International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Curated Datasets for Use in Automated Media Monitoring and Feedback System: "News Classification System" Dataset, "Government News Classification" Dataset

Dr. Deepak S. Uplaonkar[1], Santosh Sarvade[2], Harshal Paratwar[3], Vitthal Waghere[4], Chaitanya Ambekar[5]

[1]Department of Computer Engineering,
I^{2}IT
Pune, India
e-mail: uplaonkar@gmail.com
[2]Department of Computer Engineering,
I^{2}IT
Pune, India
e-mail: sarvadeadarsh7@gmail.com
[3]Department of Computer Engineering,
I^{2}IT
Pune, India
e-mail: harshalparatwar72469@gmail.com
[4]Department of Computer Engineering,
I^{2}IT
Pune, India
e-mail: wagherevitthal777@gmail.com
[5]Department of Computer Engineering,
I^{2}IT
Pune, India
e-mail: chaitnaya.a@gmail.com

Abstract— Online journalism in India, a growing field that involves news websites and Digital media, connects with the Press Information Bureau (PIB), a government agency dedicated to sharing accurate information about government policies and initiatives with journalists. While various news outlets publish diverse articles and opinions on these topics, the government seeks to leverage Artificial Intelligence and Machine Learning for gathering feedback in multiple languages. To develop such a system, a notable obstacle is the lack of a readily accessible standard dataset is required. To address this, two datasets are developed named, 'NCS' and 'GNC,' consisting of information from 2020 to 2023 and collected through web scraping tools like Parsehub and manually scrapping. NCS represents News Classification system dataset and GNC represents Government News Classification. The 'NCS' dataset includes Indian news in Hindi, Marathi, and English with categorization of Indian news as government-related or not. Then, a Machine Learning model called "Government News Classifier" to sort news articles using the 'NCS' dataset into either government-related or non-government-related categories. The objective is to use this model to figure out if a news source is discussing topics related to the government or not. Using this model, we created the 'GNC' dataset, which contains only news articles related to government schemes and policies in Hindi, Marathi, and English. In GNC dataset, Human experts manually classify each news source into three categories: "government favourable," "government non-favourable," or "neutral." In essence, this research emphasizes the importance of having access to a large dataset, which can stimulate more advanced prediction models in this complex field.

Keywords-Dataset, classification, MLP, NLP, GNC, NCS, Government Schemes, PIB, Sentiment.

I. Introduction

The Press Information Bureau (PIB) is a government agency in India responsible for sharing accurate information about government policies, initiatives, and achievements through various media channels. It acts as a bridge between the government and the media, promoting effective communication between the two. PIB publishes official government policies and facts, provides support to government organizations, holds press conferences, releases news, and interacts with the media [5]. PIB not only shares information but also plays a crucial role in ensuring that the correct information is disseminated, especially on matters of national importance. It actively identifies areas where the public seeks information and where additional details are required to meet public needs [5].

In India, there's a significant shift towards digital media, especially among younger generations who prefer getting their news from sources like search engines, social media, and news aggregator platforms. Print media is facing challenges, including increased competition, the move of advertising to digital platforms, and reliance on government advertisements. While print media isn't declining as much as it is in the West, the transition to digital is still relatively new and could change in the future. The trend clearly shows that digital media is on the rise. This highlights the growing importance of digital media in the evolving media landscape [1]. In India, there are a total of 392 news publishers, and many of these are in regional languages. These channels are mostly owned by private companies and are influenced by politics and business interests [3].

What's interesting is that in India, a lot of people get their news on their mobile phones. In fact, 73% of people use their smartphones to access news, while only 37% use computers [1]. As of 2022, there are 759 million active internet users in India, meaning people who use the internet at least once a month. By 2025, this number is expected to increase to 900 million, marking the first time that a majority of Indians are regular internet users [2].

As the Increased internet users and digital media Consumers, we are bombarded with rumors. People don't know the truth and now it's a matter of life and death. In 2018, a series of mob lynchings occurred in various parts of India, fueled by rumors spread on social media platforms like WhatsApp [8]. In one case, a video of child kidnappers was widely shared, leading to the lynching of innocent people. These incidents highlighted the deadly consequences of false information. An automatic algorithm should be developed to detect misinformation in the medical field. Fake news also affects the physical health of the public and doctors. Forged documents leading to lynching of innocent people have become a new trend in India. Social media spreads rumors and rumors quickly [14].

False information on social media poses a threat to freedoms and the whole world. Although its potential impact has been widely discussed, there is little evidence of the extent to which the problem has improved in recent years.[8] It is important to check the accuracy of news to prevent the spread of misinformation. Due to time constraints, users often believe in attractive names and images. As a result, sensational news headlines lead to misunderstandings and lies [4]. There are several motives for spreading false information, including manipulating opinions to mislead people, and influencing crucial political events. The COVID-19 pandemic has caused people to react with intense fear and increased reliance on social media, leading to the spread of misinformation [10].

