FakeVerseIndia

COL865

Prof. Abhijnan Chakraborty

By

Harshitha Chandra Jami (2019CH70171)

Harshit Bhalla (2020CH10088)

Shivam Kanojia (2019CS50131)

FakeVerseIndia

Abstract

Our research report, focused on fake news detection and news classification, encapsulates a comprehensive approach towards understanding and combating misinformation. Our journey began with an extensive review of relevant literature and engaging with authors for data acquisition. A significant breakthrough was obtaining the FactDRIL dataset, a rich compilation of fake news from 11 websites, featuring multiple entries per file. We meticulously curated additional data from these sources, successfully completing the dataset for all 10 websites. Our analysis encompassed three pivotal aspects: a thorough Statistical Analysis, a Long Term Analysis assessing topic-wise trends, and a nuanced examination of Left and Right wing narratives in news content. This multi-faceted approach has offered us deep insights into the dynamics of fake news dissemination and its classification.

Here's the github repository to our work: https://github.com/aspirin01/FakeVerseIndia/

Introduction

Over the years due to the advancement of communication technologies information can be easily accessible to all people. The credibility of the information spread on these technologies is a major issue. The spread of misinformation is increasing rapidly. Also, fake news spread during Global events like elections, COVID-19 etc. can affect people badly. In order to mitigate the spread of fake news, fake news checking websites rose from year 2017 as shown in Figure 1. Fact-checking is a systematic procedure aimed at determining the accuracy and truthfulness of widely circulated assertions based on some evidence. Many of the fact-checking websites are in English but in countries like India where people speak different languages many of them cannot communicate in English. In this work, we aim to collect fake news articles published in low-resource Indian languages. This dataset helps towards the development of strategies to address the issue of automated fact-checking in various languages, which has the potential to adversely impact a significant portion of the population.

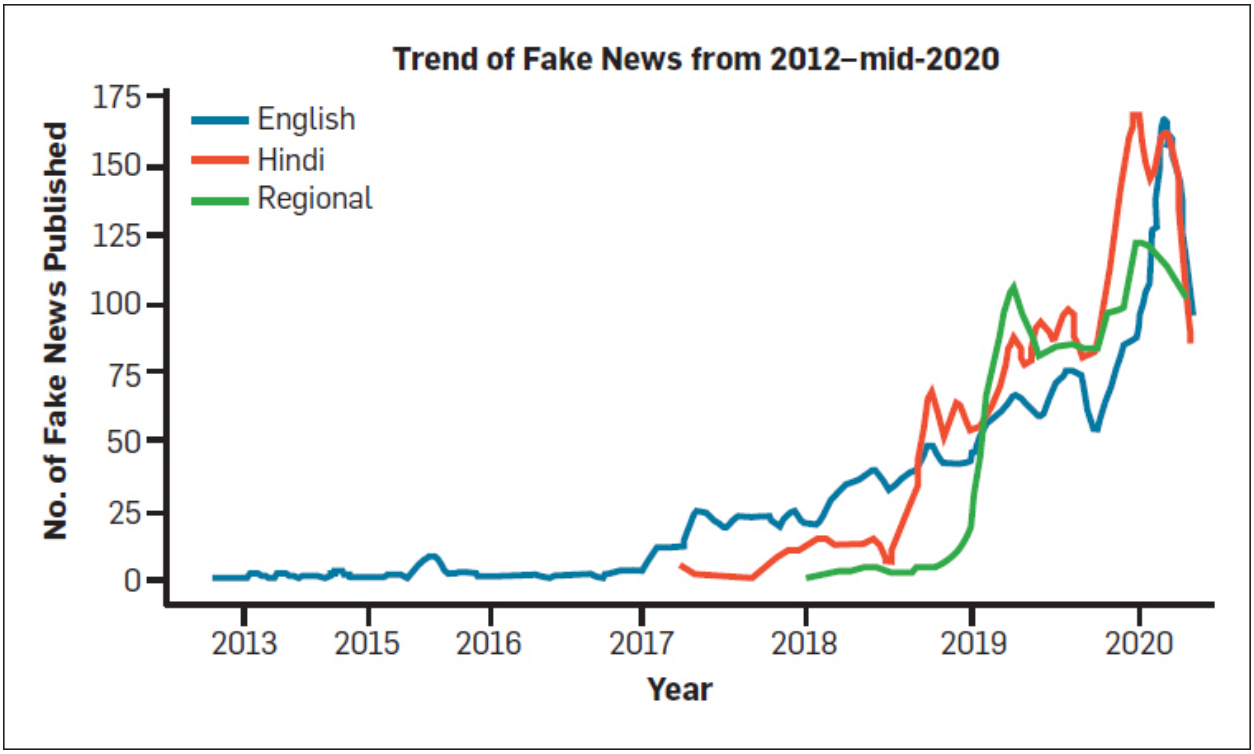


Figure 1: Fake news articles trend over the years

Literature Review:

About FakeNewsIndia

This paper [2] developed an automated data collection pipeline for fake news incidents reported by Indian fact-checkers from June 2016 to December 2019. The resulting dataset, FakeNewsIndia, encompasses 4,803 incidents, with associated 5,031 tweets on Twitter and 866 videos on YouTube. The study evaluates the impact of these incidents on Twitter and YouTube, employing popularity metrics to categorize impact levels as low, medium, and high. This work contributes crucial insights into the pervasive influence of misinformation on prominent online platforms.

About IFND Dataset

This paper [1] introduces the IFND (Indian Fake News Dataset), an extensive benchmark dataset designed for the detection of fake news spanning events in India from 2013 to 2021. The authors

examine the complexities associated with identifying fake news, such as the presence of echo chambers, deceptive writing styles, and limitations in benchmark datasets. They also assess current methods for fake news detection, encompassing text and image features, sentiment analysis, and topic modeling.

The IFND dataset comprises news articles sourced from credible Indian outlets, encompassing the period from 2013 to 2021. Each article undergoes manual verification to confirm its relevance to India and authenticity. Additionally, the dataset includes image content corresponding to each news headline. The creation of the image dataset involves text matching with a genuine news dataset, and an extra 20,000 real news datasets were scraped to eliminate biases in the proposed dataset.

To categorize news statements, the authors utilized LDA topic modeling and provided word cloud representations for visualizing the dataset's news content. In summary, the IFND dataset offers a comprehensive benchmark that includes both textual and visual elements, aiming to enhance the accuracy of current models in fake news detection. The authors explore potential applications and outline future directions for research, making a valuable contribution to advancing the field of fake news detection.

