data

MDPI

Data Descriptor

H-Prop and H-Prop-News: Computational Propaganda Datasets in Hindi

Deptii Chaudhari [1,*], Ambika Vishal Pawar [1,*] and Alberto Barrón-Cedeño [2]

[1] Department of CS & IT, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), Pune 412115, India

[2] Department of Interpreting and Translation, Alma Mater Studiorum-Università di Bologna, Corso della Repubblica 136, 47121 Forlì, Italy; a.barron@unibo.it

* Correspondence: deepti.chaudhari.phd2019@sitpune.edu.in (D.C.); ambikap@sitpune.edu.in (A.V.P.)

Abstract: In this digital era, people rely on the internet for their news consumption. As people are free to express their opinions on social media, much information shared on the internet is loaded with propaganda. Propagandist contents are intended to influence public opinion. In the mainstream media or prominent news agencies, the authors' and news agencies' own bias may impact in the news contents. Hence, it is required to detect such propaganda spread through news articles. Detection and classification of propagandist text require standard, high-quality, annotated datasets. A few datasets are available for propaganda classification. However, these datasets are mostly in English. Hindi is the most spoken language in India, and efforts are needed to detect its propagandist contents. This research work introduces two new datasets: H-Prop and H-Prop-News, which consist of news articles in Hindi annotated as propaganda or non-propaganda. The H-Prop dataset is generated by translating 28,630 news articles from the QProp dataset. The H-Prop-News dataset contains 5500 news articles collected from 32 prominent Hindi news websites. We experiment with the proposed datasets using four supervised machine learning models combined with different feature vectors and word embeddings. Our experiments achieve 87% accuracy using Logistic Regression with TF-IDF feature vectors. The datasets provide high-quality labeled news articles in Hindi and open new avenues for researchers to explore techniques for analyzing and classifying propaganda in Hindi text.

Dataset: https://zenodo.org/record/5828240#.YduLv2hBxPY.

Dataset License: Creative Commons Attribution 4.0 International

Keywords: propaganda identification; news articles analysis; Hindi text processing

check for updates

Citation: Chaudhari, D.; Pawar, A.V.; Barrón-Cedeño, A. H-Prop and H-Prop-News: Computational Propaganda Datasets in Hindi. *Data* **2022**, *7*, 29. https://doi.org/10.3390/data7030029

Received: 7 January 2022
Accepted: 15 February 2022
Published: 28 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Summary

According to [1], modern propaganda operates with many kinds of truth, such as half-truth, limited reality, and truth out of context. In recent times propaganda has been used by terrorist organizations for recruitment [2–5] and by political parties during elections [6–9], among many others. Today, abundant online news media has cropped up, some with the intent of spreading propaganda. The spectrum of a news article can range from neutral to biased [10]. Even though every news outlet/agency claims to be fair and unbiased, the personal stand of the article author and the news outlet may influence the reporting style and intent to some extent [11]. An author may use psychological and linguistic techniques to influence the readers about a specific topic. This malicious way of promoting agenda is generally referred to as propaganda.

Most of the work on creating automatic approaches to propaganda identification targets texts in English. However, most news articles are regional in a specific country context and political landscape. In India, internet users have seen a drastic surge in recent years. Hindi is the predominantly spoken language in India and the fourth most spoken language globally. Still, very little work is done to explore propaganda detection in regional languages, such as Hindi.

To allow the creation of models to identify the propaganda spread in Hindi, we introduce two datasets: H-Prop and H-Prop-News. H-Prop is produced by machine translation from an existing dataset containing news articles—propagandist vs non-propagandist in English [11,12]. A subset of the instances in QProp is translated in Hindi using IBM's Watson language translator [13]. H-Prop-News has been curated and annotated from scratch from a set of news articles originally written in Hindi, collected from prominent Indian news websites. The H-Prop corpus contains 28,630 news articles, whereas H-Prop-News contains 5500 news articles.

This research focuses on digital or computational propaganda, which will hugely contribute to the field of computational propaganda detection as no significant prior work is reported for propaganda detection in the Hindi language. Our contributions are as follows.

- We produce and release a new dataset of news articles in Hindi annotated for propaganda obtained from prominent news websites.
- We produce and release a derived dataset of news articles (originally in English) translated in Hindi and annotated for propaganda.
- We experiment with different machine learning models with the H-Prop-News dataset and show their effectiveness for propaganda classification.

Researchers can further utilize this dataset to train supervised models for the classification and detection of propaganda. These datasets can also be used for other research projects such as Hindi news articles classification and topic modeling.

Related Work

TSHP-17 [14] and Hyperpartisan News Dataset from SemEval-2019 [8] are two prominent datasets used to analyze news articles. Some studies [15–19] have worked toward rumor detection and fact-checking, whereas [12,20–22] have worked to uncover the political propaganda in news articles.