Example of public opinion is like - A new NDTV survey titled 'Public Opinion' analyses the performance of Prime Minister Narendra Modi's government in the last nine years. Despite Karnataka's recent setbacks, more than 55 per cent of the 7,000 respondents from 19 states and 71 Lok Sabha constituencies were satisfied with the government's performance and the results were surveyed. This shows that the Prime Minister's popularity is still strong and PPP polls are still stable [9].

The introduction of online media in a democratic society has both positive and negative effects. On the upside, it allows more people to access political information, share their opinions, and engage with the government. However, the rise of new media alongside the spread of false information poses a problem. There aren't strong safeguards against false news, and the focus on scandals over serious journalism weakens the press's role as a watchdog. The media's sometimes cozy relationship with politicians makes them complicit in spreading incorrect information. While there's never been a perfect era for journalism, the current situation may be a particularly challenging time for a free press in democracy [11].

Public opinion has a noticeable impact on government policy decisions, with around 75% of cases showing its effect when assessed. Furthermore, in at least one-third of these cases, public opinion significantly affects government policy, and it likely does so even more frequently. Even when taking into account the influence of interest groups, political parties, and powerful individuals, public opinion still plays a significant role in shaping policies [12].

This paper provides a dataset comprising NCS and GNC data from 2020 to 2023, with a focus on events during this time frame. The information within the dataset is gathered through web scraping tools like Parsehub. This paper introduces Government News Classifier which classifies news into government and non-government related news. GNC dataset is created using Government News Classifier and web scrapping.

In essence, this paper summarizes its major contributions as follows:

1. This paper presents the Government News Classification Dataset (GNC), a comprehensive dataset featuring Indian government news in multiple regional languages. GNC is a pioneering resource, representing the first large-scale dataset available in the Indian context for conducting sentiment analysis related to government news.

2. This paper introduces the Indian News Dataset, also known as the News Classification System dataset (NCS). NCS is the pioneering dataset in India, open to the public. It contains news articles in Hindi, Marathi, and English classified into government and non-government categories.

3. This paper introduces Government News Classifier which is ML classification model for classification of Government and non-government related news articles. This model uses NCS dataset, in this context, serves as the primary source of dataset for model building.

4. This paper also introduces News Categorization System. Categories are like sports, Education, Finance, science and health, health care, world etc.

The paper is organized into various sections, and one of them is "Related Work," where relevant prior research is discussed. “GNC Dataset” elaborates the GNC dataset. “NCS Dataset” elaborates the NCS dataset used for classifier Model Construction. "Government News Classifier" elaborates construction of Classifier model. “Analysis of the dataset” gives analysis of a proposed dataset.

II. BACKGROUND AND RELATED WORK

When it comes to analyzing government news headlines, it hasn't received as much attention as studying people's opinions about products or services. Still, there are some similar studies out there. Many of these studies have reported findings based on small datasets, or they use datasets that aren't easy for others to access. Some studies even focus on languages and regions different from what we're looking at in our Indian government news dataset.[15] One reference we found mentions a dataset related to news sentiment, but it's quite small, and it's not available for public use. So, while there is some related research, the specific area of analyzing sentiment in Indian government news in multiple regional languages isn't as well-explored as other topics like consumer opinions.[15]

Social media platforms have become really important for bringing people together and letting them share ideas, information, and knowledge. They have a lot of influence and are getting even more popular. Social media platforms are often referred to as the "Big Data" of the world because many people spend a significant amount of time on their devices using them. Research indicates that these platforms have a substantial impact on the behaviors and habits of their users. [17].

Authors in study [16] introduce a Marathi Sentiment Analysis Dataset comprising approximately 16,000 tweets. Marathi Dataset is labelled manually classified into positive, negative, and neutral tweets. They plan to establish a benchmark for future comparisons in the field of sentiment analysis.

In this paper [18], the authors introduce a foundational framework for aspect-based sentiment analysis in Hindi. They undertook a systematic approach, which involved data collection from diverse online sources, thorough data pre-processing to ensure data quality, and meticulous annotation of the dataset with aspect terms and corresponding polarity classes. This dataset is primarily composed of Hindi product reviews, sourced from a range of online platforms spanning across 12 different domains.

