PANDEMIC EFFECT ANALYSIS BASED ON NEWSPAPER REPORTS

Term Paper Leading to Thesis

by

Arka Sengupta

MCSE 1st Year 2nd Semester

Roll – 001910502028

Department of Computer Science & Engineering,

Faculty of Engineering & Technology

Jadavpur University, Kolkata, India

1. Abstract:

COVID – 19 is a worldwide pandemic disease declared by World Health Organization originated from Wuhan province of People's Republic of China which impacted the regular livelihood of people of every community. In this work, we have proposed a way to find out how much it has impacted the daily life of people based on the newspaper reports which describes all the incidents as well as the situation of people of all the classes. We have used the e-paper version of 'The Telegraph' which is a leading newspaper in India for the period of 24th March to 30th June 2020 and by applying web scraping techniques we get the news articles in image format. Then we have applied Optical Character Recognition technique to get the articles in text format. Next we have used probabilistic approach for the selection of articles for their respective categories for each day. Each article of every category is then taken individually and POS Tagging is applied so that we can detect the proper words which impacted the overall meaning of the whole statement. Next we calculate the sentiment score and average them for each category for each day and observed it throughout the aforesaid period.

2. Keywords:

Web scrapping, optical character recognition, probabilistic approach, POS Tagging, Sentiment analysis

3. Introduction:

In Twenty First century, the most disastrous situation has occurred due to a virus named SARS - nCOV 2 aka. Novel Corona Virus. The infection rate of this virus is so high that approx 1000 healthy person can be affected from one person who is the carrier of the virus and the mortality rate is high enough to be declared as a global pandemic by World Health Organization. Approx 33 million people got infected across 130 countries out of which approx 1 million people died. All the governments of their respective countries has taken measure by applying nationwide full or partial lockdown and started awareness campaign as well as economic reforms for the benefit of their people. In this situation

there is a huge impact over common people as they are the worst sufferer and all such incident from people's new mandate to Government policies, from awareness campaign to economy etc. are formed in a short article and printed over newspaper such that people can have a perception about the situation due to this global pandemic. These articles have a huge impact over the society and our work is based upon how much society is influenced over all such articles. We have categorized the articles based upon the keywords present in the text in four ways such as Impact due to Death, Economy, Education and Public Health Awareness. We calculate the sentiment score of every such article and try to visualize the impact of newspaper over common people.

4. Previous Research Work:

Azriel Rosenfeld and Tapas Kanungo in their conference paper [15], has provided a detailed survey of algorithms of different approaches of document structure analysis and provided a comparative study along with limitations of segmentation based approach in page level where both physical and logical components can be described in ordered tree structure with the help of a tree grammar.

Hidenao Abe in his journal paper [4] focused on the temporal behaviour of the Twitter service known as "retweeting". Users' tagged retweets are affected by the content of the received tweets and their history of tweets. In order to predict such targeted tweeting behaviour of followers, a model is constructed, for which one should set up more proper features to consider the history of their tweets.

Zhao Jianqiang and Gui Xiaolin in their article [5], discussed the effects of six text pre-processing method on sentiment classification performance using feature models and classifiers on Twitter datasets. To identify the sentiment polarity, most existing approaches apply text pre-processing to reduce the amount of noise in the tweets. This improved the performance of the classifier and speeded up the classification process.

Pang B. and Lee L. have elaborated on sentiment classification of two types – one in terms of either objective or subjective - and the other categorized as positive, negative or neutral [6].

A. Pappu Rajan and S.P.Victor in [7] determined the positive or negative sentiment of text which extended to strength of polarity. This included data set collection, reading of opinion dataset, and removal of noisy data, splitting of opinion sentences into opinion word, and finally finding the positive and negative opinions. The actual number of positive and negative opinions from multiple sets are being compared here. Score of opinion is measured as the difference between the number of positive words and the number of negative words.

Shailendra Kumar Singh and Sanchita Paul in [8] stated that in sentiment analysis process, negation words and negative prefixes have potential to reverse the sentiment of sentences. Part-of-Speech (POS) tagging information and opinion words and phrases are used for sentiment extraction. The opinion words and opinion phrases are used to extract positive / negative sentiments. There are two approaches – one lexicon-based and the other statistical-based.