About Aletheia: A Fake News Detection System for Hindi

This paper [4] introduces an innovative approach to address the issue of fake news in the Hindi language. The authors emphasize the escalating problem of misinformation and its societal impact before presenting Aletheia, a system employing machine learning methods for the identification of fake news articles.

To train the system, the authors curated a dataset of news articles from diverse sources, manually labeling them as either real or fake. They utilized this dataset to train multiple machine learning models, including logistic regression, decision trees, and random forests. Notably, the random forest model exhibited superior performance, achieving a noteworthy 92.5% accuracy on the test set.

Aletheia employs a blend of linguistic and contextual features for fake news identification, encompassing emotive language, sensational headlines, and the article's similarity to known fake news pieces. The system also considers the credibility of the news source and the article's engagement on social media.

Evaluation on a distinct dataset of news articles demonstrated Aletheia's effectiveness, yielding an accuracy of 89.5%. Comparative analysis with other fake news detection systems revealed Aletheia's superior performance on the Hindi language dataset.

In conclusion, the authors propose a promising solution to combat fake news in Hindi. Aletheia's integration of machine learning techniques and diverse linguistic and contextual features renders it a robust and efficient system for detecting misinformation. The authors envision its potential application by news organizations and social media platforms as a tool to mitigate the dissemination of fake news.

About FactDRIL

The existing research introduces FactDRIL [3], a significant dataset for fact-checking in Indian regional languages. It addresses the critical issue of fake news, especially in a multilingual country like India, where only a small percentage of the literate population speaks English. Their dataset, compiled over seven months, contains 22,435 samples from 11 languages. Among them, 9,058 samples belong to English, 5,155 samples to Hindi and remaining 8,222 samples were distributed in various regional languages i.e., Bangla, Marathi, Malayalam, Telugu, Tamil, Oriya, Assamese, Punjabi, Urdu, Sinhala and Burmese. FactDRIL is unique in its detailed characterization of multilingual, multimedia, and multi-domain attributes, including a novel feature - investigation reasoning. The authors aim to automate attribute extraction and encourage further research through challenges. FactDRIL serves as a vital resource for studying and combating fake news in low-resource languages, highlighting the importance of regional languages in information dissemination.

Our Aim:

As part of our ongoing efforts in our project, we recognize the need to update and expand our dataset to reflect the current landscape of fake news in India. Existing dataset includes datapoints up until June 2020, but given the dynamic nature of news and misinformation, it's crucial to extend this dataset to the present day. By doing so, we aim to maintain the relevance and effectiveness of fake news and real news, ensuring it remains a comprehensive and up-to-date resource for studying and combating fake news in Indian regional languages. This extension will enhance our understanding of the evolving trends in misinformation and aid in the development of more accurate and efficient fact-checking tools.

Dataset

We extracted 44769 fake news articles from the 10 india verified fact checking websites shown in Table 1. We collected the data from June 2020 to till date in various available languages and in

various domains

Website	Languages	Domains
Alt news	English, Hindi	Politics, Science, Religion, Society
Boom Live	English, Hindi, Bangla, Gujarati	General
DigitEye	English, Telugu, Kanada	General
FactChecker	English	General
India Today	English	General
News Mobile	English	General, Coronavirus
NewsChecker	English, Hindi, Marathi, Punjabi, Gujarati, Tamil, Urdu, Bengal	General
Vishvas News	English, Hindi, Punjabi, Odia, Assamese, Gujarati, Urdu, Tamil, Telugu, Malayalam, Marathi, Bangla	Coronavirus, Politics, Health, Society, World
The Quint-Webqoof	English, Hindi	General, Health
Fact Crescendo	English, Hindi, Punjabi, Odia, Assamese, Gujarati, Urdu, Tamil, Telugu, Malayalam, Marathi, Kannada, Bengali	General, Coronavirus

Table 1: Fact checking websites

4.1 Dataset Attributes

We extracted various features in the following categories:

Meta features: article_link, website_name, published_date

Textual features: title, content, claim, investigation

Media features: image_links, video_links, top_imge, thumbnail

Analysis

There are mainly 3 sections of our analysis:

- Statistical Analysis
- Long-term Analysis

- Left & right wind analysis

5.1 Statistical Analysis:

This is fundamental for understanding the basic characteristics of the dataset, such as distribution of news across different websites, languages, and time periods. It helps in identifying patterns and anomalies in the data, which are essential for building accurate detection models.

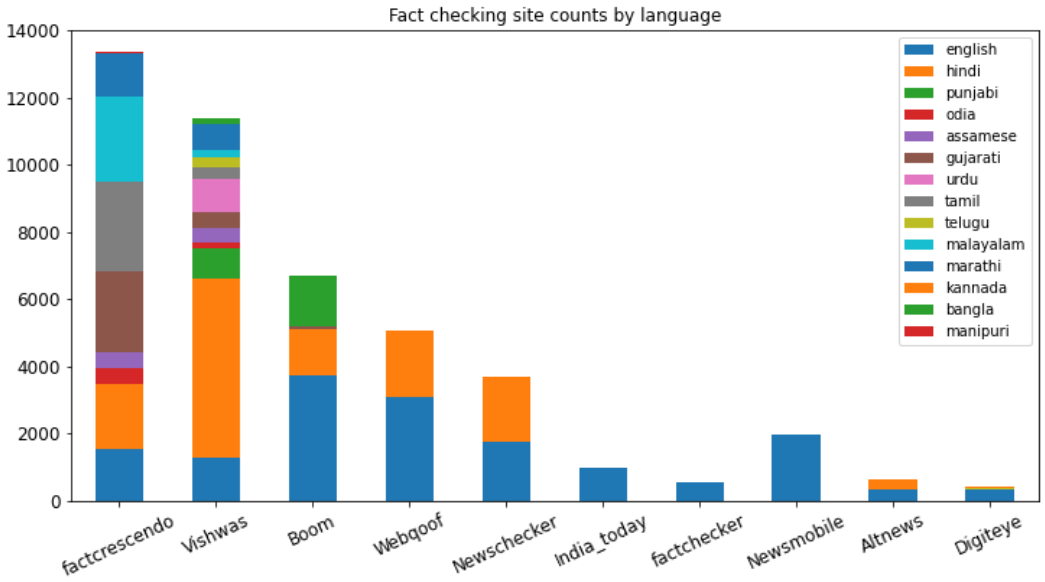


Figure 3: Fact checking site counts by language graph

This bar chart displays the number of fact-checking sites that are available in different languages. The vertical axis represents the count of sites, with numbers ranging from 0 to around 14000, and the horizontal axis lists different fact-checking platforms.