In some previous research, they used lexicon-based methods to figure out the sentiment in news articles based on who is mentioned and how they are mentioned. The best outcomes they got in classifying these mentions were around 82% accurate.[19]

In more recent work, they looked at how news can be biased by finding biased phrases and deciding whether they have a positive or negative meaning. They used a model called BERT for this, and it attained an accuracy of approximately 43% at the sentence level and 18% at the word level.[20] Addressing bias in news can be done in different ways. For instance, some studies have tried to find and fix biased phrases in Wikipedia articles and real news using BERT, and they reached around 75% accuracy in detection.[21] Another topic they looked at is "stance detection" in tweets. This means figuring out whether the person who wrote a tweet is in favor of or against a certain idea. They didn't focus on specific entities in these tweets for this research.[22]

In a study [23], they introduced a dataset for understanding how people feel about things (sentiment) in sentences from randomly chosen news articles. This dataset has about 3,163 English sentences. However, the researchers acknowledge that a single sentence isn't enough to really grasp the overall sentiment in a news story. They say you need the bigger picture. This study employs the BERT approach to investigate the stance of Indonesian people regarding government policies on vacations during the COVID-19 pandemic [17]. It creates a new dataset from news outlet stories and YouTube comment sections, classifying public sentiment as positive, neutral, or negative. Using Python and the Flask framework, an application for sentiment analysis is built, achieving an F-score of 84.33%.

In comparison, we created a dataset using Indian news headlines, Description, author, Date and Source etc. NCS and GNC datasets offer a unique and valuable resource for their specific focus on Indian government-related news and sentiment analysis. While many datasets cover broader topics, these datasets cater to a niche area, providing specialized information.

Their inclusion of multiple regional languages further distinguishes them, making them a valuable resource for research in the Indian context.

III. DATASET

A. The NCS (News Classification System) Dataset

Our most recent NCS package incorporates content in three distinct languages: Marathi, Hindi, and English. This collection is the result of contributions from a minimum of five writers, each of whom are experts in their respective domains or media specialists who employed the doccano annotation tool. Refer to Table I for an in-depth breakdown of these particulars. The NCS dataset, denoted as the News Classification System dataset, proves to be a valuable resource, catering to the needs of both organizations and individuals striving to categorize and comprehend news across multiple languages. It is made accessible in Marathi, Hindi, and English, making it particularly advantageous for businesses and other entities that seek to grasp the intricacies inherent to each language and brand.

The NCS dataset serves as a foundational resource comprising news titles with accompanying descriptions, publication dates, author information, and categories. The primary purpose of this dataset is to provide a diverse collection of news content categorized into two main segments: government and non-government. This comprehensive dataset sets the stage for the development of subsequent datasets, such as the Government News Classification (GNC) dataset. For the task of categorizing real-time news into the binary classes of government and non-government, we experimented with various machine learning models. These models include the powerful BERT (Bidirectional Encoder Representations from Transformers), Random Forest, and a Multilayer Perceptron (MLP). The results of our model in Table III for evaluations show that BERT achieves the highest accuracy among the tested models, consistently reaching 90% and above. This exceptional performance can be attributed to BERT's [24] deep contextual understanding and attention mechanism, which excels at capturing nuances in language and context, making it an ideal choice for news classification tasks.[26]. In the context of creating the GNC dataset, the NCS dataset acts as a precursor by providing a foundational framework for government news classification. By designating news articles into government and non-government categories, the NCS dataset serves as a pivotal building block for the subsequent dataset.

B. The GNC (Government News Classification) Dataset

The GNC dataset represents a significant milestone in the domain of classifying government news articles into specific departments or categories, such as education, transport, sports, and more. The objective of the GNC dataset is to facilitate the development and training of machine learning models and algorithms that can automatically classify and analyze government-related news content. By structuring the dataset into specific departments, it enables more granular classification, making it easier for AI models to understand the context and relevance of each news piece. Furthermore, the inclusion of government schemes as categories adds another layer of contextual information to the dataset. This ensures that the classification goes beyond mere department labels, allowing for a deeper understanding of how the news articles relate to the government's policy initiatives and programs. This aspect of the dataset is particularly valuable for applications in public policy analysis, government communication, and citizen engagement.

Annotating the sentiment of each news article as positive, negative, or neutral provides an additional dimension for analysis see Table II. It allows for the assessment of public perception and sentiment towards government schemes and initiatives. This sentiment analysis component has applications in assessing the effectiveness of government programs, identifying areas that may require improvement, and gauging public sentiment, which can be crucial for policymakers and government officials.