The TSHP-17 dataset [14] consists of news articles from 11 sources organized in four classes: trusted, satire, hoax, and propaganda. The dataset consists of 22,580 articles, out of which 5330 are flagged as propaganda. The authors created the dataset using distant supervision, considering the source of the news articles.

The authors of [23] released a corpus to identify fine-grained propaganda. The corpus contains 451 articles and was manually annotated by 6 people to identify 18 propaganda techniques at a fine-grained level. The authors identified 7485 propaganda technique instances from 21,230 sentences. The PTC-SemEval20 Corpus was presented by [24] as part of SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. This corpus consists of news articles gathered from 13 propaganda and 36 non-propaganda news websites identified by Media Bias/Fact Check. The corpus consists of 536 news articles with 8981 identified propaganda snippets. The annotation was performed manually by 6 professional annotators considering 18 propaganda techniques.

QCRI's propaganda corpus, known as QProp, comprises news articles focusing on two classes: Propaganda and non-propaganda [11]. The corpus was built by distant supervision, as the TSHP-17 dataset, considering the news source information published by Media Bias/Fact Check (https://mediabiasfactcheck.com, accessed on 6 January 2022) (MBFC). QProp 51,246 articles: 5714 from propagandist sources and 45,532 from non-propagandist ones. Table 1 shows statistics of the QProp dataset.

Table 1. Details of QProp Dataset.

Data	Propagandist	Non-Propagandist	Total
Development Data	575	4560	5135
Training Data	4004	31,953	35,957
Testing Data	1135	9019	10,154
Total	5714	45,532	51,246

In the TSHP-17 and QProp datasets, the articles were labeled using the distant supervision technique, which relies entirely on the source of an article to label it as propaganda. This approach did not consider the actual contents of the article to identify the propaganda in the text. Also, the number of propagandist instances in both datasets is meager compared to other classes.

All the prominent datasets proposed for propaganda detection are in the English language. To the best of our knowledge, no such dataset is available for Hindi. In this work, two datasets in Hindi for propaganda by using two different approaches are generated. In the first case, QProp is translated into Hindi. In the second case, an original dataset of Hindi news is created. As the articles are labeled by looking at the actual contents of the news articles, the annotation is more reliable.

2. Data Description

This section provides a detailed description of the H-Prop and H-Prop-News datasets. Tables 2 and 3 show statistics of the two datasets.

Table 2. Specification of Datasets.

Information	Dataset	
	H-Prop	H-Prop-News
Subject Area	NLP	NLP
Focus Area	Propaganda Classification	Propaganda Classification
File Type	tsv	csv
No. of files	3	3
Method of Dataset generation	Translation of English dataset QProp	Web Scraping from Hindi News Websites

Table 3. Dataset Details.

Dataset	Instances	Source of Data	Attributes Used
H-Prop	28,630	English News Websites	News article, Propaganda Label
H-Prop-News	5500	Hindi News Websites	News Website, News Article URL, News Headline, News Article, Propaganda Label

2.1. H-Prop Dataset

The original QProp dataset consists of 51,246 news articles. The H-Prop dataset is derived from QProp and considered only 28,630 articles. The data is split into development, training, and testing partitions. The dataset files are in tab-separated format, and UTF-8 encoding is used. This subsample of the corpus is translated into Hindi using IBM Watson Language Translator [13] (https://www.ibm.com/cloud/watson-language-translator, accessed on 13 October 2021). The translation was done over the months in 2021. Table 4 shows the details of H-Prop dataset as per partitions.

Table 4. H-Prop dataset as per partitions.

Partition	Propagandist	Non-Propagandist	Total
Development Data	574	3000	3574
Training Data	3471	15,459	18,930
Testing Data	1080	5046	6126
Total	5125	23,505	28,630

2.2. H-Prop-News Dataset

The H-Prop-News Dataset is built by extracting the news articles from 32 prominent mainstream news portals in India. The articles are fetched from September 2021 to December 2021. Its focus is national and political news in the Indian context. Table 5 shows statistics about the H-Prop-News dataset. A total of 5500 articles were scraped from these websites using the parseHub web scraping tool (https://www.parsehub.com/, accessed on 6 January 2022). These articles are annotated as propaganda or non-propaganda considering their contents and identifying the propaganda techniques observed in them.

Table 5. Statistics of the H-Prop-News dataset.

Partition	Propagandist	Non-Propagandist	Total
Development Data	275	275	550
Training Data	1850	2000	3850
Testing Data	505	595	1100
Total	2630	2870	5500

Table 6 shows the class-wise article distribution per medium. Most propagandist articles come from the news website Patrika News (available online: www.patrika.com, accessed on 6 January 2022), whereas most non-propaganda articles come from Amar Ujala (available online: www.amarujala.com, accessed on 6 January 2022).

Table 6. Class-wise article distribution.

