Weakly Supervised Learning for Textual Propaganda Detection on Twitter and a Multi-label Tweet Propaganda Dataset

Bhaskarjyoti Das
Computer Science and Engineering
PES University
Bangalore, India
bhaskarjyoti01@gmail.com

Sai Deepika Kakumani
Computer Science and Engineering
PES University
Bangalore, India
kakumani.saideepika16@gmail.com

Vridhi Goyal
Computer Science and Engineering
PES University
Bangalore, India
goyalvridhi@gmail.com

Nutheti Nikhila Priya
Computer Science and Engineering
PES University
Bangalore, India
nikhila.nutheti@gmail.com

***Abstract*—With the surge of the Internet and growing influence of social media, disinformation technique such as propaganda is increasingly getting used by both proponents and opponents of any modern-day social movement. Hence propaganda detection has emerged as an important task in modern disinformation research. However, the existing linguistic propaganda research is mostly limited to news articles, and to the best of our knowledge, there is no large labeled social media propaganda dataset available to date. The anti-CAA movement (2019) in India showcased the prominent role of social media in modern social movements and researchers have also conclusively proved the presence of inauthentic users on both sides of the discourse pursuing propagandistic goals. In this paper, we present a weakly supervised learning methodology to address the incomplete supervision challenge in supervised learning and present a large tweet-based dataset for textual propaganda detection. The data set consists of tweet ids, hashtags used, and corresponding propaganda techniques. The efficacy of the dataset is proved by building a transformer-based propaganda detection model that shows an accuracy of 87% on manually labeled test data. The dataset has been published to facilitate future research and automatic linguistic propaganda detection on Twitter.**

***Index Terms*—Weakly Supervised Learning, Data Programming, Linguistic Propaganda technique, Propaganda dataset, Twitter**

I. Introduction

Ever since people started moving online, social media has occupied a place of prominence in social movements. This is proved beyond doubt in various social movements in the past decade i.e. Egypt in 2011, Iran in 2009, Turkey in 2013, and Ukraine in 2013. In a modern protest, Twitter is able to successfully address the information asymmetry between proponents and opponents. The 2013 Brazil Vinegar protests, the 2019 Hongkong protests, and the 2020 Black Life Matter (BLM) movement in the USA are examples of this. Unlike social media mediums such as Facebook and YouTube, Twitter has become the ubiquitous medium of choice for people to rally around a cause, express their opinions as well as rally others around the same opinion. Twitter is now increasingly getting used as a tool for organizing and mobilizing a protest. India is no exception and has a lot of users who actively express their opinion and debate about ongoing events on Twitter. The 2019 protest in India against the Citizenship Amendment Act (CAA) [1]–[3] showcases the active role of social media like Twitter in a modern social movement.

The rise of social media has democratized content creation and has made it easy for everybody to share and spread information online. On the positive side, this has given rise to citizen journalism and a much bigger scope of 'virality' compared to what was possible with newspapers, radio, and TV. On the negative side, stripping traditional media from their gate-keeping role has left the public unprotected against the spread of disinformation, which could now travel at breaking-news speed. Disinformation and bias have become common features of any social event as witnessed on social media. In a modern protest movement, disinformation campaigns like propaganda take a prominent role. Propaganda is defined as an expression of opinion or specific actions by an individual or a group with the objective of influencing those of other groups or individuals. In a recent study of the anti-CAA movement based on Twitter conversations, Kumari Neha et al. [4] conclusively confirmed the presence of inauthentic users on both sides of the discourse.

With social media discourse increasingly playing a prominent role in modern-day society, disinformation research is now an important emerging area of research. However, propaganda research is still in its infancy and existing work in linguistic propaganda detection has been mostly limited to news articles. The main challenges are the scarcity and cost of

labeled propaganda datasets on social media platforms such as Twitter. As of now, there is no large labeled propaganda dataset on Twitter for linguistic propaganda detection. To address this bottleneck, a range of traditional approaches has seen renewed popularity. Semi-supervised learning uses unlabeled input in addition to a little amount of labeled data to enhance the model's performance. Transfer learning involves using a pre-trained model on one domain to enhance the performance of the model on a different target domain. But even these methods still need clean annotated data to operate satisfactorily, hence not overcoming the label scarcity hurdle. Weak Supervision, on the other hand, is a method of machine learning where one combines various noisier sources of supervision to generate considerably bigger training data much faster than manual methods.