IV. Data Collection, Selection and Processing

The headlines were sourced from a prominent online news outlet, specific to a particular language. We amassed a total of 300,000 articles in Marathi, 300,000 in Hindi, and 300,000 in English. During the collection process, it was observed that the English dataset covered news articles from as early as January 1, 2020, to the most recent articles up to October 28, 2023. Similarly, the Marathi and Hindi datasets exhibited the same time range. To maintain uniformity and ensure a consistent number of headlines from each news portal in our research, we made the deliberate choice to limit our data collection to a specific period, covering the years 2020 to 2023. During this period, we manually selected headlines that included specific named entities. Our selection criteria encompassed a variety of news categories, including government and non-government news, as well as various government departments such as Education, Sports, Road, and Transport. This decision was informed by the fact that some automated tools may overlook government departments or government news mentions, particularly when they occur at the beginning of a sentence—a frequent occurrence in news headlines.

A. Data Annotation Process

Data preprocessing, cleansing and annotation:

Annotators are provided with the following annotation guidelines:

- Classify news articles in the NCS dataset into government and non-government categories.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

- In the GNC dataset, categorize news into government departments and determine news sentiment.
- Assess whether news articles in the GNC dataset portray a negative, neutral, or positive sentiment.
- Sentiment can be expressed through clear statements or opinions about entities.
- Consider both the sentiment of the entire headline and that of the paragraph describing the news.
- Use the "neutral" label when sentiment determination is not possible, such as in cases of mixed sentiment or strong context dependency.
- Maintain a neutral and objective perspective, avoiding personal opinions when annotating news articles.

The annotation process for the NCS and GNC dataset was meticulously carried out using the doccano annotation tool [28]. A team of experts and experienced News Editors spearheaded the annotation effort [29]. Annotators were tasked with assigning labels to each headline based on its content. These labels encompassed a range of topics, including politics, sports, and entertainment. The entire annotation process adhered to the rigorous guidelines provided by the National Institute of Standards and Technology (NIST), which emphasized data quality, accuracy, and consistency. The annotation process was carried out by at least 5 annotators from the expert team or by expert news editors. Moreover, the dataset can serve as valuable training data for machine learning models, facilitating automated text summarization and headline generation. The GNC dataset is annotated with positive, negative and the NCS dataset annotated by Government News and Non-Government News.

B. Text Preprocessing

Text preprocessing is a vital task for sentiment analysis and other data analysis methods. It involves cleaning and organizing the text data to improve the results. Without preprocessing, the data may be inaccurate and hard to analyze. This article will explain the main steps of text preprocessing [30].

- The first step in preprocessing is case folding. This step standardizes all the text in the sentence or document. Lowercase or non-capitalized letters are usually preferred as they are simpler to read and analyze.
- The second step is character deletion. It removes all the irrelevant features like uniform resource locator (URL), numbers, and punctuation. This step helps to reduce the noise from the sentences.
- The third step is tokenizing. This step splits the text or sentence into individual words. This process is helpful for creating word vectors which are also known as bag of words (BOW).
- Fourth step is stop-word removal. It eliminates common but unimportant terms from sentences. This step decreases the corpus size without losing the key information.
- Fifth step is stemming. It is the process of simplifying words into their basic forms. This step reduces the complexity of the token’s words. During this step, all affixes like prefixes, infixes, or suffixes will be deleted from the word.
- The sixth step is lemmatization. It is a more advanced method than stemming. Lemmatization transforms words into their meaningful root forms. During this step, the affixes will be deleted, and the root word will be retrieved.

C. Outlier Detection

To detect bias among annotators, you can use the Bhattacharyya distance or other statistical methods. The Bhattacharyya distance measures the similarity between two probability distributions, in this case, the distribution of sentiment scores for news articles [31].

The Bhattacharyya distance (BD) between two distributions can be calculated using the following equation [32]:

$$BD = -\ln\left(\sum \sqrt{P_i Q_i}\right) \quad (1)$$

Where:

- BD is the Bhattacharyya distance.
- Pi and Qi are the probabilities of sentiment scores in two different distributions (e.g., scores for two different annotators).
- The sum is taken over all possible sentiment scores.

D. Goverenment News Classifier

The model, named the "Government News Classifier" is designed for real-time news classification into government and non-government categories, leveraging the NCS dataset as its core framework. For the purpose of Final dataset (GNC), we have to prepare model on NCS (News Classification) dataset. The process of model preparation is crucial for achieving accurate and efficient results in real-time news classification into government and non-government categories. We compared three different machine learning models: BERT, Random Forest, and a Multilayer Perceptron (MLP) to evaluate their performance and suitability for the task. BERT (Bidirectional Encoder Representations from Transformers): BERT, a transformer-based model, was the best performer among the three. It consistently achieved accuracy rates above 90%, showing its remarkable ability to understand and classify real-time news articles [33]. BERT’s main strength lies in its deep contextual understanding of language, which enables it to capture subtle nuances and context in news content. This is

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

especially important in news classification, where context and language subtleties are essential. Based on its impressive performance, BERT was chosen as the preferred model for real-time news classification. Random Forest [34]: The Random Forest model, while offering decent accuracy in the range of 80-85%, showed slightly lower performance compared to BERT. Random Forest is known for its ensemble learning capabilities, which can handle various tasks, but it did not match the accuracy of BERT for this specific news classification task. Multilayer Perceptron (MLP): The MLP model delivered accuracy within the range of 80-90%. While it showed competitive performance, BERT's consistent accuracy above 90% made it the more attractive option for real-time news classification.