The platforms themselves exhibit considerable diversity in their linguistic offerings, with "factcrescendo" and "Vishvas" providing a broad spectrum of languages, signifying a rich linguistic inclusivity. In stark contrast, platforms like "Digivijay" have a much narrower linguistic range. Hindi, Tamil, Telugu, and Malayalam are also prominent, indicating their significant online presence and potentially a higher demand for fact-checking in these languages. On the other hand, languages such as Manipuri are represented to a far lesser extent, hinting at a digital divide in the availability of fact-checking resources.

To summarize, the chart facilitates a comparative view, highlighting not only the distribution of language services across various platforms but also revealing the extent of fact-checking

infrastructure that caters to India's linguistic diversity. This suggests a landscape where some languages are well-served, while others have limited access to fact verification tools, which could impact the spread and control of misinformation among different language speakers.

Language vs Frequency

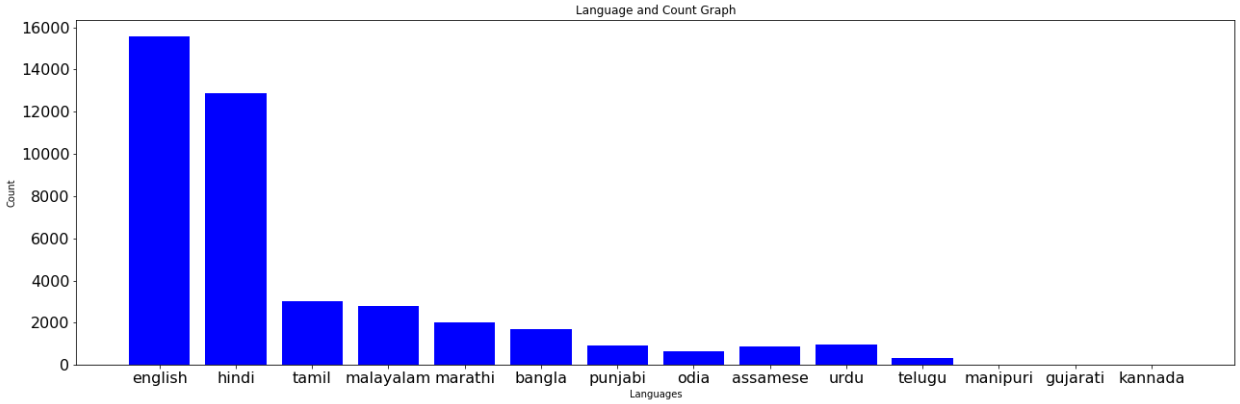


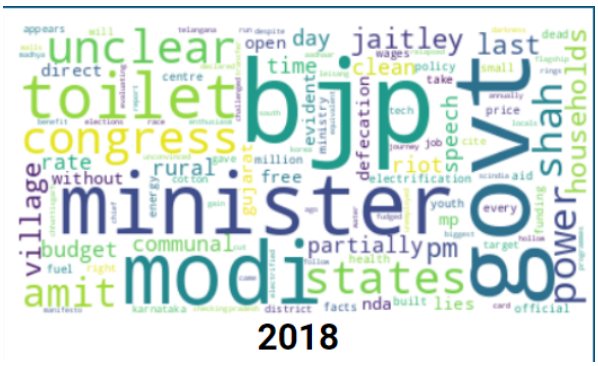
Figure 4: Language vs Frequency

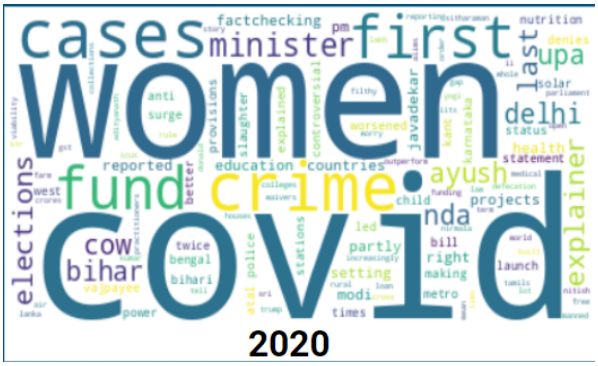
The graph presents a count of occurrences (or frequency) for various languages. The vertical axis is labelled 'count' and it scales up to 16000, while the horizontal axis lists languages. The languages featured are English, Hindi, Tamil, Malayalam, Marathi, Bangla, Punjabi, Odia, Assamese, Urdu, Telugu, Manipuri, Gujarati, and Kannada. From the graph we can observe that, English and Hindi have the highest counts, significantly more than the other languages, indicating that they are likely the most frequent in the context this data represents. Tamil and Malayalam follow as the next most common languages, with counts roughly around the lower four figures, indicative of their substantial, yet smaller presence compared to English and Hindi. Marathi and Bangla are depicted with similar frequencies, which are notably less than those of Tamil and Malayalam but still significant. The middle tier comprises Punjabi, Odia, Assamese, Urdu, and Telugu, each represented by progressively smaller bars. At the lower end of the spectrum, Manipuri, Gujarati, and Kannada have the smallest counts, suggesting a relatively infrequent occurrence within the dataset. This graph serves as a clear representation of the linguistic landscape, highlighting the dominance of certain languages and the comparative rarity of others.

