai Sharma " Literature Survey of Stastical, Deep and Reinforcement Learning in Natural Language Processing" International Conference on Computing, Communication and Automation (ICCCA 2017) IEEE.

[6] Parke Godfrey, Jarek Gryz, and Piotr Lasek "Interactive Visualization Of Large Datasets" IEEE transaction on Knowledge and Data Engineering, Vol. 28, No. 8, August 2016.

[7] Vignesh Rao, Jayant Sachdev "A Machine Learning Approach to classify news Articles based on Location" Proceedings of the International Conference on Intelligent Sustainable Systems(ICISS 2017).

[8] David Cecchini, Lina "Chinese News Classification" 2018 IEEE International Conference on Big Data and Smart Computing.

[9] Tej Bahadur Shahi, Ashok Kumar Pant "Nepali News Classification Using Naïve-Bayes, Support Vector Machines and Neural Networks" 2018 International Conference on Communication, Information & Computing Technology(ICCICT), Feb 2-3, Mumbai, India.

[10] Alwan M Ubaidillah Al-Fath, Kemas Rahmat Saleh W., M.Eng, Siti Sa'adah, M.T. "Implementation of MCL algorithm in Clustering Digital News With Graph Representation" 2016 Fourth International Conference on Information and Communication Technologies(ICoICT).

[11] Deepa Nagalavi, M. Hanumanthappa "A New Graph based Sequence Clustering Approach for News Article Retrieval System" IEEE International Conference on Power, Control, Signals and Instrumentation Engineering(ICPCSI-2017).

[12] Victoria Bobichev, Olga Kanishcheva "Sentiment Analysis in the Ukraininan and Russian News" 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).

[13] ByungSoo Ko, Chanyong Park, Dongkeon Lee, Jaewon Kim, Ho-Jin Choi, Dongsoo Han "Finding News Articles Related to Posts in Social Media: The Need to Consider Emotion as a Feature" 2018 IEEE International Conference on Big Data and Smart Computing.

[14] Matthias W. UHL "Explaining U.S. Consumer Behaviour with News Sentiment" ACM Transaction on Management Information Systems, Vol.2, Article 9. June 2011.

[15] Bin WANG, Linli GAO, Tao AN, Mei MENG, Tong Zhang " A method of Educational News Classification Based on Emotional Dictionary " 2018 IEEE.

[16] Richard Mason, Brian McInnis, Siddhartha Dalal "Machine Learning for the Automatic Identification of Terrorist Incidents in Worldwide News Media" 2012 IEEE, ISI 2012 June 11-14, Washington D.C., USA

[17] Jingyuan Zhang, Chun-Ta Lu, Mianwei Zhou, Sihong Xie, Yi Chang, Philip S. Yu "HEER : Heterogeneous Graph Embedding for Emerging Relation Detection From News" 2016 IEEE.

[18] Urooj Mohiuddin, Hameeza Ahmed, Muhammad Ali Ismail "NEWSD: A Real Time News Classification Engine for Web Streaming Data" International Conference on Recent Advances in Computer Systems (RACS 2015).

[19] Wiphada Jirasirilerd, Pikulkaew Tangtisanon "Automatic labeling for Thai News Articles Based on Vector Representation of Documents ", 2018 IEEE.

[20] Banu Inanc, Uyan Dur "Analysis of Data Visualization in daily Newspapers in terms of Graphic Design" Science Direct Proceedings of Social and Behavioural Sciences (ARTSEDU 2012).

[21] James Fairbanks, Natalie Fitch, Nathan Knauf, Erica Briscoe "Credibility Assessment in the NEWS : Do We Need to read?" MIS 2, ACM 2018

[22] Claudiu Popescue, Lacrimioara Grama, Corneliu Rusu " Automatic Text Summarization by Mean-absolute Constrained Convex Optimization" 41st International Conference on Telecommunications and Signal Processing, 2018.

[23] "https://www.expertsystem.com/machine-learning-definition"