This paper describes a weakly supervised learning approach for coming up with a large labeled linguistic propaganda dataset consisting of tweets collected from Twitter conversations during anti-CAA protests. In this work, a combination of weak supervision strategies has been used. The transformer-based supervised learning model based on this proposed dataset is shown to have an accuracy of 87%. The dataset developed has also been published [1] according to the FAIR guiding principle [5].

II. Background and Related Work

A. Propaganda detection

Propaganda as a term is rooted in history when it was used to spread [6]. Catholic faith. Content-wise, it can be defined as an expression of opinion with the specific end goal of influencing and converting the target audience. Propaganda can be both positive and negative with respect to the emotion it carries and can be white or black with respect to the truthfulness of the content. The seven main techniques of propaganda i.e. transfer, bandwagon, name calling, plain folks, testimonial, glittering generalities, and card-stacking were first identified in 1937. Subsequently, other techniques were also identified, which were essentially sub-classes of these. Modern propaganda research started with document-level binary classification and got further refined by the identification of 18 propaganda techniques at the text fragment level [7] to enable automated detection. Since propaganda on social media has an aspect of execution in a planned manner with the help of willing accomplices, modern propaganda research is additionally attempting to consider the patterns in the propaganda networks on social media. However, the work described in this paper is primarily focused on content-based propaganda detection.

Content-based propaganda research started with news article level binary classification or regression [8] followed by detection of the span of text [7], [9], [10] containing propaganda techniques. The existing work used supervised machine learning with both feature engineering [11], [12] and transformer-based [9], [10], [13] approaches. These tasks primarily used SEMEVAL (2019) dataset with strong linguistic cues and similar marginal distribution of features [11], [14] in both train and validation parts of the datasets. So, the learning model trained on this dataset does not generalize well across domains.

Content-based propaganda detection however has been mostly limited to propaganda detection on news articles even though propaganda research has become important today primarily because of its use in social media platforms such as Twitter. The popular benchmark corpus [7] for text-based propaganda detection was based on news articles. There are only a handful of content-based propaganda research [15]–[18] using social media datasets. As natural language processing tasks on article text and Tweets differ both due to socio-linguistics and platform-imposed content limitations on Twitter, a model trained on the former cannot give a satisfactory performance on the latter. This necessitates a domain adaptation [19], [20] that has its own challenges. Hence, there is a clear need for a textual propaganda detection dataset on Twitter. At the time of writing, to the best of our knowledge, there is no such textual propaganda detection dataset of tweets and the work described in this paper addresses that gap besides showcasing a weakly supervised learning methodology.

B. Weakly supervised learning

One of the key challenges of supervised learning is the lack of sufficient labeled data. Due to the sensitive nature of privacy issues involved and the prohibitive cost of annotation, getting sufficient fully labeled data on social media platforms is very difficult. In a typical scenario involving social media dataset, supervised learning is a challenge and researchers have to rely on weak supervision. Zia-Hua Zhou [21] defined three categories of weakly supervised learning i.e. 'incomplete supervision' where only a small subset of the data has labels, 'inexact supervision' where only coarse-grained labels are available, and 'inaccurate supervision' where the labels are noisy. In the case of textual propaganda detection on Twitter, it is primarily the case of incomplete supervision where only a limited number of labeled samples are available. If a particular tweet is only labeled as a propagandist but no labels are provided for the particular propaganda technique being used, then it is a case of inexact supervision. The work described in this paper attempts to address both of these scenarios.