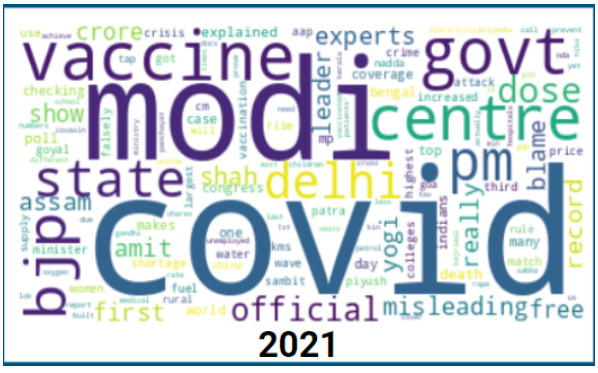
5.2 Long-term Analysis:

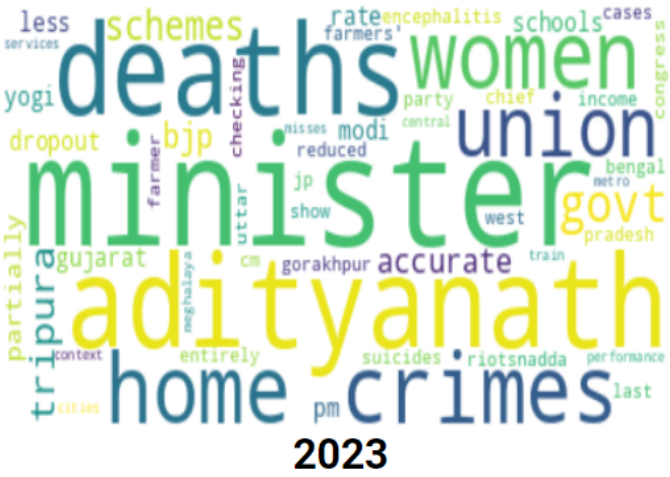
5.21 Using Wordclouds

By examining data over an extended period, we can identify trends and shifts in the nature of fake news. This analysis is vital for understanding how fake news topics evolve and how their propagation might change over time.









The word clouds from 2017 to 2023 provide an interesting glimpse into the topics that were likely prominent in discussions or media related to fake news in India. Here are some insights:

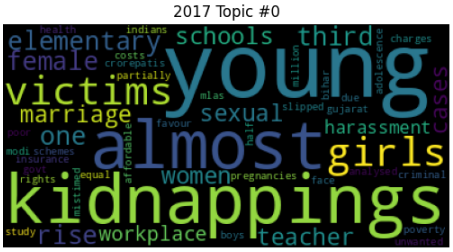
1. In **2017**, words like "women," "minister," "modi," "rural," and "bjp" were prominent, suggesting discussions were focused around political figures, gender issues, and possibly rural development. The presence of "bjp" indicates that much of the conversation may have revolved around the ruling political party.
2. The year **2018** emphasizes "minister," "modi," "congress," and "bjp," along with "toilet," pointing to continued political discourse, perhaps with a focus on sanitation initiatives, which may have been misrepresented in some news.
3. In **2019**, "minister," "modi," "attack," and "shah" are notable, possibly reflecting a political climate fraught with conflicts or allegations of violence that were being discussed or misreported.
4. The word "covid" appears in **2020** and becomes central in **2021**, showing the pandemic's overwhelming presence in fake news. Words like "vaccine," "cases," and "health" in these years indicate a focus on public health and the government's response to the crisis.
5. For **2022**, "health" remains a key term, along with "omricon" [presumably a misspelling of "Omicron"], suggesting that health misinformation continued to be a significant issue, likely in relation to COVID-19 variants.
6. In **2023**, "deaths," "crimes," "women," and "adityanath" are prominent, indicating a possible focus on law and order and gender issues. "Adityanath" might refer to the Chief Minister of Uttar Pradesh, suggesting regional political discourse was a significant source of fake news.

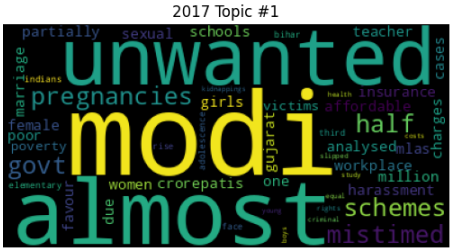
Across the years, the recurrence of political terms, especially names of prominent leaders and parties, suggests that a lot of fake news may be politically motivated or related to political events. The shift from more general political terms to specific issues like sanitation, health, and COVID-19 indicates that the nature of fake news is responsive to current events and societal concerns. The consistent presence of terms like "minister" and "modi" suggests that figures in high political offices are common subjects of misinformation. The introduction of "covid" related terms in 2020 and onwards underscores the pandemic's impact on the spread and focus of fake news.

5.22 Using Latent Dirichlet Allocation (LDA)

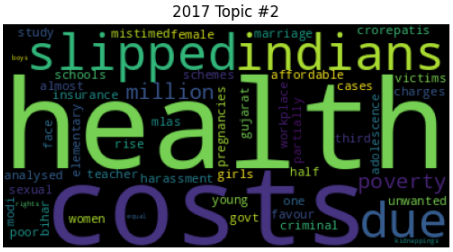
We performed LDA on titles of the articles for years 2017 to 2023 to find out the top 5 underlying topics present in the fake news articles. In every year we can mainly observe topics like politics, health, gender, crime. In year 2020 and 2021 we can observe one more topic which is related to coronavirus.

