Textual Analysis of Indian English Newspapers

(A Corpus Based Study)

Dr. Sandhya[1] & Ms. Neelam[2]

[1]Assistant Professor, Department of Communication Management and Technology, Guru Jambheshwer University of Science & Technology, Hisar .

[2]Researcher (UGC NET qualified).

ABSTRACT

The study was largely conducted content analysis of websites covering the widely claimed most popular online newspapers; Times of India (TOI), Hindustan Times (HT) and The Hindu (TH).Very few research studies have been conducted on textual analysis of news content of Indian newspapers. The study was borrowed from amplification effect of media of (Watson, 1998) that says "by giving intensive coverage to certain stories and issues, their importance is amplified".The objectives determined for the study were interpreted and analyzed with Antconc statistical tools to get accurate answers. It is discovered that content of the story published by the different newspapers unambiguously reflect or at least indicate their agenda and perception. This study finds this quality in line with general understanding of press being the responsible body of society and also the fourth pillar of democracy. This study examines the effective agenda setting role of the Indian online newspapers, keeping in background the emotions of their readers. In a bid to develop empirical result, while scrutinizing the news stories, it focuses on the classification of different categories of news in terms of political, social ,economic etc. polarity. This is primarily to explore and establish the connection between content and newspapers. In order to identify and substantiate the study, a total of 8,640 news stories spanning over a decade from 2006 to 2016 were evaluated. Considering the huge quantum of data, it was realized that automatic system must be resorted to put them in orderly fashion. Lastly, we elaborate on the different approaches for the machine treatment of subjective communication (opinion mining) and present our findings while bringing out their implications.

Key words –textual analysis, online newspaper, corpus.

INTRODUCTION

Media and society share a strong bond with each other. Media reflects the true images of a society as it shapes and influences the thought process of people who constitute the society. The newspaper is one of the forms of media that has remained as a popular and conventional medium to develop an informed opinion, and this informed opinion is an important ingredient for nourishing the democracy in a country. The healthy democracy and well informed public go hand in hand in the functioning of statecraft and it is important to highlight that newspaper play an important part in the informing process of population (Honderich, 2016). The newspaper serves as a two-way communication between the Government and people wherein the former disseminates and apprise the latter about the new policies and schemes, and on the other hand it provides platform to the latter to vent out their criticism or appreciation for the actions and initiatives of the former (Chakraborty, 2018). In order to feed the diverse intellect hunger in a vast country like India, there are 70,000 newspapers which make it the largest newspaper market of the world wherein 100 million copies are sold each day (Biswas, 2012).

Theoretically, the news media is expected to address as well as reach every region of the society on an equitable basis. However, this utopian definition of news media is not truly reflected in the society wherein on the ground there are islands created by the media giving preferential treatment to certain regions. Such bias adopted by the media has been a source of numerous studies across the globe. The Indian news media have come under fire from different corners for adopting the prejudiced behavior while representing the different regions in mainstream media. The languages spoken in India are widely effect it's administrative division, it also has effects on political structure and socio-cultural that ultimately affects the coverage of the media. The federal divisions and their coverage in the media are found to be linked (Arya, 2006). However, the online version of the newspaper has added new dimension to the conventional news media industry. Thus, the emerging online newspaper trend and its modus operandi to give representation to the different regions in the extremely diverse fabric of India is the starting point of this research wherein it is aimed to explore into the content and coverage of the news coverage of online English dailies in all the states of India.

Broadly, the factors that provide matrix to the news media to operate include the various dimensions of Indian society such as its social, geographic, political and economic system and lastly the evolution of online news papers will be covered to provide context for the subsequent analysis.

For the purpose of analysis of Indian states' news coverage in online English dailies, this research has selected three online newspapers, namely Times of India (TOI), Hindustan Times (HT) and The Hindu (TH). As per the survey conducted by the Indian Readership Survey in 2017, Times of India has more readers than the Hindustan Times and The Hindu put together in terms of readership of printed newspapers in India (TNN, 2018). The survey states that out of the total readership of English newspapers which is over 1.3 crore, Hindustan Times account for 68.47 lakh and The Hindu readership stands at 53 lakh, whereas Times of India has more readership than put together of both the newspapers (ibid). In terms of online readership, the research conducted in 2014 brings out that 43 percent of respondents preferred The Times of India, 33 percent opted for Hindustan Times and 20 percent selected The Hindu (Tiwari, 2015).

SIGNIFICANCE OF THE STUDY

Very few research studies have been conducted on textual analysis of news content of Indian newspapers. The study was borrowed from Daniel Learner (1958) as well as amplification effect of media. Watson (1998) discusses the amplification effect of media which says "by giving intensive coverage to certain stories and issues, their importance is amplified". It is the constitutional obligation of the central government to treat every state government equally in terms of budgetary allocations, approval of special schemes, financial packages, etc. (Rao, 2017). However, a ruling political party at the center tends to favor the states which have governments of its own party. This undue favoritism result in the disproportionate development of states. There have been different views attributed to undemocratic and biased treatment meted out to news stories placing under or over prominence of news coverage from some states (Arya, 2011). Undoubtedly, an undemocratic and biased representation of content will hamper the developmental aspect of the nation. Media are powerful instruments of development. “We seem to be painting a jaundiced view of society, which is already marred by countless cracks and socio-political divides” (Dwivedi, 2007as cited in Arya, 2011). “The media should serve as purveyors of egoless egalitarianism and modesty and attempt to create a saner national discourse” (ibid.). The need for the present study was felt in order to determine the English newspapers’ undue stress in highlighting issues in certain newspapers and probable reasons thereof. The types of news related to politics, crime, business, sports, agriculture, science and technology, natural disaster, developmental issues and defense, etc. project the focus and approach of the newspapers. On the other, hand according to Walter Lippmann (1922) the journalists select the news stories they publish not by chance or because they are motivated by personal interests, but according to the predictable importance that would have for their readers. So, present research would endeavor to find the accurate scenario of content and coverage analysis of state news in online newspapers using corpus linguistic approach.