News Source	Non-Propaganda Articles	Propaganda Articles	News Source	Non-Propaganda Articles	Propaganda Articles
Amar Ujala	360	244	Dainik Navjyoti	68	56
Bhaskar	293	204	Pratahkal	63	85
Dainik Jagran	269	238	Jansatta	61	93
One India	166	120	Asianet News	51	21
Zee News	150	85	BBC Hindi	36	21
Patrika	133	347	Nai Duniya	32	15
News On AIR	131	20	Oulook Hindi	25	54
India	130	78	Univarta	16	3
Punjab Kesari	121	78	The Quint	15	28
Dainik Tribune	114	79	Bebak Post	12	16
Khas Khabar	111	63	The Wire	12	34
TV9 Hindi	111	99	Live Hindustan	11	8
Samay Live	107	111	Aaj Tak	10	168
Lokmat News	95	52	DW	9	1
Tehelka Hindi	77	156	Times Now	5	8
ABP Live	76	40	India TV	0	5

3. Methods

This section elaborates on the methods and techniques used for data collection and dataset generation of H-Prop and H-Prop-News datasets.

3.1. H-Prop Dataset Generation

A portion of the QProp dataset for preparing the H-Prop dataset is considered, as explained in Section 2.1. IBM Watson Language Translator is used for translation purposes. The English translation process introduces several special characters due to the encoding conversion. These special characters are then removed to clean the data. Figure 1 shows the methodology used to generate the H-Prop dataset.

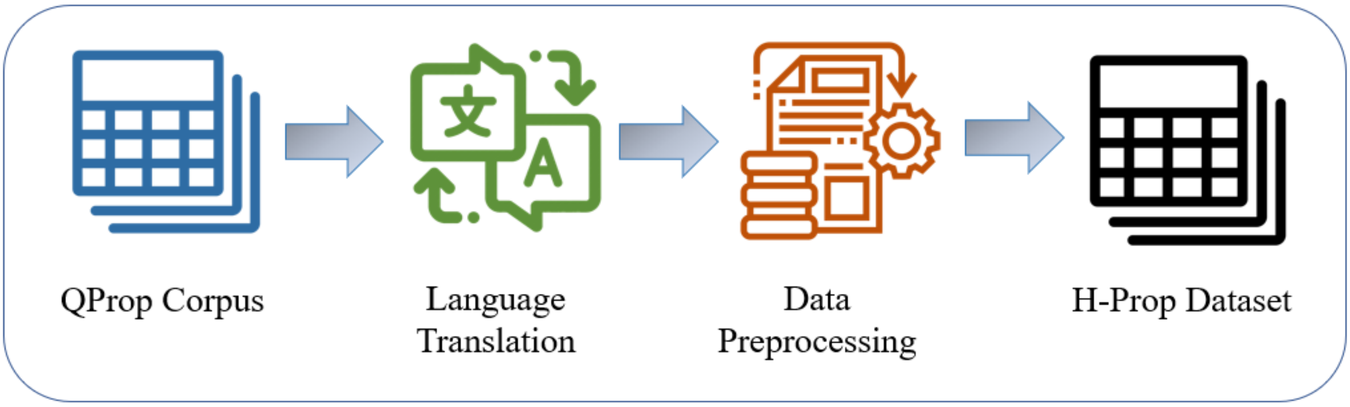

Figure 1. Generation of H-Prop Dataset.

3.2. H-Prop-News Dataset Generation

First, 32 prominent Hindi news websites were selected, reporting national and political news. Collecting data from different websites is a challenging task. Each website follows a different page layout. Parsehub is a cloud-based, free web-scraping tool that extracts data from a website in a few steps. We extracted News headlines, News URLs, and Article Texts from the websites. Figure 2 shows the process of H-Prop-News dataset creation.

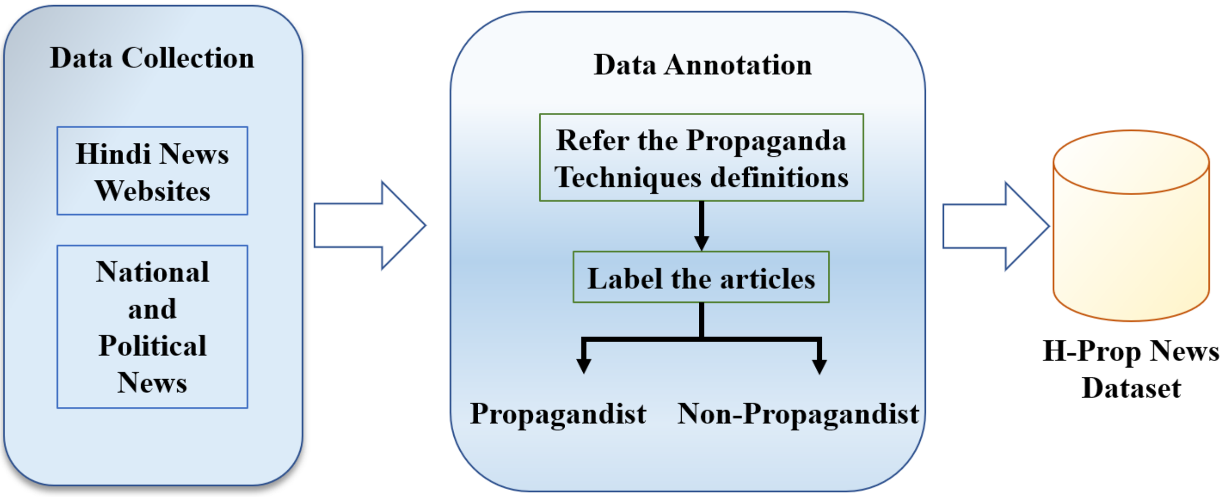

Figure 2. Creation of H-Prop-News Dataset.