From the perspective of incomplete supervision, semi-supervised learning that tries to utilize the availability of a large amount of unlabeled data with a very small amount of labeled data is a possible strategy to address incomplete supervision. Semi-supervised learning attempts learning either in a transductive or an inductive setup using some assumptions about the data space such as smoothness assumption, low-density assumption, cluster assumption, and manifold assumption. Semi-supervised learning works only when these assumptions are valid. Also, semi-supervised learning will not work if no labeled dataset is available as in the case of the propaganda dataset of tweets. Hence, the work described in

[1] Bhaskarjyoti Das, Nutheti Nikhila Priya, Sai Deepika Kakumani, & Vridhi Goyal. (2023). Multi-label Tweet Dataset for Textual Propaganda Detection related to anti-CAA protest in India (2019-2021) (1.0.0) https://doi.org/10.5281/zenodo.7797035

this paper follows an approach relying on generating weak supervision.

Joshua Robinson et al. [22] showed that a large number of available weak labels can easily surpass the performance of a small number of available strong labels. The existing work that relies on this strength of weak supervision, can be roughly divided into the following categories:

1) **Ensemble** [23] of weak learners deliver a learner that is stronger than one particular weak learner. However, this method is more appropriate for inaccurate supervision.
2) **Crowdsourcing** [24] has been tried out as an alternative to active learning. The challenge in this approach is to combine the labels provided by different human annotators who may have different biases. Web knowledge extraction to generate sufficient weak labels is also an approach but it is more suitable for specific domains such as health informatics.
3) **Adversarial label learning** [25] has seen a lot of work in recent years but this is more suitable for inaccurate supervision rather than incomplete supervision.
4) **Data Programming** [26] can be used to effectively address the case of incomplete supervision. Data programming allows users to define labeling functions that generate noisy labels and the data programming framework Snorkel [27] uses these generated noisy labels with the help of generative models to finally generate [28] probabilistic labels for unlabeled data. Snorkel also provides an estimate of the accuracy of such generated weak supervision.

For the reasons mentioned above, the work described adopts a data programming approach for weakly supervised learning. However, it is not possible to come up with an effective labeling function for all eighteen propaganda techniques due to the fine differences involved. For the remaining few labels where data programming could not be used, a model pre-trained on a limited number of samples from propagandist news articles from the SemEval 2019 [7] dataset has been used to generate weak labels. Finally, these two sets of labels generated from these two weakly supervised learning approaches have been combined to generate a large multi-label tweet dataset labeled for propaganda techniques.

III. Data Collection and pre-processing

For collecting anti-CAA protest-related tweets, the following steps were followed :

1) Anti-CAA protest-related tweets identified by related hashtags were scraped using Twitter search API by using the open source library 'snscrape' [29]. This library produces a text file containing the list of tweet ids.
2) As a next step, the tweets were retrieved using Twitter REST API using Tweepy.
3) Since individual tweets use multiple hashtags, the collected tweets had a lot of duplicates. Duplicate tweets were removed.
4) Mixed language tweets where the primary language was English, were translated into English, and tweets in other languages (Urdu, Hindi, etc,) were dropped using a Python language detection tool called Fasttext.
5) Hashtags were extracted from the tweets and made into a separate column. Special characters and URLs were removed.

IV. Exploratory Analysis

To understand the collected tweets better, a series of analyses were done. The collected tweets were in various languages and had to be translated into English for uniformity and ease. Before the translation process, a total of 48 languages were observed, English and Hindi being the major ones.

After analyzing the Web scraped data, 24 attributes related to a single tweet were found. Counts of the unique values of these attributes were calculated. As shown in Figure 4, 27232 unique user ids were found active during the protest and 8136 user locations were observed.

Out of all the locations, as shown in Figure 5, New Delhi was the most active with Mumbai following closely and Karachi in the neighboring country Pakistan was the third.

The dataset contains a total of 15790 different hashtags but only some of them are prevalent across tweets. This indicates that the other hashtags were primarily meant to achieve a larger audience and mask the real intent of the message. Figure 7 displays the count of the top 10 hashtags found.

From Figure 2, the count of each propaganda technique used can be observed. Name Calling has the highest number of instances whereas thought-terminating cliche has the least.