THEORETICAL FRAMEWORK

This study is based on the theoretical framework of the agenda setting theory (Shaw & McCombs, 1977). The agenda setting theory states that “the mass news media have a large influence on the audience by their choice of what stories to be considered newsworthy and how much prominence and space are allocated to them”. The agenda setting theory focuses its clarification on how news content in the media shapes the public’s beliefs regarding what is important in society. This theory explains that once the media presented certain issues more significantly than others, those prominent issues became the spotlight of the campaign. Over time, this agenda setting analysis has enclosed conclusion that the media conjointly tell us “what to think about”, that referred to as the second level agenda setting. The second- level agenda setting research has found that media messages don’t simply highlight issues, however they present informational components regarding those issues; and those informational elements tell us what to think about the issue. The agenda-setting theory has been one of the main theories in mass communication research, since the seminal study by McCombs and Shaw (1972). A core concept of this theory is the transfer of issue salience from the media to the public (McCombs & Shaw, 1977), which theorizes that the media persuades the public to determine which issues are important. It can be said that the more intensively media cover an issue, the more importantly public perceive that issue (Lee & Hahn, 2014).

RESEARCH PROBLEM

In the beginning of the media, it was very aptly termed as “Watchdog” of the society, as its aim and obligations to check the events of a social order. But as time changed, due to rise of capitalism, cultural imperialism ushered in the practices adopted by media houses. It eventually changed the views and working styles of newsrooms as well. It became biased as catered to the needs of the shareholders in the business of news making. As a result, in Developing countries like India, where media were thought to be the guardian of poor, destitute and underrepresented it also succumbed to the motive of profit making or business model of newsrooms. The role of media has become very important to the society by now, that without realizing the trend of biased coverage of facts and ideas, we have become dependent on media for almost every aspect of life. So, we all know very well that in today's scenario the agenda setting function of media has become more prominent than ever before. Though there are others who are sure about the democracy in the use and expression of ideas due to the advents of new media. They espouse that there are so many platforms available which not only speak different opinions, but also provide a wide array of choices to the users to meet their educative, emotive or professional needs with the added bonus of expressing themselves freely and reaching millions and billions instantly. But still no one can deny that mass media in one or other way molds the policies (be it social, political or economic)and practices of a nation by governments via business and media conglomerates. Vested interests of the shareholders in media houses lead to media houses, picking the issues and contents selectively to support their profit making. This leads to representation of few state's events and issues related to them more as compared to others. That creates a certain gap in the society that silently breaks the harmony and democratic setup. But there are others who are very much sure about the freedom of news picking and making in mass media.

The media should serve as purveyors of egoless egalitarianism, modesty and it should attempt to produce a peaceful unbiased national discourse (Arya, 2011). So, present research would endeavor to inquire as to what kind of representation English language online newspaper in India gives to different states despite of their difference of socio-culture, religious, geographical, economical and political interest using corpus linguistic method.

LITERATURE REVIEW

Not abundant literature is obtainable on the account of states within the Indian news media. However, A few noteworthy studies have been done related to corpus linguistic method:-

> Weir, G. R., & Anagnostou, N. K. (2007) conducted a study "exploring newspapers: a case study in corpus analysis". Based upon our tagging, counting and data extraction, then they detailed several dimensions of results. The most frequently cited countries, towns and cities, male and female forenames were noted and discussed. Thereafter, they considered the gender-specific references within the newspaper corpus and noted the significant disparity between male and female referents. They compared the data statistics from the newspaper corpus against data drawn from the British National Corpus. In another study carried out by Bednarek & Caple (2012), titled "Depictions of strike as battle and war and comments to a South African online newspaper, with particular reference to the period following the Marikana massacre of August 2012". A corpus was constructed from the articles, editorials and comments with in online news from the online newspaper of The Guardian and The National Weekly Mail in the period of four years from 2008 to 2012. The results of the analysis revealed that the articles, editorials and comments of readers represent the protestors as anti social elements. These representations show that the protestor and participants of such wars and

battles are considered to be opponents of the society and they don't get included as a part of South African society. Whereas Paul Baker, Costas Gabrielatos and Tony McEnry (2013) in their study "Muslim representation in the British newspapers" which was conducted after 9/11 terrorists attack on world trade centre in United State America, found that a gradual move towards news stories that are personalizing (referring to Muslims), rather than the more abstract concept of Islam the religion. Additionally, there is a gradually increasing focus on stories about Muslims in the UK context, as opposed to Muslims in other countries. It is interesting to note how the concept of extremism appears to be fairly prevalent across the corpus, although this is referred to by changing terms: hardliner, fanatic, militant, radical and extremist. Other concepts are restricted to particular periods, such as veiling and references to tolerance or hatred, which tend to be more common towards the end of the corpus data, although they do not necessarily indicate that the British press has become more tolerant of Islam. In another study Monika Bednarek, James Curran, Tim Dwyer & Fiona Martin (2014) described in their study "All the news that's fit to share" for the analysis of most shared news items,100 articles from English newspapers were selected and most shared news stories were selected on the basis of their share and likes on social media. The researcher aimed to find the common news value and its language of most shared news stories by applying corpus linguistics method. They opine that this study can give an outline to the conceptual framework and its application to the qualitative and quantitative analysis of online news stories. Brindle (2015) carried out a study "A corpus analysis of discursive constructions of the Sunflower Student Movement in the English language Taiwanese press" Brindle carried out a study of two newspapers with an objective to develop an ideological understanding of Taiwanese society through analyzing the news media covering the anti-establishment protests in Taiwan. In order to understand the emerging patterns of student protests, the keywords and lexical frequency in the news corpora were examined to analyze the corpus based content analysis in two newspapers. The paper brought out findings which broadly pointed out that the Taipei Times projected the student protests as part of their ongoing struggle in having democratic set up and also to Taiwan's independence.