3.3. Data Annotation

The news articles in the QProp corpus were labeled using distant supervision. The authors [11] rely on the news outlet information provided by Media Bias Fact Check (MBFC) (https://mediabiasfactcheck.com/, accessed on 6 January 2022). The labels were obtained by considering news coming from propagandist news outlets as propagandist and news coming from non-propaganda news outlets. We retain the annotations as provided in the original QProp dataset.

The annotation task for the H-Prop-News dataset involved identifying propaganda methods used and labeling the articles as propaganda or non-propaganda. The definitions of 14 propaganda techniques are followed as listed in Table 7. The annotation task was done in two phases (i) two annotators labeled the articles independently as propaganda or non-propaganda class, and (ii) the annotations were then reviewed for conflicts. We used the LightTag text annotation tool [25] for the annotation and analysis. With reference to the annotation guidelines provided by the authors of [24], we present the flowchart for the article label decision process at the document level. As shown in Figure 3, the propaganda techniques are grouped as per specific indications. For example, the articles showing the addition of irrelevant data along with problem simplification may have propaganda techniques such as casual oversimplification, appeal to authority, black-and-white fallacy, or thought-terminating cliché. The annotators further referred to the more detailed definition of these techniques as listed in Table 7 for technique identification. If more than one technique is spotted in the article, the annotator labeled the article as propaganda.

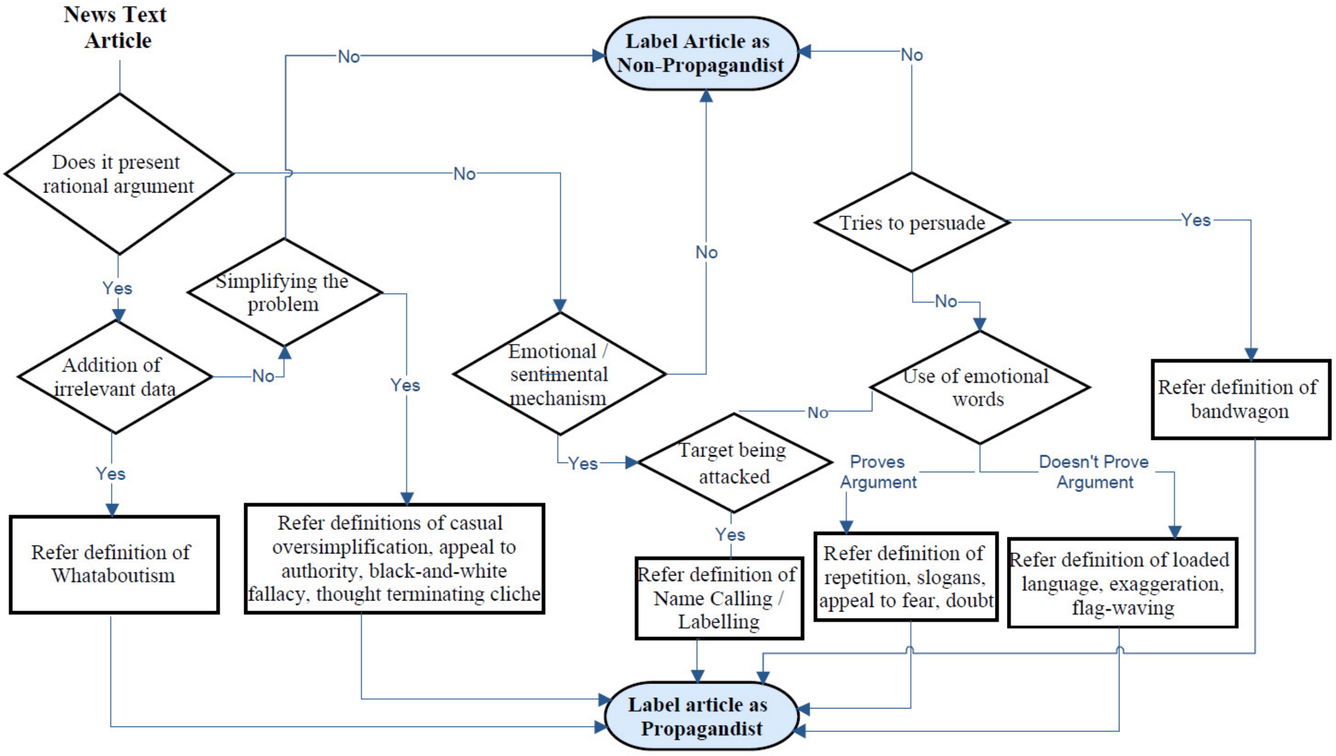
News Text Article
Does it present rational argument
Yes
No
Addition of irrelevant data
No
Yes
Simplifying the problem
No
Yes
Label Article as Non-Propagandist
Emotional / sentimental mechanism
No
Yes
Tries to persuade
No
Yes
No
Use of emotional words
No
Proves Argument
Doesn't Prove Argument
Target being attacked
Yes
Refer definition of Whataboutism
Refer definitions of casual oversimplification, appeal to authority, black-and-white fallacy, thought terminating cliche
Refer definition of Name Calling / Labelling
Refer definition of repetition, slogans, appeal to fear, doubt
Refer definition of loaded language, exaggeration, flag-waving
Refer definition of bandwagon
Label article as Propagandist