Topics 2017

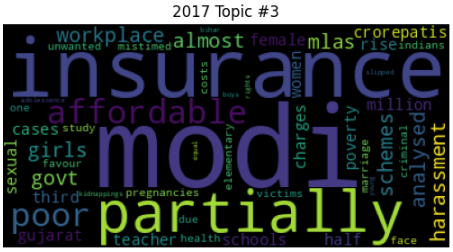




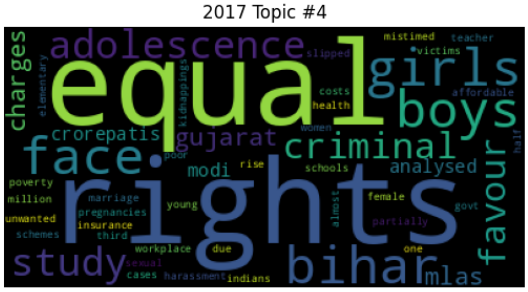
2017 Topic #2



2017 Topic #3

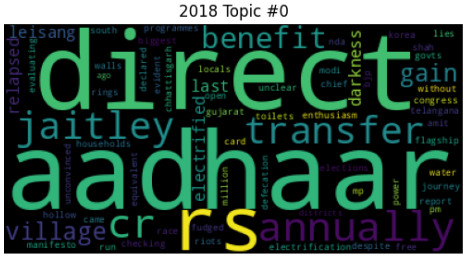


2017 Topic #4

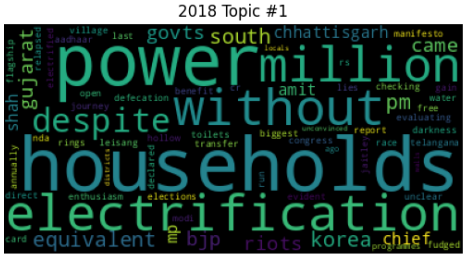


Topics 2018

2018 Topic #0

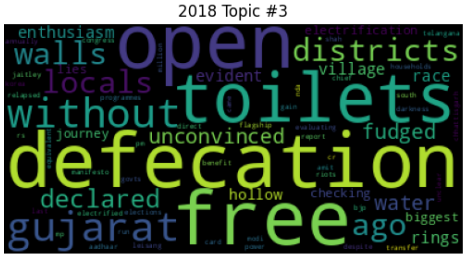


2018 Topic #1

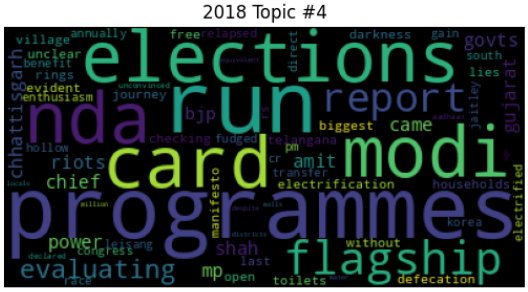


2018 Topic #2

2018 Topic #3

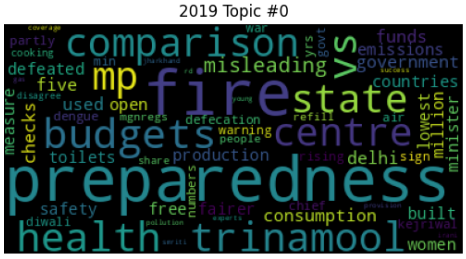


2018 Topic #4

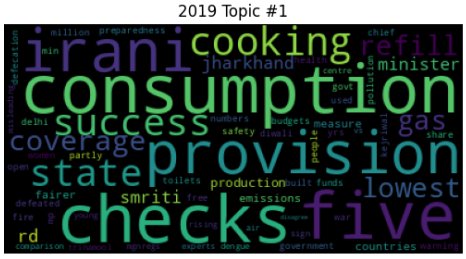


Topics 2019

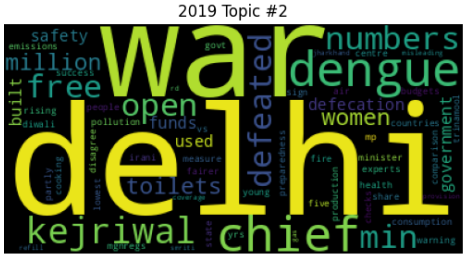
2019 Topic #0



2019 Topic #1

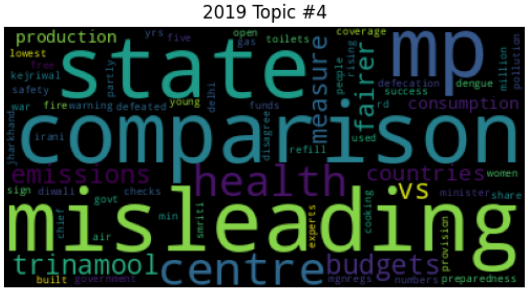


2019 Topic #2



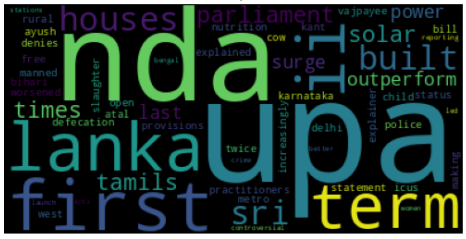
2019 Topic #3

2019 Topic #4

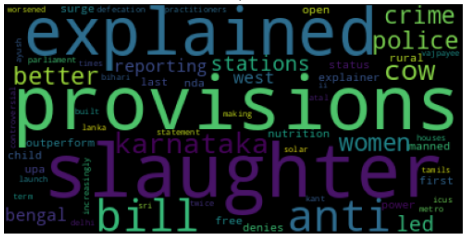


Topics 2020

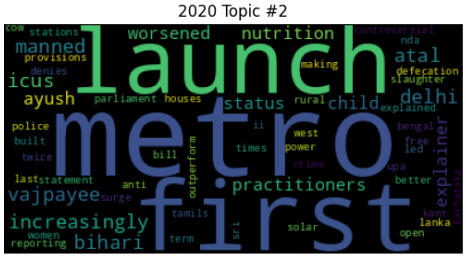
2020 Topic #0



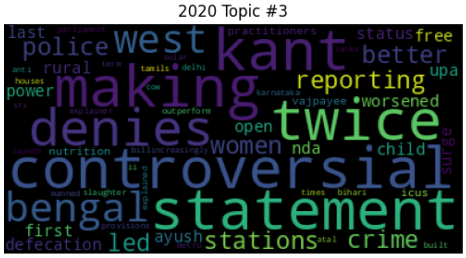
2020 Topic #1



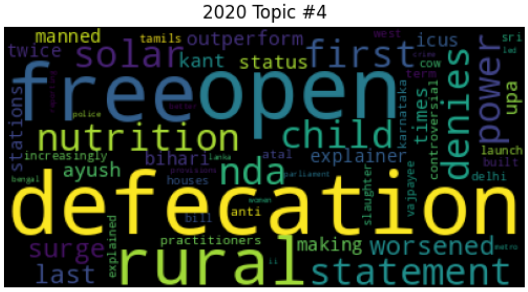
2020 Topic #2



2020 Topic #3



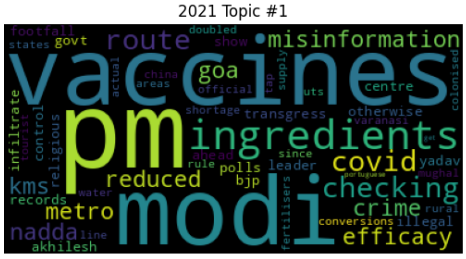
2020 Topic #4



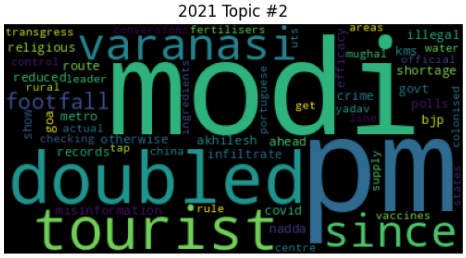
Topics 2021

2021 Topic #0

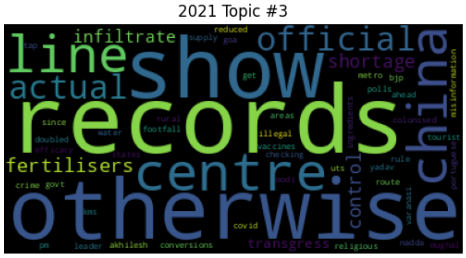
2021 Topic #1



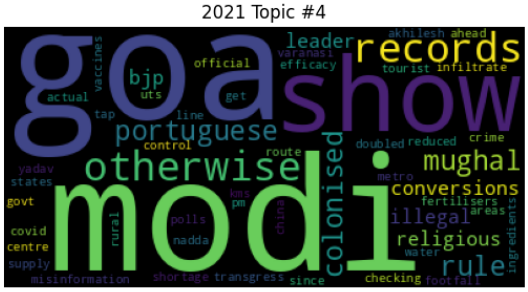
2021 Topic #2



2021 Topic #3

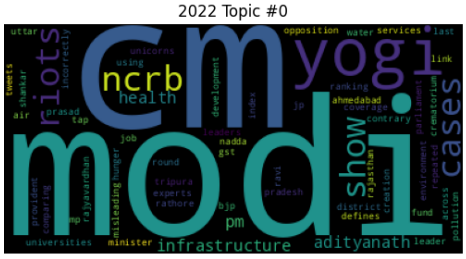


2021 Topic #4

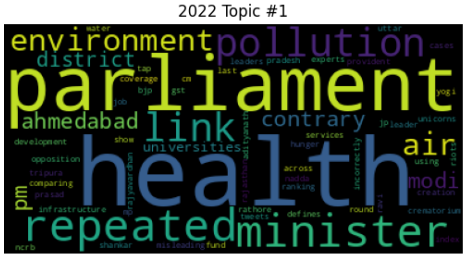