METHODOLOGY

The study was largely conducted content analysis of websites covering the widely claimed most popular online newspapers. A mixed method was opted for the study, which combines the qualitative and quantitative approach (Tashakkori & Teddlie, 1998) to study the content of the news items appearing in online national newspapers. This method has been advocated by various most prevailing and influential methodologists in the social sciences (Dörnyei, 2007 cited in Bednarek, 2009). Quantitative method emphasizes on "objective measurements and the statistical, mathematical or numerical analysis of data and focuses on gathering numerical data and generalizing it across groups of people or to explain a particular phenomenon" (Babbie 2010) whereas, Brennen (2017) discussed that *"qualitative research is interdisciplinary, interpretive, political and theoretical in nature*. Using language to understand concepts based on people's experience, it attempts to create a sense of the larger realm of human relationships".

OBJECTIVES

Following broad and specific objectives were deciphered by qualitative and quantitative method in present study:-

Broad Objective

The Broad objective of the study was "to determine what kind of representation English newspapers in India give to different news stories."

Specific Objectives

1. To perform a detailed textual analysis on the news corpus in terms of - lexical density (word token and word types), word list and word frequency.
2. To generate the prominent words of the newspapers, news corpus and to perform a detailed qualitative study on the news corpus by finding the collocates and KWIC (Key Word in Context) in the concordances.
3. To generate the word list of news corpus filtering for various forms of words (positive and negative).

STATISTICAL TOOLS

The objectives given above were interpreted and analyzed with **Antconc** (using a corpus of text) statistical tools to get more accurate answers.

Corpus (Corpora: plural) is an electronic based authentic words database that can be available through the internet or as software installed on the desktop (Hasselgard, 1997). Speech in a corpus can be either a collection of written or spoken texts; for example, written texts, collected from newspapers, movies, political speech, business letters, popular fiction, books, or magazines, published or unpublished school essays etc (Witton, 1993). Collections of spoken texts can be any recorded formal or informal conversations, radio shows, talk show, weather broadcast or even business meeting setc. Usually, for the contextual analysis of large number of language, a searching tool the concordance is used by corpus users (ibid); the concordance in corpus provides users better quality of examples and more exposure to an unknown word (Cobb, 2003). By using the concordance tool of corpus to search forword contexts, researchers are involved in a more speedy and efficient text analysis experience.

The Corpus Linguistics method was initially designed for researchers in English Language, linguistics and Semantics Study, but today is a well recognized and vastly used approach for content analysis in all the fields of Social Sciences. By applying the Corpus Linguistics method, the researcher can get both qualitative as well as quantitative analysis of the data. Krippoendorff (1980) states, "though the text analysis approach in mass media research has now become relatively accepted as an alternative or addition to conventional content analysis". Corpus- linguistic studies are about the computer-aided exploration of lexical, pragmatic or syntactical, which generates a semantic similarity and relatedness that occur in a specific text collection (Baker, 2006, Simon et al, 2002, Stubbs, 2001). Following Stubbs (2001) features of semantic relatedness can be formed "as clusters of Lexis (node and collocates), grammar (colligation), semantics (preferences for words from particular lexical fields) and pragmatics (connotations or discourse prosodies)". Dörnyei (2007) gives an overview of researchers in social science using 'mixed methods' approaches involving both quantitative and qualitative research. This has been advocated by some of the most eminent researchers in the social sciences.

WEB TOOLS

www.site:(newspaper site name)intitle:(state name) - Google power searching command has been used to collect the relevant data from the web and saved in a notepad file for corpus analysis.

http://www.raosoft.com/samplesize.html - An application for authentic sample selection according to the selected level of confidence and a certain error margin.

https://resoomer.com/en/ - This application was adopted to summarize the huge data collected in note pad file for corpus analysis.

PROCEDURES

The research was conducted in following steps:

Sample selection - The online newspapers of three most read and circulated English Media, The Times of India TOI), Hindustan Times (HT) and The Hindu, were chosen for the study. TOI has a readership of 13.4 million and a circulation of 31, 98, 449, HT has a readership of 6.8 million and a circulation of 11, 68, 613 and The Hindu has 5.3 million readership and 15, 48,660 circulation (ABC, 2017 & IRS, 2017). These are the three most read and nationally circulated English newspapers. English newspapers were chosen because English is considered to be a major language in India after Hindi. It is typically used among a nation's educated class and expatriate community; and second because newspapers that publish in this shared language often is among a nation's most influential (Massey and Levy, 1999et al.)

Unite of analysis - News Item, in which the name of the respective state appears in the headlines, was counted as a unit of analysis. This would be found by applying Google power searching commands, for example; (www.site:thehindu.comintitle:haryana)

Data collection- The data for this study is retrieved from the online database of mentioned above newspaper's websites;https://timesofindia.indiatimes.com/,www.hindustantimes.com and www.thehindu.com with the use of Google power searching tool.

MAINTAINING THE CODE BOOK AND CORPUS CREATION

a) First, along with the data collection the code sheet was maintained with all the related information mentioned in the code book. Information of every unit of analysis was taken and noted in the code sheet for further analysis.

b) Secondly, a corpus of words was created from the news at www.timesofindia.com, www.hindustantimes.com and www.thehindu.com. The steps in creating the corpus are as follows:

Every states news from the website was copied and pasted in to a notepad file separately. Dates, months and the name of writer in the news were deleted from the texts. Only headlines and the body of the news were copied into a master file. For instance, there is a reference on the top of every page saying, "published on (date/month/year)." These phrases would affect the frequency of word list. The texts were collected into state wise and newspaper wise file. Then the corpus was fully developed.

Data analysis

Antconc-The entire corpora was analyzed as per the defined objectives of the study. Various corpus tools have been used like collocates, keywords, Key Word in Context (KWIC), type and token ratio, lexical density, concordance, prominent words generation and positive and negative words.