Figure 3. The process followed for the Article Annotation decision.

To evaluate annotation quality in terms of inter-annotator agreement, Cohen's Kappa [26] is used. Cohen's Kappa measures the agreement between two annotators, classifying articles in n mutually exclusive categories. The inter-annotator agreement (K) observed is on average 0.81.

Table 7. Propaganda Technique definitions.

No.	Propaganda Technique	Definition
1.	Loaded Language	Use of strong emotional words and phrases [27]
2.	Name Calling/Labelling	Labeling the object of the propaganda with something the audience fears, hates, finds undesirable, or loves or praises [28]
3.	Repetition	Repeating the same message repeatedly [28,29]
4.	Exaggeration/minimization	Representing something excessively or making something seem less important or smaller than it actually is [30]
5.	Doubt	Questioning the credibility of someone or something
6.	Appeal to fear/prejudice	Infusing anxiety and/or panic towards an alternative, possibly based on prejudiced conclusions
7.	Flag-waving	Playing on strong national feeling to justify or promote an action or idea [31]
8.	Causal oversimplification	Transfer of the blame to one person or group of people without investigating the complexities of an issue
9.	Slogans	A concise and dramatic phrase that may include labeling and stereotyping
10.	Appeal to authority	Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence [32]
11.	Black-and-white fallacy	Presenting two alternative options as the only possibilities, when in fact, more possibilities exist [29]
12.	Thought-terminating cliche	Words or phrases that discourage critical thought and meaningful discussion on a topic [33]
13.	Whataboutism	Discredit an opponent's position by charging them with hypocrisy without directly disproving their argument [34]
14.	Bandwagon	Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action" [31]

Sample news articles and the respective labels are shown in Table 8. The English translation is provided here for the understanding of our international readers. The first article does not contain any propaganda technique. In the second news article, propaganda techniques such as loaded language, exaggeration, and casual oversimplification can be observed.

Table 8. Sample news articles text and labels.

Sample News Article Text	English Translation of Article Text	Article Label
रिपोर्ट में कह□गय□है कि शहर गैस वितरण क□पनियो□को कीमतो□में 10–11 प्रतिशत की बढ़ोतरी करनी होगी।'' अ□तरर□ष्ट्रीय ब□ज□रो□क□रुख क□अनुरूप अप्रैल, 2022 स□सित□बर, 2022 क□दौर□न एपीएम गैस क□द□म बढ़कर 5.93 डॉलर प्रति इक□ई हो ज□एग□ अक्टूबर, 2022 स□म□र्च, 2023 तक यह 7.65 डॉलर प्रति इक□ई होग□ इसक□मतलब है कि अप्रैल, 2022 में सीएनजी और पीएनजी की कीमतो□में 22–23 प्रतिशत की वृद्धि होगी। अक्टूबर, 2022 में द□म 11 स□12 प्रतिशत और बढ़ेंग□ रिपोर्ट में कह□गय□है कि एपीएम गैस मूल्य में बढ़ोतरी की वजह स□अक्टूबर, 2021 स□अक्टूबर, 2022 क□दौर□न एमजीएल और आईजीएल को कीमतो□में 49 स□53 प्रतिशत की बढ़ोतरी करनी होगी।	The report said that the city gas distribution companies will have to increase the prices by 10–11 percent. From October 2022 to March 2023, it will be $7.65 per unit. This means that the prices of CNG and PNG will increase by 22–23 percent in April 2022. In October 2022, the price will increase by another 11 to 12 percent. MGL and IGL will have to increase prices by 49 to 53 percent between October 2021 and October 2022 due to the hike in APM gas prices, the report said.	Non-Propaganda
Privatization को ल□कर केंद्र पर भड़क□Rahul Gandhi, बोल□द□श की 70 स□ल की पू□जी को PM न□ब□च दिय□ By लोकमत न्यूज़ ड□स्क | Published: 24 August 2021 09:59 PM क□ग्र□स न□त□र□हुल ग□ंधी न□मोदी सरक□र पर हमल□बोल□है. र□हुल ग□ंधी न□र□ष्ट्रीय मौद्रिकरण योजन□(एनएमपी) की घोषण□को युव□ओ□क□'भविष्य पर हमल□कर□र द□त□हुए सरक□र पर निश□न□स□ध□ उन्हो□न□आरोप लग□य□कि प्रध□नमन्त्री नरेंद्र मोदी न□70 स□ल में बनी द□श की पू□जी को अपन□कुछ उद्योगपति मित्रो□क□ह□थो□में ब□च दिय□ र□हुल ग□ंधी न□य□द□व□भी किय□कि कुछ क□पनियो□को यह 'उपह□र द□न□स□उनक□एक□धिक□र बन□ग□जिस क□रण द□श क□युव□ओ□को रोजग□र नही□मिल प□एग□	Rahul Gandhi, furious at the Center over privatization, said-PM sold the country's 70 years of capital! By Lokmat News Desk | Published: 24 August 2021, 09:59 p.m. Congress leader Rahul Gandhi has attacked the Modi government. Rahul Gandhi targeted the government, terming the announcement of the National Monetization Plan (NMP) as an "attack on the future of the youth". He alleged that Prime Minister Narendra Modi sold the country's capital built in 70 years to some of his industrialist friends. Rahul Gandhi also claimed that giving this "gift" to some companies will make them a monopoly, due to which the youth of the country will not be able to get employment.	Propaganda