In conclusion, the selection of BERT as the preferred model for the NCS dataset was influenced by its superior accuracy and the critical ability to handle the complexity of real time news articles. BERT's deep learning and contextual comprehension capabilities are key in accurately classifying government and non-government news content.

V. Analysis of Dataset

The GNC (Government News Classification) and NCS (News Classification System) datasets are integral components of a broader effort to enhance news classification and analysis. These datasets offer distinct but complementary features that contribute to a more comprehensive understanding of news articles. The GNC dataset focuses on classifying government-related news articles into specific departments and categorizing them based on related government schemes. By also annotating sentiment, it provides insights into public perception. This dataset's applications span policy analysis, government communication, and public sentiment evaluation. In contrast, the NCS dataset acts as a bridge, offering a binary classification of news articles into government or non-government categories. This distinction serves as a foundational step before further granular categorization. BERT, Random Forest, and MLP model evaluations on NCS highlight the superiority of BERT for real-time news classification.

Together, these datasets pave the way for more precise and context aware news analysis. GNC takes a finer-grained approach, while NCS simplifies the initial categorization. Combining these datasets and harnessing advanced models like BERT ensures accurate classification, contributing to more informed decision-making and public communication. In summary, GNC and NCS datasets collectively support the development of an advanced news classification system, encompassing government and non-government articles, departments, schemes, and sentiment analysis. These resources empower data scientists.

A. *Figures and Tables.*

The "NCS dataset" is a tabular compilation of news articles, featuring essential attributes such as title, language, URL, source, author, date, category, description, and classifiers for each entry. This structured dataset facilitates the efficient organization and analysis of news articles, serving as a valuable resource for research and data-driven insights. With columns providing insights into article content, language, source, authorship, publication date, and category, it proves versatile for tasks such as sentiment analysis for the GNS dataset, content categorization, source tracking, and temporal trend analysis. This dataset empowers researchers and analysts to delve into news content, gain an understanding of sentiments, and identify emerging trends, thereby enhancing their comprehension of the ever-evolving news landscape. Furthermore, the dataset's flexibility allows for easy expansion to encompass a diverse range of news entries, making it an adaptable and versatile tool for investigations related to news and media.

The "GNS dataset" is an extensive tabular dataset comprising a diverse set of attributes for each news article it contains. These attributes encompass the title, language, URL, news source, author, associated departments, government initiatives discussed, publication date, a descriptive summary of the article's content, and an evaluation of the expressed sentiment within the article. This dataset enables in-depth exploration of news articles across various languages, source tracking, authorship examination, departmental categorization, identification of government programs, temporal analysis of news publication, content analysis, and sentiment assessment, facilitating a holistic comprehension of the news landscape. The dataset serves as a valuable resource for research, sentiment analysis, and data-driven investigations into news articles spanning a wide range of topics and languages.

TABLE I. NCS Dataset

Title	*Language*	*URL*	*News Source*	*Author*	*Date*	*Description*	*Classifiers*
COVID-19 Vaccine Distribution Begins	English	https://example.com/news/123	NDTV	John Doe	2023-01-15	The distribution of COVID-19 vaccines has started in various regions.	Government
Tech Innovation Summit	English	https://example.com/news/101	TechInsider	Sarah Johnson	2023-04-05	The latest technological innovations were revealed at the Tech Innovation Summit,	Non-Government

Unveils Future Trends						showcasing future trends in the industry.	
रोजगार के अवसरों की खोज	Hindi	https://example.in/news/789	सुचना समाचार	आदित्य शर्मा	2023-02-20	रोजगार के लिए नई अवसरों की खोज करने के लिए सरकार का आलान।	Government

TABLE II. GNC DATASET

Title	***Languag e***	***URL***	***News Source***	***Author***	***Departments***	***Government Scheme***	***Date***	***Description***	***Sentiment***
New Infrastructure Investment Bill	English	https://example.com/news/789	Infrastruc ture Times	Sarah Johnson	Infrastructure	Infrastructure Investment Bill	3/25/2023	The government has introduced a new bill to boost infrastructure development.	Positive
Stock Market Records All-Time High	English	https://example.com/news/987	Financial News	Emma Brown	Finance, Economy	Stock Market Growth	4/5/2023	The stock market has reached a new all-time high, reflecting strong economic growth.	Negative
गरीबों के लिए आवास योजना	Hindi	https://example.in/news/567	गरीबी मुक्ति समाचार	राजेश कुमार	Housing	Affordable Housing Scheme	3/10/2023	सरकार ने गरीबों के लिए आवास योजना की घोषणा की है।	Positive