DATA ANALYSIS AND INTERPRETATION

Objective -1

The objective one was formed to perform a detailed text analysis on the news corpus in terms of - lexical density (word token and word types), word listing and word frequency.

Word Type to word Token ratio was obtained by using the following formula:

$$\text{Type to Token Ratio (TTR)} = \frac{\text{Total word tokens X 100}}{\text{Total word types}}$$

$$\text{Token to Type Ratio (TTR)} = \frac{\text{Total word type}}{\text{Total word Token}}$$

Table (1)- TTR list of TOI, HT and TH online newspapers.

	Word Token	word Types	Types to Token ratio	Token to types Ratio
TOI	369429	24715	6.69%	14.95%
HT	159871	15910	9.95%	10.05%
TH	534980	30629	5.73%	17.47%
TOI+HT+TH	1064280	43713	4.11%	24.35%

With the advent of electronic corpora and corpus processing tool, it has become much easier to transfer a set of texts and a set of complete list. A word-list is essentially a list of word-types. A word list program goes through a text or a set of text and reduces all repeated tokens to types, that is, each instance (token) of the word THE is counted, but the completed list display THE only once as "type", usually together with its frequency(the number of tokens found). As the table shows that TOI's Types to token ratio is (6.69%) which has very poor lexical diversity. Although this ratio is more than The Hindu's types to token ratio that is (5.73%) while the token to the types ratio of TOI is (14.95%) however The Hindu's token to type ratio is richer than TOI with (17.47%).Whereas HT has the maximum type to token ratio with (9.95%) than other two online newspapers. On the other hand TOI is lagging behind in token to type ratio with (10.05%) lexical diversity.

OBJECTIVE -2

Objective two was aimed to generate the prominent words of the news corpus and to perform a detailed qualitative study on the news corpus by finding the collocates and KWIC (Key Word in Context) in the concordances.

Table -2 - The 50 most frequently used words in the corpus of The Hindu Online Newspaper.

Rank	Freq	Word	Rank	Freq	Word
1	7586	State	26	1068	Centre
2	4561	government	27	1041	Water
3	3190	Minister	28	1031	woman
4	2991	Year	29	1022	Crore
5	2524	District	30	1015	Power
6	2248	People	31	993	Make
7	2126	Chief	32	987	Country
8	2026	India	33	951	Case
9	1802	Party	34	945	Assembly
10	1787	Area	35	943	Lakh
11	1546	Police	36	923	Court
12	1408	Pradesh	37	917	Number
13	1385	Bjp	38	890	Leader
14	1381	Official	39	875	Group
15	1258	Land	40	873	Farmer
16	1233	Congress	41	872	Department
17	1196	Delhi	42	848	Political
18	1188	time-share	43	839	Indian
19	1165	Village	44	796	Family
20	1160	Project	45	781	Child

21	1124	Issue	46	779	Demand
22	1113	National	47	754	Provide
23	1104	Report	48	725	Road
24	1103	Election	49	718	Forest
25	1078	development	50	700s	Increase

Words Frequency– Most lexical software packages generate a list of keywords in combination with their particular frequency counts. A keyword can be used as a key trait and occurrence counts can be measured as the strength of that attribute hidden in the word.

Frequency of words is "one of the most basic way" to explain the attitude or discourse of a corpus (Baker, 2010). "Frequency can be an indicator of markedness" (ibid). Word frequency can point out a possible partiality in text and can be investigated of the text, therefore its importance may not be taken granted. The table shows the relative frequency of words appearing at least more than 500 times in the corpus.

In the Table -2 we can see 50 most frequently used words in the news of The Hindu Online newspaper. In which the researcher explored that words related to politics are leading in the most frequent word list. However the word list of The Hindu represents that the development and social concern dominating words like people, village, issue, development, water, woman, make, farmer, family, child, provide, road and forest etc. has been published more after politics related words. The Hindu has used mostly progressive, positive, and development related words most frequently.

The top frequently occurred word list of The Times of India is almost same as The Hindu's word list. As we can see in the table -3 of most frequently used words in The Times of India that there is almost same set of words used in The Hindu . Despite the similarity in most words used in most frequently used word list of The Hindu and The Times of India, there were few words which are different from other corpus list e.i. Gujarat, Bihar that depicts that these two states have been getting extra attention from other Indian states. There were the occurrence of word "city" more than "village" that indicates that the times of India has been publishing more news from the city than related to village issues. There were also words like education, students, school, health, death, and population in the top frequently used words list of The Times of India. While these words were missing in the word list of The Hindu that represents that The Times of India is publishing more news regarding education and health than The Hindu newspaper.

Table -3 - The 50 most frequently used words in the corpus of The Times of India Online Newspaper.

Rank	**Freq**	**Word**	**Rank**	**Freq**	**Word**
1	6158	State	26	753	National
2	3461	government	27	740	Land
3	2434	Year	28	734	Child

4	1938	Minister	29	724	Court
5	1652	District	30	709	Congress
6	1544	People	31	696	Issue
7	1483	Police	32	691	Student
8	1305	Chief	33	689	Development
9	1300	Area	34	667	Make
10	1249	India	35	663	Village
11	1229	Official	36	636	Water
12	1093	Report	37	634	Bihar
13	1061	Case	38	634	Road
14	976	Delhi	39	610	Increase
15	970	Bjp	40	527	Health
16	934	Crore	41	463	Death
17	908	Project	42	492	Children
18	903	department	43	456	Education
19	879	Gujarat	44	449	Village
20	865	City	45	431	Road
21	843	Country	46	405	Case
22	804	Party	47	401	Population
23	768	Lakh	48	379	Forest
24	761	Power	49	338	School
25	757	Woman	50	318	Family

Table -4 - The 50 most frequently used words in the corpus of Hindustan Times Online Newspaper.