To understand the correlation between each label, a heat map, as shown in Figure 6, was drawn that displays the frequency of tweets containing each pair of propaganda techniques. Name-calling and Loaded Language appeared to be labeled together the most.

It was observed that some tweets in the dataset had multiple propaganda techniques. Figure 3 exhibits the count of tweets with multiple propaganda techniques as their labels.

Sentiment analysis was conducted on the tweet context to get a better understanding of positive and negative propaganda. As shown in Figure 1, the numbers of tweets carrying positive and negative sentiments are almost balanced. It indirectly shows the intense messaging effort put in by both sides of the discourse.

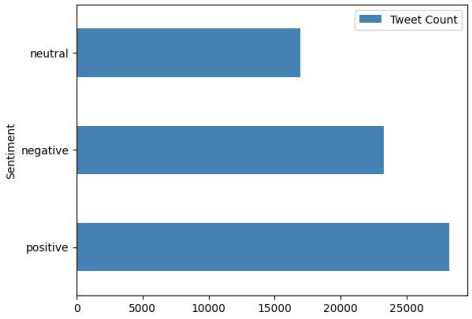

Fig. 1. Sentiment polarity of tweets.

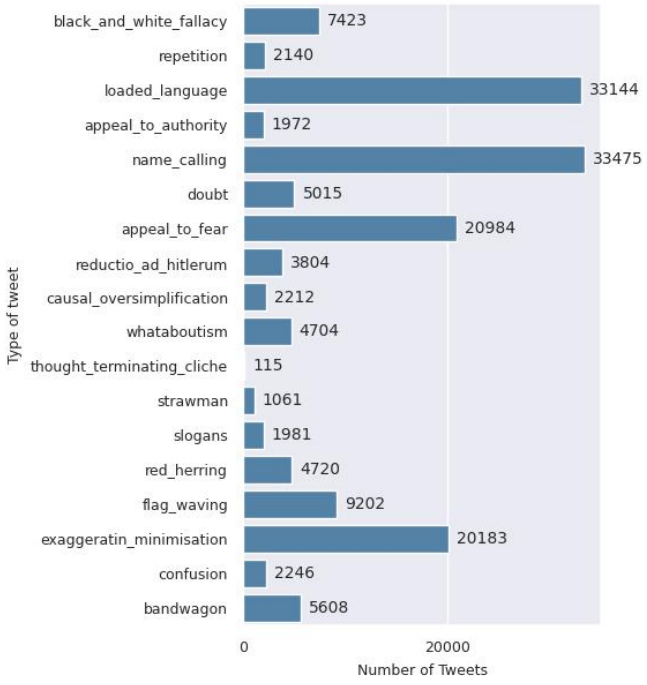

Fig. 2. Frequency of each label in the dataset.

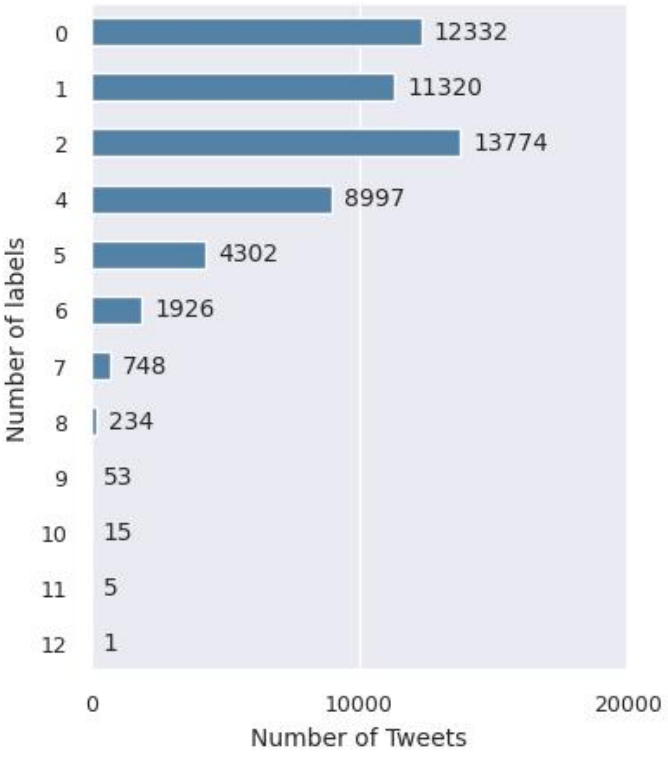

Fig. 3. Count of tweets with multiple labels.