TABLE III. CLASSIFICATION MODEL

ML Models	***Accuracy (on average)***
Random Forest	80% - 85%
Multilayer Perceptron	80% - 90%
BERT	Exceeding 90%

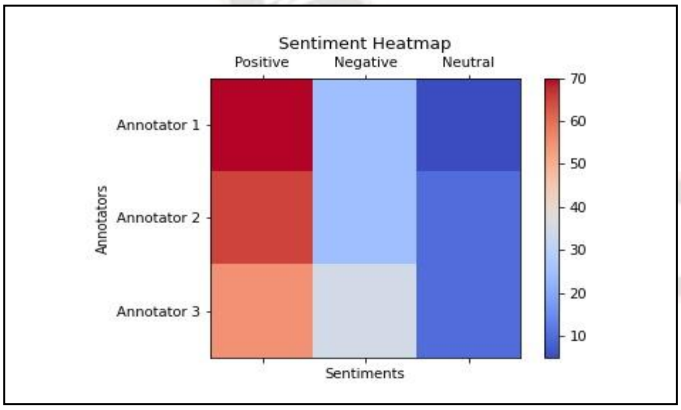

Figure 1. Sentiment Heatmap.

It represents the distribution of sentiments among three annotators: "Annotator 1," "Annotator 2," and "Annotator 3." Sentiments are categorized as "Positive," "Negative," and "Neutral." The heatmap's color intensity varies to depict the percentage of each sentiment category, with warmer colors indicating higher percentages and cooler colors indicating lower percentages. The color scale helps quickly identify sentiment preferences for each annotator, allowing for easy comparison.

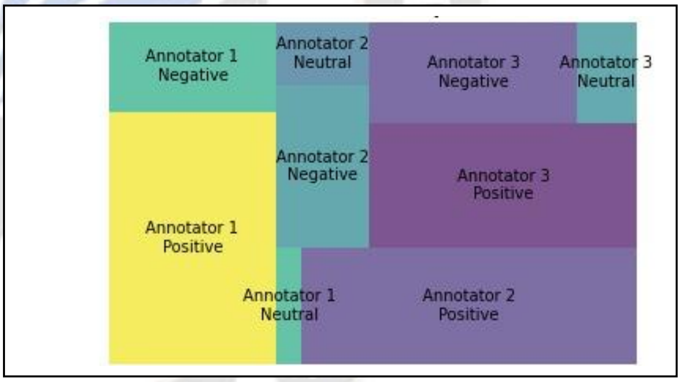

Figure 2. Distribution of Sentiment by Annotator.

In Figure 2, there is a treemap visualization representing the sentiment distribution among three annotators (Annotator 1, Annotator 2, and Annotator 3) across three sentiment categories: Positive, Negative, and Neutral. Each square within the treemap corresponds to a specific annotator-sentiment combination, with the square's size indicating the proportion of that sentiment category for the respective annotator.

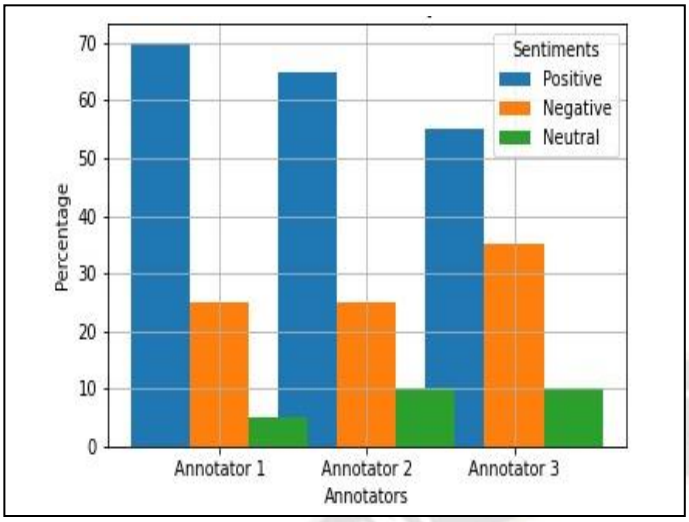

Figure 3. News Sentiment Analysis by Annotator.