Rank	**Freq**	**Word**	**Rank**	**Freq**	**Word**
1	2187	State	26	307	woman
2	1235	government	27	305	Case
3	1214	Year	28	290	Party
4	714	Minister	29	289	Assam

5	695	India	30	284	Bjp
6	692	People	31	282	Make
7	667	Police	32	280	congress
8	657	District	33	272	Crore
9	555	Official	34	267	Delhi
10	498	Nagaland	35	261	Lakh
11	464	Chief	36	261	officer
12	463	Area	37	254	arunachal
13	404	Student	38	244	Girl
14	379	department	39	241	Power
15	375	Group	40	235	Plan
16	362	Village	41	234	Forest
17	359	jharkhand	42	234	School
18	351	National	43	233	family
19	347	Naga	44	231	Leader
20	343	Issue	45	223	China
21	340	Child	46	212	Water
22	336	uttarakhand	47	209	Land
23	334	Report	48	208	Border
24	307	Country	49	206	Health
25	307	Project	50	206	School

As the table - 4 shows that Hindustan Times word list could not be any different from The Hindu and TOI’s word list. Though, these both corpus files share almost the same words in top most frequent word list. However, there were few words which were different from other two newspapers. In the list of Hindustan Times the words that are different from other two newspapers were Nagaland, naga, Jharkhand, Uttrakhand, Assam and Arunachal, which clearly shows that news related to these states were published more in Hindustan Times than other two newspapers. Other unusual words that have been noticed in the word list of Hindustan Times were “China” and “border” that represents the issues related to Indo-China and disputes related to Indian border with neighbor countries have been published more than other two newspapers. The word “girl” has mentioned (243) times in the Hindustan Times newspaper while it did not make in the list of The Hindu and The Times of India. Whereas, like the Times of India, Hindustan Times has also published issues related to student and school, but the word “education” was missing.

Table -5 The 50 most frequently used words in the corpus of All Three Online Newspapers including The Hindu, The Times Of India and Hindustan Times .

Rank	Freq	Word	Rank	Freq	Word
1	15931	State	26	1968	Development
2	9257	government	27	1942	Make
3	6639	Year	s28	1889	Water
4	4833	District	29	1855	Child
5	3970	India	30	1777	Court
6	3696	Police	31	1716	Group
7	3550	Area	32	1716	Student
8	3165	Official	33	1663	Gujarat
9	2896	Party	34	1660	Election
10	2639	Bjp	35	1553	Leader
11	2531	Report	36	1552	Road
12	2439	Delhi	37	1485	Family
13	2375	Project	38	1429	Forest
14	2317	Case	39	1346	Farmer
15	2228	Crore	40	1340	Health
16	2222	Congress	41	1281	Home
17	2217	National	42	1312	Lead
18	2207	Land	43	1273	Political
19	2190	Village	44	1250	Poll
20	2163	Issue	45	1248	Modi
21	2154	Department	46	1225	Union
22	2137	Country	47	1219	Secretary
23	2095	Woman	48	1193	Border
24	2017	Power	49	1126	Education
25	1972	Lakh	50	1102	North

Though the higher frequency of the first few words is quite evident owing to the domination of the political news in newspapers, but the unusual high occurrence of the other words like “people”, “village”, “project”, “development” etc shows that the development and society related issues have been published widely. We can see 50 most frequent words used in the news of all three online newspapers. In which the researcher revealed that words related to politics like government, state, minister, district, bjp, congress etc were leading in the most frequent word list. However village, project, issue, report and development kind of words are representing the development dominating news has been published more than any other news category. If we take a close look at the list we will find that these newspapers have mostly progressive, positive and development related words most frequently like power, make, leader, increase women, child, water, provide, project etc. The words like women, child and water are related to some basic issues any country is facing and being national online newspapers, they have been doing their duty.

OBJECTIVE -3

Objective three was carried out to perform a detailed qualitative study on the news corpus by finding the KWIC (Key Word in Context) in concordances and collocates words.

KWIC (Key Word in Context) in concordances

Study revealed that words related to politics like government, state, minister, district bjp; congress etc were leading in the most frequent word list. However village, project, issue, report and development kind of words were representing the development dominating news has been published more than any other news category.The Hindu has mostly progressive, positive, and development related words most frequently like power, make, leader, increase, women, child, water, provide and project etc.

Word Collocates

Most frequently used word “*land*” has been friend with words like; inteqal, gair (gairmumkin land), down, cabinet, becharam (land reform minister), will, better. While the frequency of the word land is mostly used with “will” with more than 50 which indicate hope, expectations and future plans.“*Village*” has been collocated with the word “near” with more than 50 repetitions with the word “village” followed by remote, nearby, eco, panchayat, block and smart. The collocates for word “*project*” were chief project scientist, electrification, water, ambitious, pilot and reservoir etc. That indicates that the development issues have been taking place in news.“*Power*” was one of the most frequently used word in the corpus so the high intensity collocates of the word “ power” were; roof top solar, SunEdisonAdani enterprise, chief minister , palatana (tripura), vidhyut, water, thermal and hydro etc. The collocates with the word “*Women*” with most stats were old, unborn, empowerments, streedhan, uneducated, follower and smuggler etc.“*Water*” word’s collocates were three, major, air, will, potable, safe, depletion (reduction), sewerage, scarcity (shortage), surplus, sanitation, shortage, acute, irrigation and Cauvery etc. The high intensity collocates of “*development*” at two words left and two words right (against a flexible five word standard practice accepted internationally)showed up the words chief, major, will, quick, better, millennium, infrastructural, summary, front and composite etc.The collocate words with the word "*student*" were old, placement, class, ratio, organizations, teacher and college etc.With high state collocate words of the "*leader*" in the corpus of online newspaper were opposition, spiritual, senior, top, mukti and legislature etc.Collocate words with the word "*Family*" were welfare, health, minister, royal, union, poor, diseased, compensation, hospital, farmer and income etc.The word "*Health*" was collocating with words mostly minister, state, care, department, education, public, family, welfare, national, mission, medical, primary, union, people, scheme, private and infrastructure etc."*Forest*" was mostly collocated with the department, cover, state, area, land, dense, reserve, environment, increase and conservator etc.Mostly associated words with the word *"Farmer"* was suicide, village, land, family, small, seed extension, sugarcane, paddy, debt, cotton and progressive etc. "*Road*" collocated with transport, connectivity, construction, traffic, rail, people, project and infrastructure etc."*Border*" was associated mostly with Indo, Myanmar, security, china, force, trade, international, dispute, across and Manipur etc. "Election" was collocated with mostly commission, assembly, campaign, BJP, congress and constituency etc words with highest stat percentage in the corpus of all three online newspapers.