Anusha K S and Radhika A D in [9] discussed about the sentiment analysis of twitter data. This involves data collection, data pre-processing, feature extraction, sentiment analysis, and ultimately polarity classification into positive, negative and neutral.

5. Literature Survey:

5.1. Optical Character Recognition:

Optical Character Recognition, or OCR, is a technology that enables one to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.
Assuming there is a paper document – such as a magazine article, brochure, or a PDF contract obtained via email - a scanner is not sufficient to make this information available for editing on the editor software. All a scanner can do is create an image or a snapshot of the document that is nothing more than a collection of black and white or color dots, known as a raster image. In order to extract and repurpose text from scanned documents, camera images or image-only PDFs, what is needed is the OCR software that would single out letters on the image, put them into words and then - words into sentences, thus enabling one to access and edit the content of the original document.

5.2. Web Scraping:

Web Scraping is a technique of extracting relevant information from web pages throughout the internet. It is a data mining approach.
Relevant information regarding a topic can be found over internet in the form of web pages. We can parse this information by going to the structure of the page i.e. a HTML format page, where we can check the appropriate div block where the necessary information is kept either in text format or in the image file format.

5.3. Sentiment Analysis:

Sentiment Analysis is the computational study of people's opinions, appraisals, and emotions toward entities, events and their attributes.

Sentiment Analysis involves subjectivity analysis of a statement and then emotion identification that get expressed through the statement.