Topics 2022

2022 Topic #0

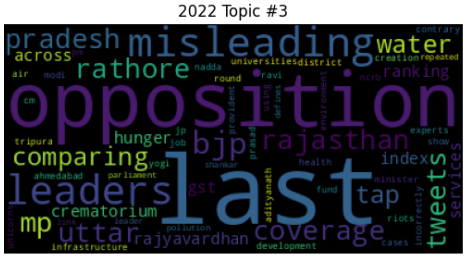


2022 Topic #1

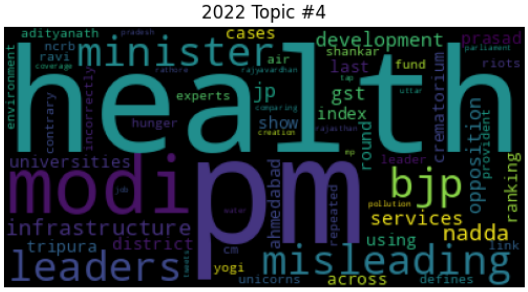


2022 Topic #2

2022 Topic #3

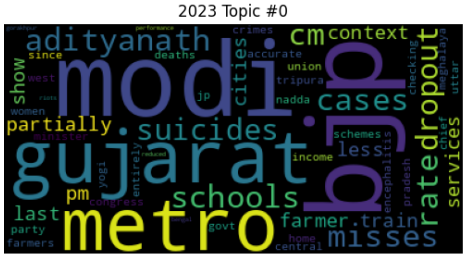


2022 Topic #4

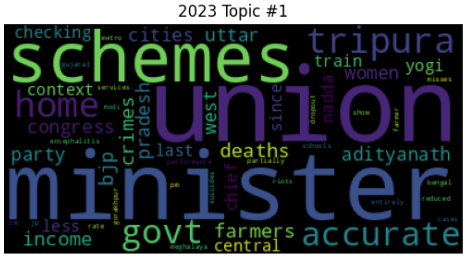


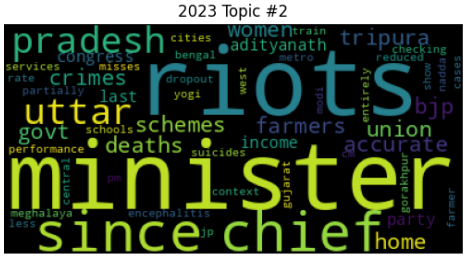
Topics 2023

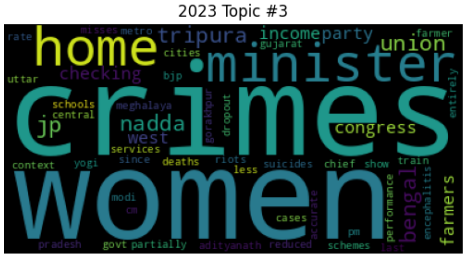
2023 Topic #0

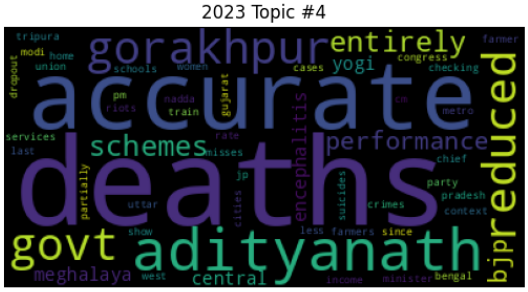


2023 Topic #1









5.3 Left & Right Wing Analysis:

Analyzing the political bias in news articles is crucial for understanding the landscape of misinformation. It helps in detecting any partisan slant in the news, which is a significant factor in the spread of fake news. This analysis is key for developing models that can neutrally and effectively classify news content without bias.

In our political analysis, we aimed to dissect the spread and political biases of fake news across different fact-checking websites. We meticulously curated a dataset, labeling entries as right-wing, left-wing, and neutral based on the frequency of popular keywords that emerged in fake news claims. Leveraging this dataset, we developed a text classification model using a Support Vector Machine (SVM) approach, which is adept at sorting complex text data into distinct categories.