OBJECTIVE - 4

Objective four is related to the opinionated words such as negative and positive used in the text of all three online newspapers. To achieve the objective the word list was generated in the news corpus filtering for various forms of words (positive and negative).

Frequency of Negative Words versus Positive Words

The positive and negative words have been analyzed on the basis of Minqing Hu and Bing Liu's (2004) opinion word. Which are characterized according to the social media word usage. The corpus of The Hindu had 2300 negative words out of total 565609 words (0.40%) whereas The Times of India had 1880 (0.47%) negative words out of total 394144 words. Hindustan Times had 1450 (0.83%) negative words out of total 175781 words in its corpus. Table shows the top 20 negative words. Most repeated negative words in all three newspaper's corpus are loss, kill, death, poor, problem, fail, rape etc. The researcher must mention here that words per se may not fully qualify for being positive or negative unless the related context is read into. But prima facie, the negative words seemed unequivocally coherent in conveying the impression here in the corpus (Hu and Liu, 2005).

On expected lines, negative words (kill, loss, damage, died, problem, rape, protest, crime, attack etc) consistently dominatated the news stories. Every word has its own connotation here, for example, negative word "kill" has been occurring (283) times in TOI, (250) times in HT and (327) times in The Hindu corpus, which has been used to describe the end of life legally or illegally, sometimes in the case of murder by criminals or non- criminal persons, knowing or unknowingly ending someone's life or assault someone for personal reasons or sometime without reason. The word "kill" has been used in the concordance of the level of crime, fear in the society, regulation by law &order and timeliness of police and media. Second, most frequently used negative word is "loss" that has been used (269) times mostly, in the sense of losing a competition, an event or election and in lost and found case.

To generate the top 20 Positive and negative words in all three online newspaper's corpus files.

Positive words-

Out of Total word 394144 in the corpus of The Times of India, there are 1110 positive words (0.28%). Hindustan Times' total word in the corpus was 175781 in which there are 771positive words (0.43%). On the other hand The Hindu corpus has 1275 positive words (0.22%) in the complete corpus of 565609 words. We have taken the top 20 most frequently used words to find the relevance of the words in the corpus of all three newspapers separately. Then we have studied the concordance of the selected 20 positive words of each newspaper's corpus.

Top 20 positive words of The Times of India newspaper were new (878), working (663), lead (465), good (349), win (337), support (302), expected (196), improve (171), award (139), promise (137), relief (137), award (135), benefit (131), strong (116), positive (107), protection (107), great (102), safe (79), peace (74), skill (70) and success (64).

Top 20 positive words of Hindustan Times newspaper were new (434), lead (200), peace (164), good (143), high (139), support (108), relief (93), promise (81), win (80), strong (64), great (55), improve (51), rich (50), love (49), award (47), grand (47), benefit (42), protection (41), popular (39), rich (38) and smart (36).

Top 20 positive words of The Hindu newspaper were new (1256), lead (788), support (509), win (458), good (733), expect (271), benefit (247), strong (237), peace (222), improve (210), promise (194), relief (182), award (168), great (166), protection (164), grow (155), significant (145), fast (142), gain (142), interest (135) and victory (122).

Negative words-

Sometime, to depict the situation when one is present physically but not mentally. It has been used in the context of loss of transaction, e.g. business, land or money. "Loss" has been used to symbolically present a situation of winning or losing. After "loss" the next most frequently used negative word is "died" in the corpus of TOI, HT and TH .That describes the state of death of a living thing. It has been used mostly in the context of death, murder or killing someone or group of people. It also represents the death due to criminal act and life loss because of natural or man-made calamities. The word "died" has been extensively used in the context of loss of life in violence between groups, classes, across the border or within the society and war with external or eternal enemies and militancy. It has been mentioned in the context of most hospitals and medical facilities and willful or knowingly killing someone.

Top 20 negative words of The Times of India newspaper were died (676), loss (453), kill (375), problem (247), rape (209), damage (173), poor (251), protest (162), crime (157), attack (132), complaint (181), fall

(162), fail (160) waste (134), illegal (115), disaster (100), cancer (98), corruption (89), suffer (86) and risk (75).

Top 20 negative words of Hindustan Times newspaper were kill (250), loss (185), death (195), rape (129), problem (119), accused (116), crime (109), disaster (94), protest (85), conflict (79), attack (78), fall (73), crime (72), illegal (71), poor (70), fail(68), damage (65), war (65), complaint (63) and threat (45) .

Top 20 negative words of The Hindu newspaper were lose (524), die (494), problem (430), kill (327), poor (273), protest (249), fail (195), rape (195), fall (194), illegal (100), disaster (174), attack (167), damage (164), complaint (156), difficult (155), war (152), crime (142), conflict (134), corruption (134) and crisis (132).

CONCLUSIONS

The federal structures of the India and administrative divisions carved therein have a profound bearing on the emergence of diverse socio-cultural and political fabric of the country. This factor is amply reflected in the uneven coverage of media in different states and the emerging digital technology in the media has added a new dimension to be explored for the scholars in the field. Thus, the extremely diverse fabric of India and the emerging online newspaper trend was the starting point of this research wherein it was intriguing to explore into the content and coverage of the news coverage of online English dailies in all the states of India. In the process of research, the variations were traced in the reporting pattern of national online newspapers.