The process is depicted in the following figures 1 and 2

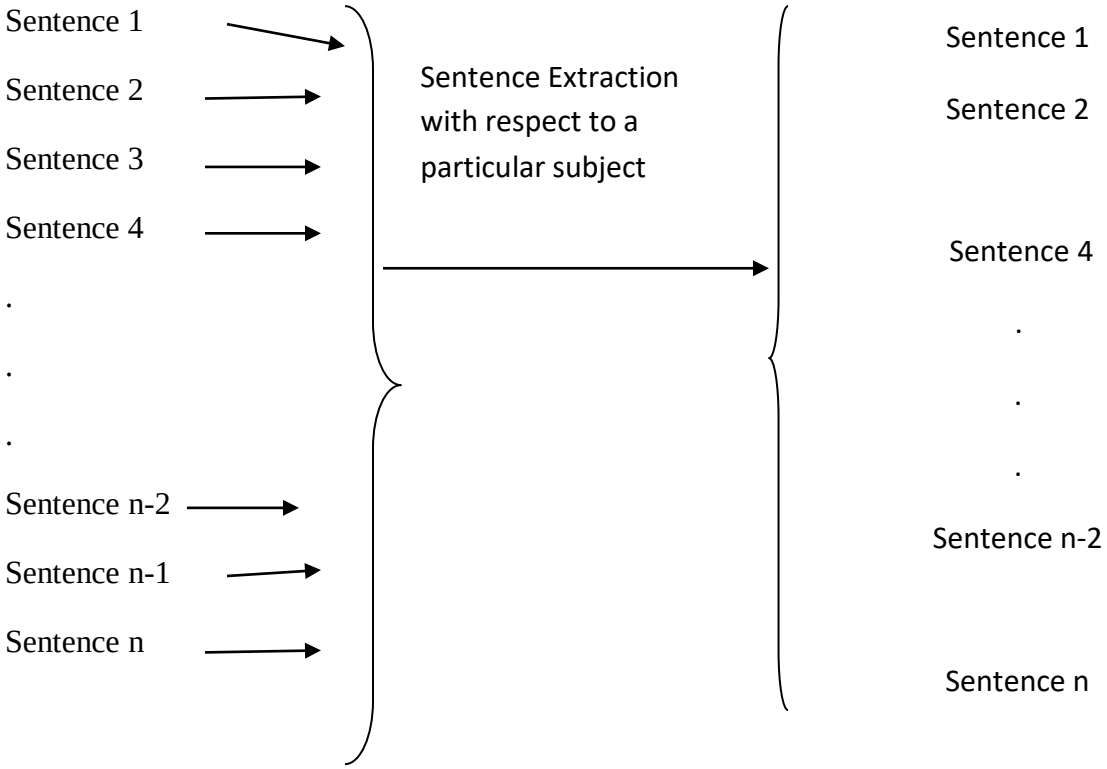

Fig.1 Subjectivity Extraction

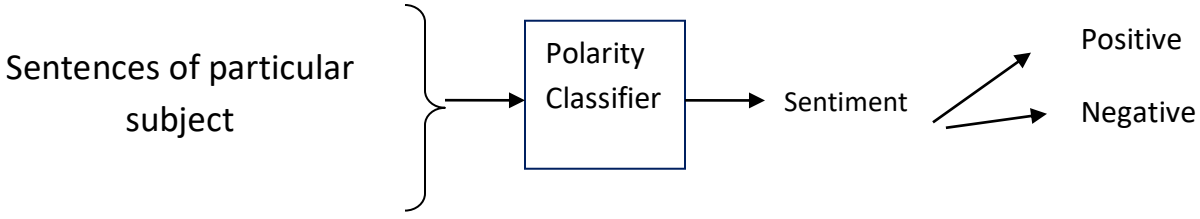

Fig.2 Sentiment Analysis

Subjective Sentences express people's beliefs.

Components needed for identifying sentiments:

- Text containing the attitudes (sentence or entire document)
- Emotional expressions (eg. Positive, Negative)

The actual task of sentiment analysis can be broken up into several subtasks, of which two major ones are discussed next.

5.4. Tokenization and Part-Of-Speech (POS) Tagging:

Tokens are individual words in a text and Tokenization is the process of breaking up into its individual words. Next, the POS Tagger software assigns a specific part of speech to each word in a text using the Penn Treebank tag set.

- CC Coordinating conjunction
- CD Cardinal Digit
- DT Determiner
- EX existential there (like: "there is" … think of it like "there exists")
- FW foreign word
- IN preposition/subordinating conjunction
- JJ adjective 'big'
- JJR adjective, comparative 'bigger'
- JJS adjective, superlative 'biggest'
- LS list marker 1)
- MD modal could, will
- NN noun, singular 'desk'
- NNS noun plural 'desks'
- NNP proper noun, singular 'Harrison'
- NNPS proper noun, plural 'Americans'
- PDT predetermine 'all the kids'
- POS possessive ending parent's
- PRP personal pronoun I, he, she
- PRP$ possessive pronoun my, his, hers
- RB adverb very, silently,
- RBR adverb, comparative better
- RBS adverb, superlative best

- RP particle give up
- TO, to go ‘to’ the store.
- UH interjection
- VB verb, base form take
- VBD verb, past tense took
- VBG verb, gerund/present participle taking
- VBN verb, past participle taken
- VBP verb, sing. present, non-3d take
- VBZ verb, 3rd person sing. present takes
- WDT wh-determiner which
- WP wh-pronoun who, what
- WP$ possessive wh-pronoun whose
- WRB wh-abverb where, when

Thus, POS-tagging is also known as grammatical tagging or word-category disambiguation. It may be described as the process of marking up a word in a text to correspond to a particular part of speech, based on its relationship with adjacent and related words in the text.

6. Proposed Work:

News on a specific day are collected from e-newspaper portal of ‘The Telegraph’. A probabilistic approach to fetch the news related to each event, in turn, is applied here with the help of a set of keywords formed earlier with samples from both datasets picked out manually. A threshold value is also set up for each event to facilitate the process of classification in choosing the correct event from the news-data.

Algorithm: Newspaperdata_Extraction

Input:

1. URL of chosen e-newspaper
2. Set of predefined keyword for each Event
3. Time period
4. Probabilistic threshold value for each Event

Output: Event-wise News Text

Method:

[1] Import the following packages:

pytesseract, requests, BeautifulSoup, datetime, Image.

[2] Fetch all pages of news in image format.

[3] For each page

[4] For each image of the article

[5] Convert the news image to text format.

[6] For each event

[7] Count= No. of key word present in this text.