In Figure 3, a visual comparison is provided for sentiment percentages among three annotators: "Annotator 1," "Annotator 2," and "Annotator 3." Sentiments are categorized into "Positive," "Negative," and "Neutral." Each annotator's distribution is represented by side-by-side bars, facilitating an easy visual comparison of how sentiment preferences vary among annotators. The chart's x-axis displays annotators, while the y-axis shows the percentage of sentiment categories. The legend distinguishes the sentiment categories with distinct colors, enhancing the chart's readability. This visualization offers a clear and concise overview of sentiment distribution among the annotators.

VI. CONCLUSION

This paper introduces novel datasets with an Indian perspective for Government News Classification (GNC) and News Classification System (NCS). These datasets represent an innovative collection of news articles from India, encompassing multiple regional languages and focusing on government policies and schemes. These are first Datasets for Indian news Articles which are in Multiple Regional languages and related to government policies and schemes. These datasets will form important pillars for research of Indian Government news for researches. GNC dataset consists of news that are classified using Government News Classifier to extract news only related to government scheme, policy, standards of India. Government News Classifier is a Classification model for government related news. Proposed approach to model is using BERT Classification is giving more accuracy than Random Forest and Multi-layer Perception. Dataset are verified and labelled by Human Annotators. GNC dataset will be valuable Dataset for Sentiment analysis for Indian news. This will help government to closely analyze feedback of public. This will result in substantial influence of public feedback on Government Policies.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ajitkumar Shitole, head of department Computer Engineering, I^{2}IT for continuous guidance and support.

REFERENCES

[1] The future of news in India, report https://vidhilegalpolicy.in/wphttps://vidhilegalpolicy.in/wp-content/uploads/2020/07/Vidhi_12052020_Edited.pdfcontent/uploads/2020/07/Vidhi_12052020_Edited.pdf

[2] Over 50% Indians are active internet users now; base to reach 900 million by 2025 https://www.moneycontrol.com/news/business/internet-users-in-india-set-to-reach-900https://www.moneycontrol.com/news/business/internet-users-in-india-set-to-reach-900-million-by-2025-report-10522311.htmlmillion-by-2025-report-10522311.html

[3] https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/india

[4] Political Effects of the Internet and Social Media Ekaterina Zhuravskaya,1 Maria Petrova,2,3,4,5,6 and Ruben Enikolopov3,2,4,5,6 1Paris School of Economics, École des Hautes Études en Sciences Sociales, 75014 Paris, France; 2Department of Economics and Business, Universitat Pompeu Fabra, 08002 Barcelona, Spain 3New Economic School, Moscow 121353, Russia 4Institute of Political Economy and Governance, 08005 Barcelona, Spain 5Graduate School of Economics, 08005 Barcelona, Spain 6Catalan Institute for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.

[5] PIB (Press Information bureau) https://en.wikipedia.org/wiki/Press_Information_Bureau https://www.pib.gov.in/aboutpibn.aspx

[6] Allcott H, Gentzkow M, Yu C. 2019. Trends in the diffusion of misinformation on social media. Res. Politics 6. https://doi.org/10.1177/2053168019848554

[7] Kastrati, Z.; Dalipi, F.; Imran, A.S.; Pireva Nuci, K.; Wani, M.A. Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. Appl. Sci. 2021, 11, 3986. PY - 2021/04/28 VL - 11 DOI - 10.3390/app11093986 JO - Applied Sciences

[8] https://timesofindia.indiatimes.com/readersblog/world-of-words/fake-news-and-socialmedia-33975/

[9] https://www.ndtv.com/india-news/ndtv-public-opinion-9-years-of-modi-government-over55-people-satisfied-with-centres-work-on-various-fronts-4059500

[10] COVID-19 pandemic: An era of myths and misleading advertisements Tarif Hussian1 , Manjusha Choudhary2 , Vikas Budhwar1 and Garima Saini1 PY - 2021/01/20 SP - 174113432098832 VL - 17 10.1177/1741134320988324 Journal of Generic Medicines: The Business Journal for the Generic Medicines Sector

[11] https://www.bbvaopenmind.com/wp-content/uploads/2018/03/BBVA-OpenMind-Diana-Owen-The-New-Medias-Role-in-Politics.pdf

[12] The Impact of Public Opinion on Public Policy: A Review and an Agenda Author(s): Paul Burstein Source: Political Research Quarterly, Vol. 56, No. 1 (Mar., 2003), pp. 29-40 Published by: Sage Publications, Inc. on behalf of the University of Utah Stable URL: http://www.jstor.org/stable/3219881

[13] Making What Government Does Apparent to Citizens: Policy Feedback Effects, Their Limitations, and How They Might Be Facilitated By Suzanne Mettler Suzanne Mettler is John L. Senior Professor of American Institutions in the Government Department at Cornell University.