Lexical density is a measure of a newspaper that shows how meaningful, informative and descriptive the text is. The type and token ratio suggested that leading online newspaper's lexical diversity is very poor. Although HT newspaper has high lexical diversity, while The Hindu and TOI has quite a low lexical density. That reflects less unique words and more repeated words meaning thereby a repeated vocabulary has been used to disseminate the information in the news of online newspapers. Most lexical software packages generate a list of keywords along with their respective frequency counts. A keyword can be considered a key attribute, and frequency counts can be considered the strength of this attribute embedded in the text. The study explored that top most frequent words used in the news of all three online newspapers were related to politics like government, state, minister, district, BJP, congress etc. are leading in the most frequent word list. However village, project, issue, report and development kind of words are representing that the development dominating news has been published more than any other news category. They used mostly progressive, positive and development related words most frequently like power, make, leader, increase, women, child, water, provide and project etc. The words like women, child, water are related to some basic issues any country is facing and being on the list of top 50 most frequent word list of leading online newspapers it can be concluded that the national media has been doing its duty by publishing issues widely about development and society. It was also noted that almost same set of words with the hierarchy of most frequent words has been chosen for The Hindu, HT and TOI in their text. It was revealed that the corpus of The Hindu, HT and TOI has more negative words then positive words. It can be viewed in both the ways, either news stories containing negative annotations sells faster and attract more readers. Or, the other viewpoint may be the declining situation of our society which in turn gets reflected in the newspapers.

References and Bibliography

- A Shifting Global Economic Landscape. (2017). *World Economic Outlook*. Retrieved from https://www.imf.org/external/pubs/ft/weo/2017/update/01/
- Agarwal, S. (2018). Internet users in India expected to reach 500 million by June: IAMAI. Retrieved from https://m.economictimes.com/tech/internet/internet-users-in-india-expected-to-reach-500-million-by-june-iamai/amp_articleshow/63000198.cms
- Akoijam, I. (2013). Coverage of the North East declines. Retrieved on 25.5.2018from http://www.thehoot.org/web/Coverage-of-the-North-Eastdeclines/6918-1-1-54-true.html
- Arya, U. (2011). Uneven Cultivation of States and Union Territories' News Coverage on Indian Print Media Landscape A Content Analysis of English Dailies. Media Asia, 38(2).
- Babbie, Earl R. *The Practice of Social Research*. 12th ed. Belmont, CA: Wadsworth Cengage, 2010.
- Baccianella, S., Esuli, A. &Sebastiani, F. (2010, May). Sentiwordnet 3.0: An Enhanced Lexical Resource For Sentiment Analysis And Opinion Mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).
- Bagchi, A. K. (1991). Industrialization And Growth: A Comparative Study. *Structural Change and Economic Dynamics*, *2*(1), 247-249. doi:10.1016/0954-349x(91)90016-l
- Baker, P. 2006. Using Corpora in Discourse Analysis.London: Continuum.
- Balahur, A., Steinberger, R., Van der Goot, E., Pouliquen, B., Kabadjov, M. (2009). Opinion Mining on Newspaper Quotations. Proceedings of the Workshop 'Intelligent Analysis and Processing of Web News Content' (IAPWNC), held at the 2009 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Milano, Italy, 2009.
- Bansal, K. (2003). International News Coverage In Four Indian Newspapers: A Content Analysis study. Media Asia, 30(1), 31-40.
- Basnett, P. (2010). Coverage of Northeast India in the Indian Mainstream Media: A Study of the Perception of Northeast Indians Living in Bangalore. Retrieved on 20.5.15 from http://repository.christuniversity.in/1816/5/Initial_Pages.pdf
- Bednarek, M. (2009). Corpora and discourse: A Three-Pronged Approach To Analyzing Linguistic Data. In Selected Proceedings of the 2008 HCSNet Workshop On Designing The Australian National Corpus (pp. 19-24). Somerville, MA: Cascadilla Proceedings Project.
- Bednarek, M., Curran, J., Dwyer, T., Martin, F.&Nothman, J. (2015). “All the News That’s Fit To Share”: Investigating The Language of “Most Shared” News Stories. In *Abstract Book,Corpus Linguistics 2015* (p. 44).
- Benton, M. & Frazier, P. J. (1976). The Agenda Setting Function Of The Mass Media At Three Levels Of" Information Holding". Communication Research, 3(3), 261-274.