4. Experimental Setup

This section provides an overview of the experiments performed for the propaganda classification task using the H-Prop-News dataset. We trained four machine learning models: Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost. Figure 4 shows the propaganda classification framework. After preprocessing data by removing URLs, we remove the Hindi stopwords from the article text. The tokenization of the text is performed using the nlp-indic library. For representation, we use four different feature vectors and word embeddings: Bag-of-words, TFIDF (Term Frequency-Inverse Document Frequency), word2vec, and doc2vec. Each machine learning model is fed with each of the word embeddings. The entire dataset of 5500 articles is considered for the experimental setup. The dataset is split into training, testing and validation set using an 70:20:10 ratio. The resulting training set contains 3850 articles, testing set contains 1100 articles and validation set contains 550 articles.

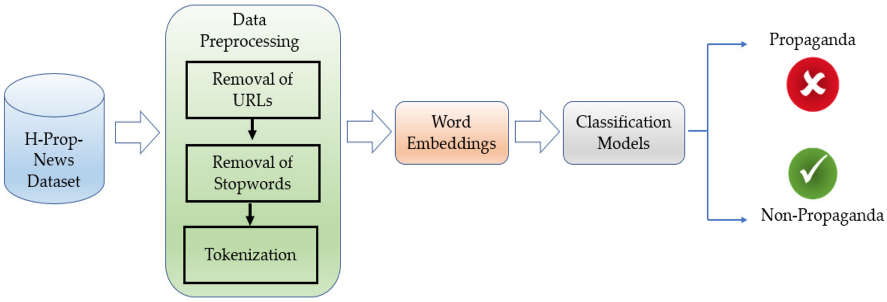

Figure 4. Propaganda classification methodology.

5. Results and Discussion

Table 9 shows the performance of all the machine learning models using different features and word embeddings on training, testing and validation sets. The Logistic Regression with TF-IDF feature vectors gives the best results on the testing as well as validation dataset. The F1 score and accuracy obtained on validation set is 87.46 and 87.45 respectively. All the classifiers show least performance with doc2vec word embeddings.

Table 9. Classifier performance.

Feature Vectors/Word Embeddings	**Classifier**	**Performance on Training Set**		**Performance on Testing Set**		**Performance on Validation Set**	
		F1 Score	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**
Bag-of-Words	Logistic Regression	99.93	99.93	86.37	86.36	85.22	85.27
	Random Forest	99.98	99.98	82.70	82.73	86.82	86.91
	SVM	95.50	95.50	83.76	83.82	86.08	86.18
	XGBoost	87.08	87.11	83.58	83.64	83.33	83.45
TF-IDF	Logistic Regression	95.45	95.45	87.06	87.09	87.46	87.45
	Random Forest	99.93	99.93	86.12	86.18	84.20	84.17
	SVM	98.02	98.02	86.84	86.91	85.61	85.64
	XGBoost	88.95	88.98	83.90	84.00	83.23	83.27
word2vec	Logistic Regression	74.53	74.52	73.48	73.45	75.10	75.09
	Random Forest	99.98	99.98	72.54	72.55	71.54	71.82
	SVM	72.57	72.57	72.01	72.00	73.31	73.27
	XGBoost	87.95	86.75	83.93	84.29	83.76	83.86
doc2vec	Logistic Regression	44.45	49.70	43.78	49.82	44.29	48.91
	Random Forest	51.09	51.09	50.64	50.55	51.01	50.91
	SVM	48.38	51.25	47.62	51.09	51.10	53.27
	XGBoost	50.23	52.50	49.36	51.45	51.14	52.55

The main aim of this work was to develop a propaganda dataset in the Hindi language and a machine learning model for the classification of propaganda text. The annotation process required rigorous and time-consuming inspections of the news articles by the annotators. The annotation reliability is established by using Cohen Kappa as the measure. The most frequent propaganda techniques observed during the annotation process were loaded language and labeling or name-calling. Our observations are similar to the findings of the work [23].