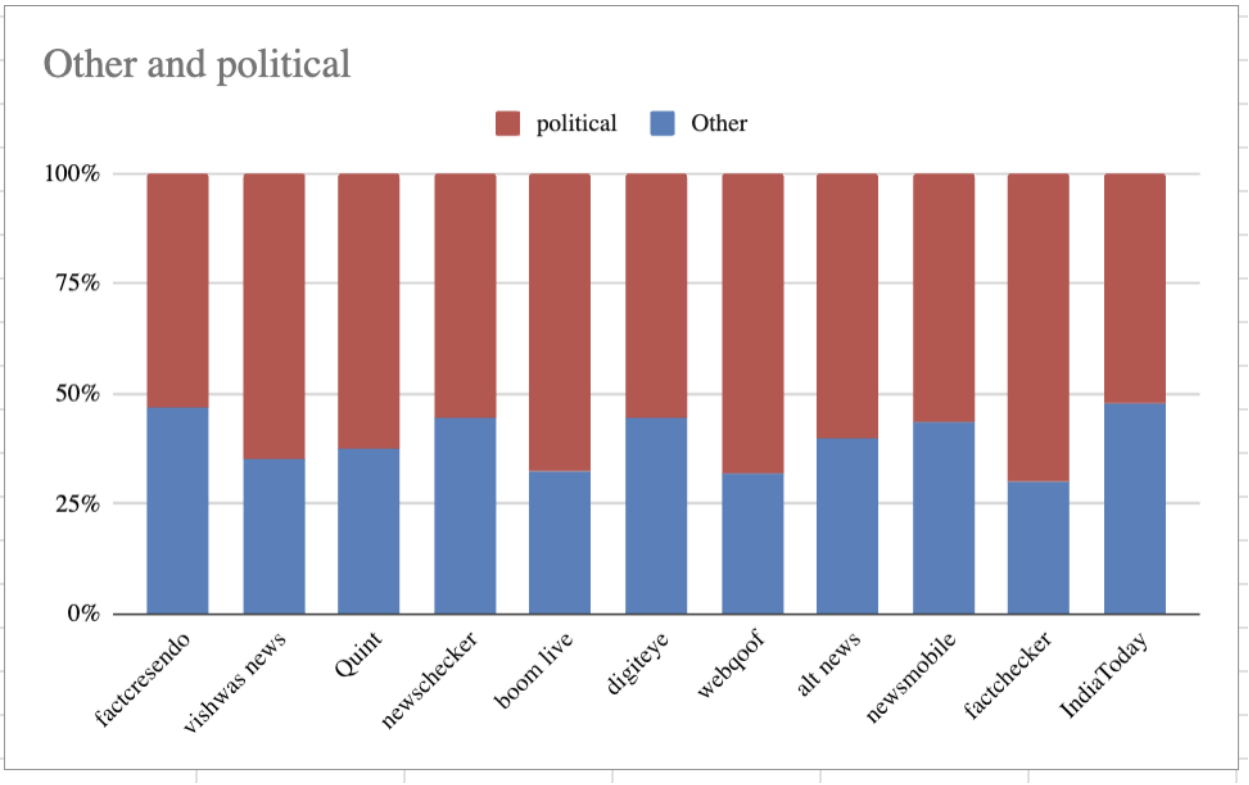


Figure 5

Our results as shown depict the prevalence of political versus non-political fake news, as well as the distribution between left-leaning and right-leaning fake news across various platforms. These visualizations indicate that political fake news constitutes a significant share of the total fact-checked articles, with a notable presence on all surveyed fact-checking sites. The distribution between left and right biases showed variability across these platforms, suggesting differing patterns in the dissemination or perhaps in the fact-checking focus of these sites.Our SVM model achieved an accuracy of 81%, a testament to the model's capabilities in classifying political orientations of fake news content. However, we recognize that there is room for improvement. We plan to expand our dataset to include more examples, which could help in enhancing the model's precision. Moreover, we intend to transition from the SVM model to a more sophisticated transformer-based model, which could significantly improve our understanding of the nuances in language used in fake news.

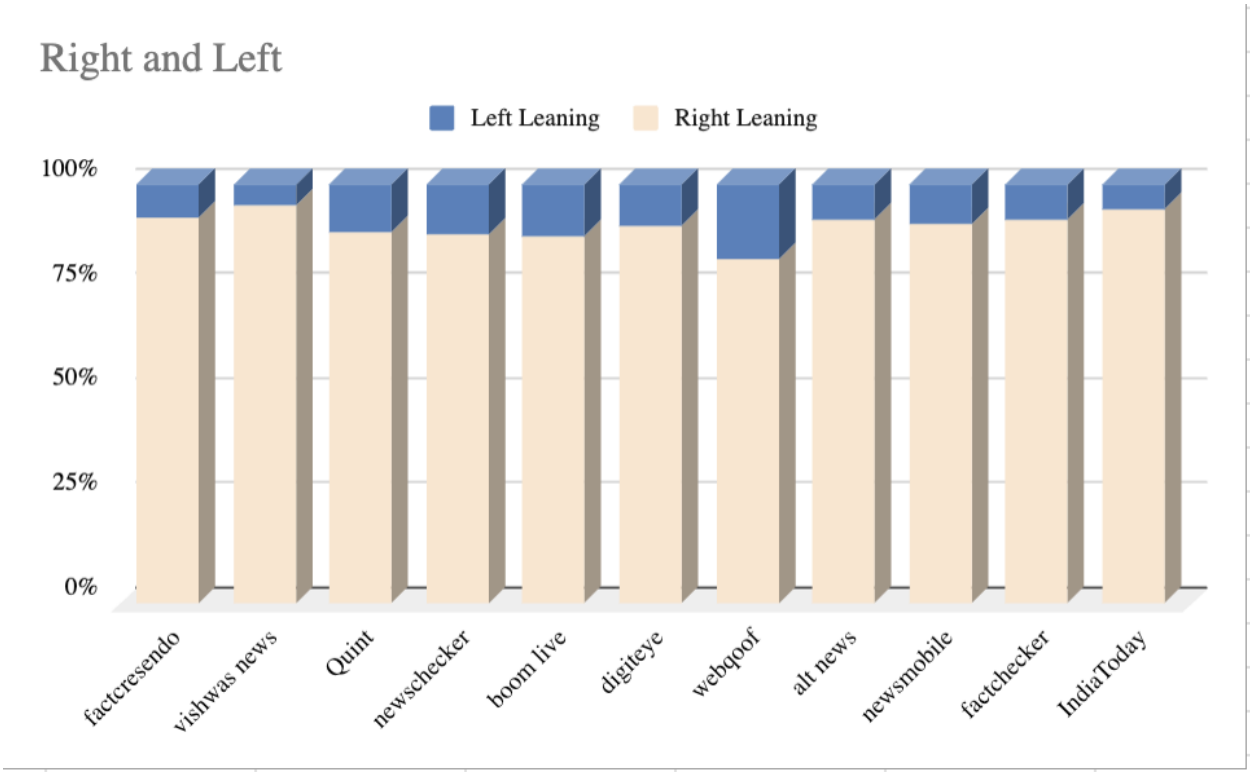


Figure 6

The insights from our analysis are revealing. The balance of left and right-leaning fake news varies by site, which could indicate biases in either the production of fake news or the fact-checking process itself. The significant proportion of political content highlights the critical nature of political discourse in the misinformation ecosystem. As we move forward, we are motivated to refine our approach, harnessing the potential of advanced AI and machine learning technologies to better detect, categorize, and understand the landscape of fake news, with the ultimate goal of bolstering the integrity of information dissemination.