- Bhargava, Y. (2014). India Has Second Fastest Growing Services Sector. *The Hindu*. Retrieved from http://www.thehindu.com/business/budget/india-has-second-fastest-growing-services-sector/article6193500.ece
- Boczkowski, P. J. (2005). Digitizing The News: Innovation In Online Newspapers. Mit Press.
- Bolboacă, S. D., Jäntschi, L.,Sestraş, A. F., Sestraş, R. E.&Pamfil, D. C. (2011). Pearson-Fisher chi-square statistic revisited. Information, 2(3), 528-545.
- Boutron Y.A. et al. (2012). "Misrepresentation Of Randomized Controlled Trials In Press Releases And News Coverage: A Cohort Study". PLoS Med 9(9): e1001308.
- Bowker, R.& Hunt, S. (2015). Depictions Of Strikes As "Battle" And "War" In Articles In, And Comments To, A South African Online Newspaper, With Particular Reference To The Period Following The Marikana Massacre Of Aug. 2012. In *Abstract Book,Corpus Linguistics 2015* (p. 60). Lancaster: UCREL.
- Chakraborty, S. (2018). Only 5 states account for 70% of exports, Economic Survey shows. Retrieved from https://www.business-standard.com/budget/article/only-5-states-account-for-70-of-exports-economic-survey-shows-118012900344_1.html
- Chakraborty, S.&Chakma, N. (2016). Economy and Social Development of Rural Sikkim. *Space and Culture, India*, *4*(2), 61. doi:10.20896/saci.v4i2.198
- Digirolamo, G. J., and Hintzman, D. L. (1997). First Impressions Are Lasting Impressions: A Primacy Effect In Memory For Repetitions. Psychonomic Bulletin & Review 4(1):121–124.
- Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, 85(1), 87-94.
- Gupta, K. (2015). Transgender identities in the UK mainstream media in a post-Leveson context. In *Abstract Book,Corpus Linguistics 2015* (p. 143). Lancaster: UCREL.
- Haneefa, M.&Nellikka, S. (2010). Content Analysis of Online English Newspapers in India. *DESIDOC Journal of Library & Information Technology, Vol. 30, No. 4, July 2010, pp. 17-24.*
- Hashima, N. H.& Meloche, J. A. (2007). Australian online newspaper: an exploratory study on internet savvy users using Q-Methodology.
- Heeter, C (1989), 'Implications of New Interactive Technologies for Conceptualizing Communication', in JL Salvaggio& J Bryant (eds), Media Use in the Information Age: Emerging Patterns of Adoption and Computer Use, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 217-35.
- Ihlström Eriksson, C. (2005). The e-newspaper innovation-converging print and online. In International Workshop on Innovation and Media: Managing changes in Technology, Products and Processes, Stockholm, November 11-12, 2004. DigiNews.
- K, A. A. (2016). The Economist. Why India's newspaper business is booming. Retrieved from https://www.economist.com/blogs/economist-explains/2016/02/economist-explains-13
- Kothari, C. R. (2004). Research methodology: Methods and techniques. New Age International.
- Lee, J. H.& Hahn, K. T. (2014). Factors Influencing the Agenda-Setting Effects of Newspapers on Their Subscribers. 미디어경제와문화, 12(1), 192-233.

- Liu B., Zhang L. (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA
- Mahtaney, P. (2007). India: Her Tryst with Globalization. *India, China and Globalization*, 18-29. doi:10.1057/9780230591547_4
- Menezes, B. (2015). Indian IT services exports seen growing 12-14% in year ahead. *Live Mint*. Retrieved from http://www.livemint.com/Industry /bCLOgyaLGiIi6TuhmN0S7J/Indian-IT-services-exports-seen-growing-1214-in-year-ahead.html
- Minqing Hu and Bing Liu.(n.d.) "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge
- Nehru, J. (2008). Discovery of India. Penguin UK.
- Peter Williams, David Nicholas, (1999) "The migration of news to the web", Aslib Proceedings, Vol. 51 Issue: 4, pp.122-134, https://doi.org/10.1108/ EUM0000000006971
- Quandt, T. (2008). News on the World Wide Web? A comparative content analysis of online news in Europe and the United States. Journalism Studies, 9(5), 717-738.
- Ray, R. K. (2017). India's economy to become 3rd largest, surpass Japan, Germany by 2030. Retrieved from https://www.hindustantimes.com/business-news/india-s-economy-will-become-third-largest-in-the-world-surpass-japan-germany-by-2030-us-agency/story-wBY2QOQ8YsYcrIK12A4HuK.html
- RNI. (n.d.). *Registrar of newspaper for India*. Retrieved from http://rni.nic.in/#
- Roy, J. (2015, February 13). Role of media in our society – Global Ethics Network. Retrieved from https://www.globalethicsnetwork.org/m/blogpost?id=6428686%3A BlogPost%3A53995
- Sengupta, A. (November 2, 2007). Northeast India: Through the Prism of the National Media. Retrieved on 22.5.18 from http://www.ipcs.org/article/terrorism-innortheast/northeast-india-through-the-prism-of-the-national-media-2409.html
- Census. (2011). *population*. Retrieved from https://web.archive.org/web/20131116 020014/http://mospi.nic.in:80/mospi_new/upload/SYB2013/CH-2-POPULATION/ Chapter%20No-2-Population.pdf
- Tcn News. (2011). Print Media grows by 6.25%; Urdu at No. 3. *Two Circles News*. Retrieved from http://twocircles.net/2011dec30/print_media_grows_625_urdu_no_3.html
- Tesch, R. (1990). Qualitative research: Analysis types and software tools. Bristol, PA: Falmer
- Thangkhal, B. K. (n.d.). NorthEast needs space in Mainstream Media. Retrieved from http://e-
- The Constitution (Sixty-Ninth Amendment) Act. (1991). *Ministry of Law and Justice,Government of India*. Retrieved from http://indiacode.nic.in/coiweb/amend/amend69.htm
- The Race-Newspapers have a bright future as print-digital hybrids after all -- but they'd better hurry. (2007). Columbia Journalism Review. Retrieved from http://archives.cjr.org/cover_story/the_race.php?page=1
- Thiel, S. (1998). The Online Newspaper: A Postmodern Medium. *The Journal of Electronic Publishing*, *4*(1). doi:10.3998/3336451.0004.110

- Tiwari, P. (2015). The habits of online newspaper readers in India. Journal of Socialomics 2015, 4:2. doi:10.4172/2167-0358.1000124
- UNPD. (n.d.). India Becomes a Billionaire. *Population Division Department of Economic and Social Affairs*. Retrieved from https://www.un.org/esa/population /pubsarchive/india/ind1bil.htm
- Van Dijk, T. 1988. News as Discourse. Hillsdale, NJ: Lawrence Erlbaum.
- Van Dijk, T. 1991. Racism and the Press. London:Routledge.
- Veglis, A. (2005). Implementation of a computer supported collaborative work systemin a newspaper. WSEAS Transactions on Information Science and Applications, 2(7), 891-901.
- Weber, R. P. (1990). Basic content analysis. Beverly Hills, CA: Sage.
- Webindia123. (n.d.). India Government. Retrieved from http://www.webindia123.com/government/intro.htm
- YijunGao, Liwen Vaughan, (2005) "Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada, and China", Aslib Proceedings, Vol. 57 Issue: 5, pp.398-411,