[8] P_value=count/length(predefined keyword set)

[9] If P_value >= Probabilistic threshold value

[10] Then

[11] Record text news data for the event

[12] End For

[13] End For

[14] End For

Table: Key word sets for all events from News Data

Sl. No.	Tweets and News on event	Key word sets
1.	Public Health Awareness	'social distancing ', 'quarantine', 'isolation ', 'community spread ', 'lockdown', 'W.H.O guideline', 'guideline', 'awareness ','covid-19 spread', 'Sanitization', 'PPE N95 mask', 'confirmed case', 'prevent', 'stay safe from corona', 'hygiene maintain', 'trauma', 'child affected'
2.	Economy	'economy effect', 'covid-19 economy', 'lockdown economy', 'industry', 'corporate sector effect ','shutdown economy', 'economic growth', 'companies', 'market', 'jobless in corona time', 'economic fallout', 'manufacturing sector', 'rural agriculture economy', 'RBI', 'unemployment in corona time', 'farmer ','capital investment in corona time', 'atmanirbhar bharat package'
3.	Death	'covid-19 death', 'corona death', 'co-morbidities death ','death rate in corona'
4.	Education	'Impact of Coronavirus on Education', 'school college off', 'education ', 'lockdown education','covid-19 education', 'school', 'college education', 'examination', 'exam postpone', 'university education ','final year exam', 'education

		session ', 'lockdown education', 'online education', 'online class ','online child education'

Once all the news articles for a day for all the categories are extracted based upon the aforesaid algorithm, they are taken individually and POS Tagging is applied such that we can eliminate unwanted words which does not perform over the overall sentiment of the sentence. Then we can generate both positive and negative sentiment score to check how an article impacts the society. Now we need apply the same algorithm for all the news articles for a particular day. Next we calculate the average positive score and negative score for that day. Then we can visualize the score on a graph week wise so that we can check the insight of the news impact over society during the mentioned timeline over this pandemic situation.

7. Future Scope:

- As lot of data will come in future regarding this pandemic situation, the analysis result will give more insights which are overlooked at present scenario.
- We will be able to apply the same techniques to other news media where more numbers of regional news will be used for analysis.
- We can also able to generate sentiment score on social media platform where more number of people shares their personal views.
- It will be beneficiary if we can correlate the newspaper extracted sentiment with people's thinking shared over social media. It will also reveal how a speech or topic impacted individual thinking process.
- Also we can perform a statistical analysis over the topic with referencing to the predicted keywords.
- Finally with the help of statistical measurement we can train the machine to predict the sentiment of unprecedented events.

8. References:

[1] Tanu Singhal, "A Review of Coronavirus Disease-2019 (COVID-19)",2020

[2] Gopalkrishna Barkur,Vibha,Giridhar B. Kamath. "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India",2020

[3] Christiana Loana Muntean, Gabriela Andreea Morar, Darie Moldovan: Exploring the Meaning behind Twitter Hashtags through Clustering. Business Information Systems Workshop

[4] Hidenao Abe: Extracting User Behavior-related Words and Phrases using Temporal Patterns of Sequential Patterns Evaluation Indices. Vietnam J Comput Sci, 2017

[5] Zhao Jianqiang, Gui Xiaolin: Comparison Research on Text Pre-Processing Methods on Twitter Sentiment Analysis. Supported by: NSFC under Grant 1472316(in part), Shaanxi Science and Technology Plan Project under Grants 2016ZDJC-05 and 2013ZS16 - Z01/P01/K01 (in part) and Fundamental Research Funds for Ministry of Education of China under Grant XKJC2014008, February,2017.

[6] Pang B. and Lee L. Opinion Mining and Sentiment Analysis. Journal Foundation and Trends in Information Retrieval. 2008; 2(1-2): 1 – 135

[7] A Pappu Rajan and S.P.Victor, " Web Sentiment Analysis for Scoring Positive or Negative Words using TweeterData", International Journal of Computer Applications (0975 – 8887) Volume 96– No.6, June 2014

[8] Shailendra Kumar Singh and Sanchita Paul , "Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes", International Journal of Applied Engineering Research · June 2015

[9] Anusha K S , Radhika A D, "A Survey on Analysis of Twitter Opinion Mining Using Sentiment Analysis", International Research Journal of Engineering and Technology (IRJET)

[10] http://www.expertsystem.com/machine-learning-definition/

[11] J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, An imprint of Elsevier, © 2012

[12] C. C. Aggarwal, C. X.Zhai, "Mining Text Data", Springer, 2012

[13]Shiv Kumar Goel, Sanchita Patil, "Twitter Sentiment Analysis of Demonetization on Citizens of INDIA using R",2017

[14] https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

[15] Rosenfield. A, Kanungo. T, "Document structure analysis algorithms: A literature survey", Document Recognition and Retrieval X, 22-23 January 2003, Santa Clara, California, USA