Propaganda detection remains a challenging task with fine-grained analysis of the text. This work provided an opportunity to develop machine learning models that detect propaganda at the document level.

Use Cases of the H-Prop and H-Prop News Dataset

The proposed datasets have the following practical implications.

- These datasets can be used for propaganda classification tasks at the article level.
- The datasets can be further enriched for fine-grained propaganda labeling to identify various propaganda techniques.
- The H-Prop-News dataset can be further utilized to explore various topics and events related to propaganda, such as the target of propaganda, source of propaganda, etc.

6. Conclusions and Future Work

The research presents two propaganda datasets. H-Prop consists of news articles translated from the English propaganda dataset QProp. H-Prop-News contains original Hindi News articles gathered from Hindi mainstream news websites. The H-Prop dataset contains 28,630 news articles, and the H-Prop-News dataset contains 5500 news articles. The annotations of articles are retained from the original QProp corpus. In contrast, the H-Prop-News dataset is manually annotated, considering the definitions of propaganda techniques. To the best of our knowledge, no significant work is reported in the area of propaganda detection in Hindi text. Hence, these newly created datasets are the first publicly available datasets of their kind. This work also explains the process for dataset creation and provides statistical details. Also, the propaganda classification using machine learning techniques is explored, obtaining an accuracy of 87%. Thus, this work is a contribution in this direction. As computational propaganda detection and analysis is an upcoming field of research, this work will help researchers explore natural language processing and machine learning techniques in this area.

As the future scope of this work, the aim is to augment the size of the H-Prop-News dataset by covering more news websites. Currently, the news articles are collected under the national and political categories. The dataset can also be included to evaluate the use of propaganda in opinion and editorial articles. As the dataset is manually annotated, it might have the annotators' bias. More annotators can be employed to dim it. It is also observed that even though the news articles are collected from Hindi news media, the text is not purely in Hindi. Some amount of code-mixing or use of English words is observed.

Author Contributions: Conceptualization, A.V.P. and A.B.-C.; methodology, D.C.; resources, A.V.P. and D.C.; data curation, D.C.; writing—original draft preparation, D.C.; writing—review and editing, A.V.P. and A.B.-C.; funding acquisition, A.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Research Support Fund of Symbiosis International (Deemed University).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available on zenodo platform with doi:10.5281/zenodo.5828240. [HProp] Deptii Chaudhari, Ambika Pawar, and Alberto Barrón-Cedeño. 2021. H-Prop and H-Prop-News: Computational Propaganda Datasets in Hindi.