Correspondence: suzanne.mettler@cornell.edu DOI: 10.1177/0002716219860108

[14] IFND: a benchmark dataset for fake news detection Dilip Kumar Sharma1 · Sonal Garg1 Received: 2 June 2021 / Accepted: 21 September 2021 / Published online: 16 October 2021 © The Author(s) 2021 Complex & Intelligent Systems (2023) 9:2843–2863 https://doi.org/10.1007/s40747-021-00552-1

[15] A dataset for Sentiment analysis of Entities in News headlines (SEN) Katarzyna Baraniaka,*, Marcin Sydowa,b aPolish-Japanese Academy of Information Technology, Koszykowa 86, Warsaw 02-008, Poland Institute of Computer Science, Polish Academy of Sciences, Poland

[16] L3CubeMahaSent: A Marathi Tweetbased Sentiment Analysis Dataset Atharva Kulkarni1,3, Meet Mandhane1,3, Manali Likhitkar1,3, Gayatri Kshirsagar1,3, and Raviraj Joshi2,3 1Pune Institute of Computer Technology, Pune 2Indian Institute of Technology Madras, Chennai 3 L3Cube, Pune {k.atharva4899,meetmandhanemnm,manalil1806,gayatrimohan7}@gmail.com ravirajoshi@gmail.com

[17] Analysis of Government Policy Sentiment Regarding Vacation during the COVID-19 Pandemic Using the Bidirectional Encoder Representation from Transformers (BERT) Intan Nurma Yulita 1,* , VictorWijaya 2, Rudi Rosadi 2, Indra Sarathan 3 , Yusa Djuyandi 4 and Anton Satria Prabuwono 5

[18] Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation Md Shad Akhtar, Asif Ekbal and Pushpak Bhattacharyya Department of Computer Science and Engineering Indian Institute of Technology Patna Patna, India {shad.pcs15,asif,pb}@iitp.ac.in

[19] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J., 2010. Sentiment analysisin the news, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta. URL: http://www.lrecconf.org/proceedings/lrec2010/pdf/909_Paper.pdf.

[20] Fan, L., White, M., Sharma, E., Su, R., Choubey, P.K., Huang, R., Wang, L., 2019. In plain sight: Media bias through the lens of factual reporting, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6344–6350.

[21] Pryzant, R., Martinez, R.D., Dass, N., Kurohashi, S., Jurafsky, D., Yang, D., 2019. Automatically neutralizing subjective bias in text. arXiv preprint arXiv:1911.09709

[22] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C., 2016. Semeval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41

[23] Hamborg, F., Donnay, K., Gipp, B., 2021. Towards target-dependent sentiment classification in news articles, in: Proceedings of the iConference 2021.

[24] Sentiment analysis classification system using hybrid BERT models Talaat Journal of Big Data (2023) 10:110 https://doi.org/10.1186/s40537-023-00781-w Journal of Big Data

[25] Sae-Mi Lee, Seung-Eui Ryu, Soon-Jae Ahn, "Auto-Classification of Government Department-Specific News Articles" 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 10.1109/CSCI49370.2019.00286, 2019.

[26] Salsabila Mazya Permataning Tyas, Riyanarto Sarno;, Agus Tri Haryono, Kelly Rossa Sungkono,"A Robustly Optimized BERT using Random Oversampling for Analyzing Imbalanced Stock News Sentiment Data","International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)",2023

[27] Sanjeev Verma,"Sentiment analysis of public services for smart society: Literature review and future research directions",Government Information Quarterly,july 2022

[28] Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X., 2018. doccano: Text annotation tool for human. URL: https://github. com/doccano/doccano. software available from https://github.com/doccano/doccano.

[29] Martin Sarnovský; Viera Maslej-Krešňáková; Nikola Hrabovská,"Annotated dataset for the fake news classification in Slovak language",18th International Conference on Emerging eLearning Technologies and Applications (ICETA),2020

[30] Text Preprocessing for Machine Learning & NLP - Kavita Ganesan, PhD1, https://kavita-ganesan.com/text-preprocessing-tutorial/

[31] Bhattacharyya distance, https://handwiki.org/wiki/Bhattacharyya_distance

[32] "Introduction to the Practice of Statistics" by David S. Moore, George P. McCabe, and Bruce A. Craig.

[33] Bihui Yu; Chen Deng; Liping Bu ,"Policy Text Classification Algorithm Based on Bert",11th International Conference of Information and Communication Technology (ICTech)) ,February 2022

[34] Gunasekar Thangarasu; Kesava Rao Alla,"Detection of Cyberbullying Tweets in Twitter Media Using Random Forest Classification",13th Symposium on Computer Applications & Industrial Electronics (ISCAIE),2023