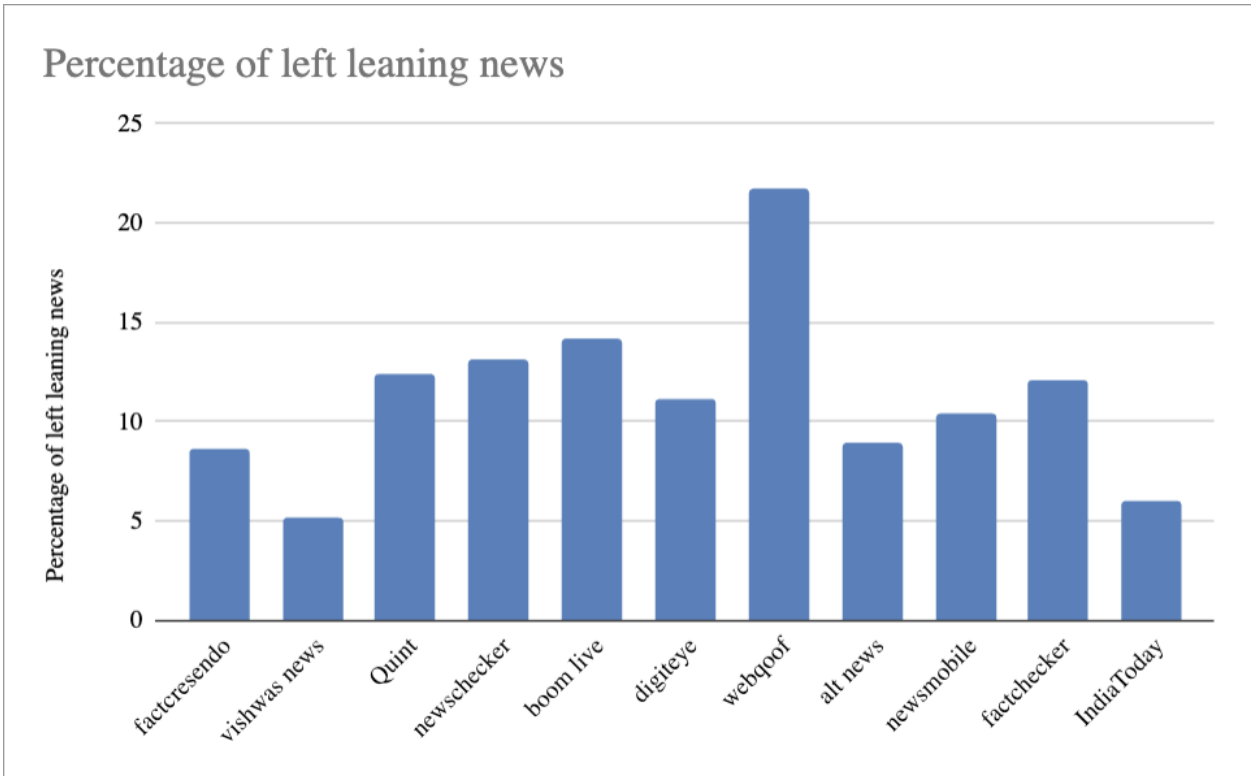


Figure 7

Conclusion

Throughout our dialogue, we've delved into the multifaceted nature of fake news and its implications in the political sphere. Our project's journey began with the meticulous extraction and categorization of data, which we then visualized to discern trends and patterns in the dissemination of fake news. These visualizations brought to light a notable shift in the subjects of fake news over time, reflecting the dynamic interplay between current events and the topics that tend to be misrepresented in the media.

Our analysis took a deep dive into the political polarity of fake news, examining how misinformation may be skewed towards left-wing or right-wing biases. Utilizing an SVM model, we achieved an 81% accuracy rate in classifying the political leanings of fake news content. This model served as a cornerstone for our exploration, providing a quantitative measure of the political charge embedded within the fake news narratives.

Our findings from the word clouds and bar charts highlighted the prevalence of political content, especially in relation to specific individuals and parties, which suggests a deliberate targeting within the political domain. Moreover, the emergence of terms related to health and pandemics in recent years underlines the adaptability of fake news to current global crises.

In conclusion, our report stands as a testament to the complex and ever-evolving nature of fake news. It underscores the necessity for ongoing vigilance and sophistication in our approach to detecting and combating misinformation. The insights garnered point towards the significant impact of fake news on public opinion and the pressing need for more advanced tools to enhance the accuracy of fact-checking mechanisms. As we look to the future, increasing our dataset size and transitioning to transformer-based models holds the promise of elevating our analytical capabilities, enabling us to contribute more effectively to the fight against the scourge of misinformation.

References

[1] Sharma, D. K., & Garg, S. (2021). IFND: a benchmark dataset for fake news detection. Complex & Intelligent Systems, 9(3), 2843–2863. https://doi.org/10.1007/s40747-021-00552-1

[2] Dhawan, A., Bhalla, M., Arora, D., Kaushal, R., & Kumaraguru, P. (2022). FakeNewsIndia: A benchmark dataset of fake news incidents in India, collection methodology and impact assessment in social media. *Computer Communications, 185*, 130-141. https://doi.org/10.1016/j.comcom.2022.01.003

[3] Singhal, S., Shah, R. R., & Kumaraguru, P. (2021). Factorization of Fact-Checks for Low Resource Indian Languages. *ArXiv*. /abs/2102.11276

[4] Badam, Jathin & Bonagiri, Akash & Raju, Kvln & Chakraborty, Dipanjan. (2022). Aletheia: A Fake News Detection System for Hindi. 255-259. 10.1145/3493700.3493736.

[5] Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., ... & Visnansky, G. (2018, May). Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 17-21).

[6] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *ArXiv*. /abs/1705.00648