Acknowledgments: We are thankful to Symbiosis International (Deemed University) for providing research facilities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ellul, J.; Merton, R.K.; Kellen, K.; Lerner, J. *Propaganda: The Formation of Men's Attitudes*; Vintage Books: New York, NY, USA, 1965.
2. Caluza, L.J.B. Deciphering published articles on cyberterrorism: A latent Dirichlet allocation algorithm application. *Int. J. Data Min. Model. Manag.* **2019**, *11*, 87–101. [CrossRef]
3. Alharbi, A.R.; Aljaedi, A. Predicting rogue content and arabic spammers on twitter. *Futur. Internet* **2019**, *11*, 229. [CrossRef]
4. Heidarysafa, M.; Kowsari, K.; Odukoya, T.; Potter, P.; Barnes, L.E.; Brown, D.E. Women in ISIS Propaganda: A Natural Language Processing Analysis of Topics and Emotions in a Comparison with Mainstream Religious Group. In *Science and Information Conference*; Springer: Cham, Germany, 2019; pp. 610–624.
5. Nizzoli, L.; Avvenuti, M.; Cresci, S.; Tesconi, M. Extremist propaganda tweet classification with deep learning in realistic scenarios. In Proceedings of the 11th ACM Conference on Web Science, Boston, MA, USA, 30 June–3 July 2019; pp. 203–204. [CrossRef]
6. Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; Menczer, F. Truthy: Mapping the spread of astroturf in microblog streams. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–April 2011; pp. 249–252. [CrossRef]
7. Kellner, A.; Rangosch, L.; Wressnegger, C.; Rieck, K. Political Elections Under (Social) Fire? Analysis and Detection of Propaganda on Twitter. *arXiv* **2019**, arXiv:1912.04143.
8. Stukal, D.; Sanovich, S.; Tucker, J.A.; Bonneau, R. For Whom the Bot Tolls: A Neural Networks Approach to Measuring Political Orientation of Twitter Bots in Russia. *Sage Open* **2019**, *9*, 1–16. [CrossRef]
9. Neyazi, T.A. Digital propaganda, political bots and polarized politics in India. *Asian J. Commun.* **2020**, *30*, 39–57. [CrossRef]
10. Chaudhari, D.D.; Pawar, A.V. Propaganda analysis in social media: A bibliometric review. *Inf. Discov. Deliv.* **2021**, *49*, 57–70. [CrossRef]
11. Barrón-Cedeño, A.; Jaradat, I.; Da San Martino, G.; Nakov, P. Proppy: Organizing the news based on their propagandistic content. *Inf. Process. Manag.* **2019**, *56*, 1849–1864. [CrossRef]
12. Barrón-Cedeño, A.; Da San Martino, G.; Jaradat, I.; Nakov, P. Proppy: A System to Unmask Propaganda in Online News. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 9847–9848. [CrossRef]
13. Watson Language Translator-India | IBM. Available online: https://www.ibm.com/in-en/cloud/watson-language-translator (accessed on 13 October 2021).
14. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2931–2937. [CrossRef]
15. Popat, K.; Mukherjee, S.; Strötgen, J.; Weikum, G. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 1003–1012. [CrossRef]
16. Wang, L.; Wang, Y.; De Melo, G.; Weikum, G. Five shades of untruth: Finer-grained classification of fake news. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 593–594. [CrossRef]
17. Qazvinian, V.; Rosengren, E.; Radev, D.R.; Mei, Q. Rumor has it Identifying Misinformation in Microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 1589–1599.
18. Baly, R.; Karadzhov, G.; Alexandrov, D.; Glass, J.; Nakov, P. Predicting factuality of reporting and bias of news media sources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3528–3539. [CrossRef]
19. Kwon, S.; Cha, M.; Jung, K.; Chen, W.; Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of the 2013 IEEE 13th International Conference on Data mining, Dallas, TX, USA, 7–10 December 2013; pp. 1103–1108. [CrossRef]
20. Saleh, A.; Baly, R.; Barrón-Cedeño, A.; Da San Martino, G.; Mohtarami, M.; Nakov, P.; Glass, J. Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 1041–1046. [CrossRef]
21. Yoosuf, S.; Yang, Y. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Hong Kong, China, 19 August 2019; pp. 87–91. [CrossRef]
22. Baisa, V.; Herman, O.; Horák, A. Benchmark dataset for propaganda detection in Czech newspaper texts. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 77–83. [CrossRef]
23. da San Martino, G.; Yu, S.; Barrón-Cedeño, A.; Petrov, R.; Nakov, P. Fine-grained analysis of propaganda in news articles. 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), Hong Kong, China, 3–7 November 2019; pp. 5636–5646.

24. Da San Martino, G.; Barrón-Cedeño, A.; Wachsmuth, H.; Petrov, P. SemEval 2020 Task 11: Detection of Propaganda Techniques in News Articles. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 12–13 December 2020; pp. 1377–1414.
25. Perry, T. LightTag: Text Annotation Platform. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing System demonstration, Santo Domingo, Dominican Republic, 7–11 November 2021; pp. 20–27.
26. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
27. Weston, A. *A Rulebook for Arguments*, 5th ed.; Hackett Publishing: Indianapolis, Indiana, 2018; pp. 1–86.
28. Miller, C.R. The Techniques of Propaganda. *How Detect. Anal. Propag.* **1939**, *10*, 27–29. Available online: https://www.cengage.com/resource_uploads/downloads/0534619029_19636.pdf (accessed on 6 January 2022).
29. Torok, R. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In Proceedings of the 8th Australian Security and Intelligence Conference, Joondalup, Australia, 30 November–2 December 2015; pp. 58–65. [CrossRef]
30. Jowett, G.S.; O'Donnell, V. What Is Propaganda, and How Does It Differ From Persuasion? In *Propaganda and Persuasion*, 4th ed.; Sage Publications: Thousand Oaks, CA, USA, 2006.
31. Hobbs, R. Teaching about Propaganda: An Examination of the Historical Roots of Media Literacy. *J. Media Lit. Educ.* **2014**, *6*, 56–67. [CrossRef]
32. Goodwin, J. Accounting for the force of the appeal to authority. *Argumentation* **2011**, *25*, 1–9. [CrossRef]
33. Hunter, J. Brainwashing in a Large Group Awareness Training?: The Classical Conditioning Hypothesis of Brainwashing. Ph.D. Thesis, Psychology University of Kwazu-Natal, Durban, South Africa, September 2015.
34. Richter, M.L. The Kremlin's Platform for 'Useful Idiots' in the West: An Overview of RT's Editorial Strategy and Evidence of Impact. *Eur. Values* **2017**, *31*, 53. Available online: http://www.europeanvalues.net/wp-content/uploads/2017/09/Overview-of-RTs-Editorial-Strategy-and-Evidence-of-Impact.pdf (accessed on 6 January 2022).