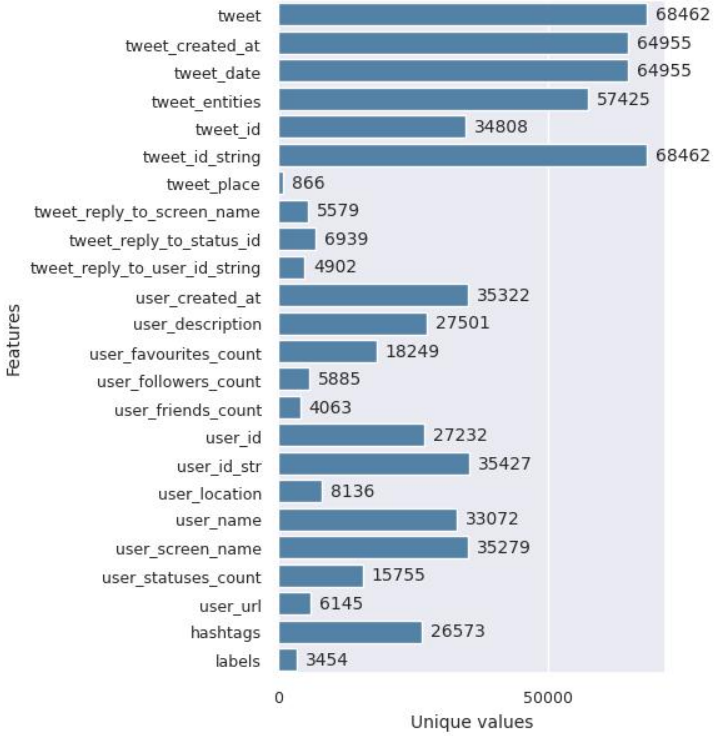

Fig. 4. Dataset Columns with unique values.

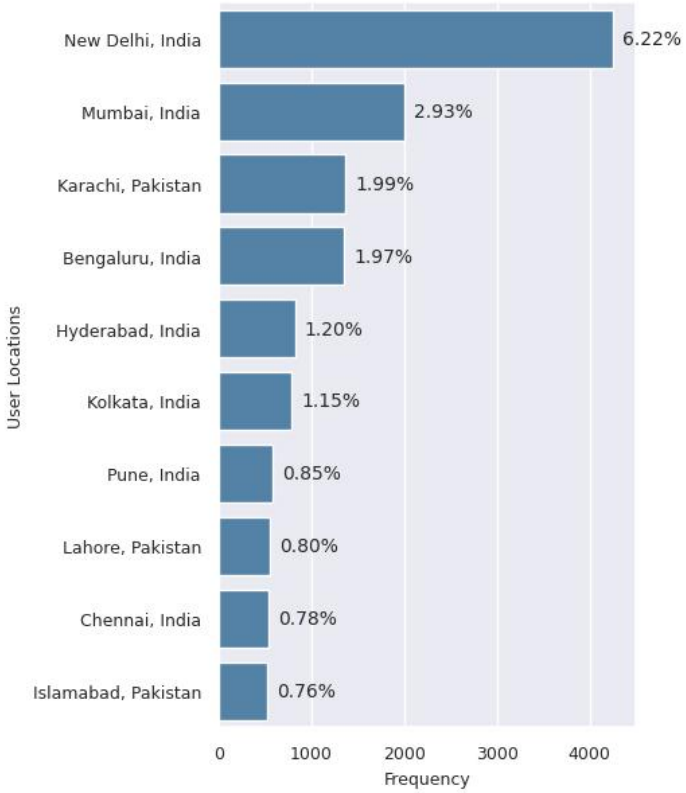

Fig. 5. Location of Users.

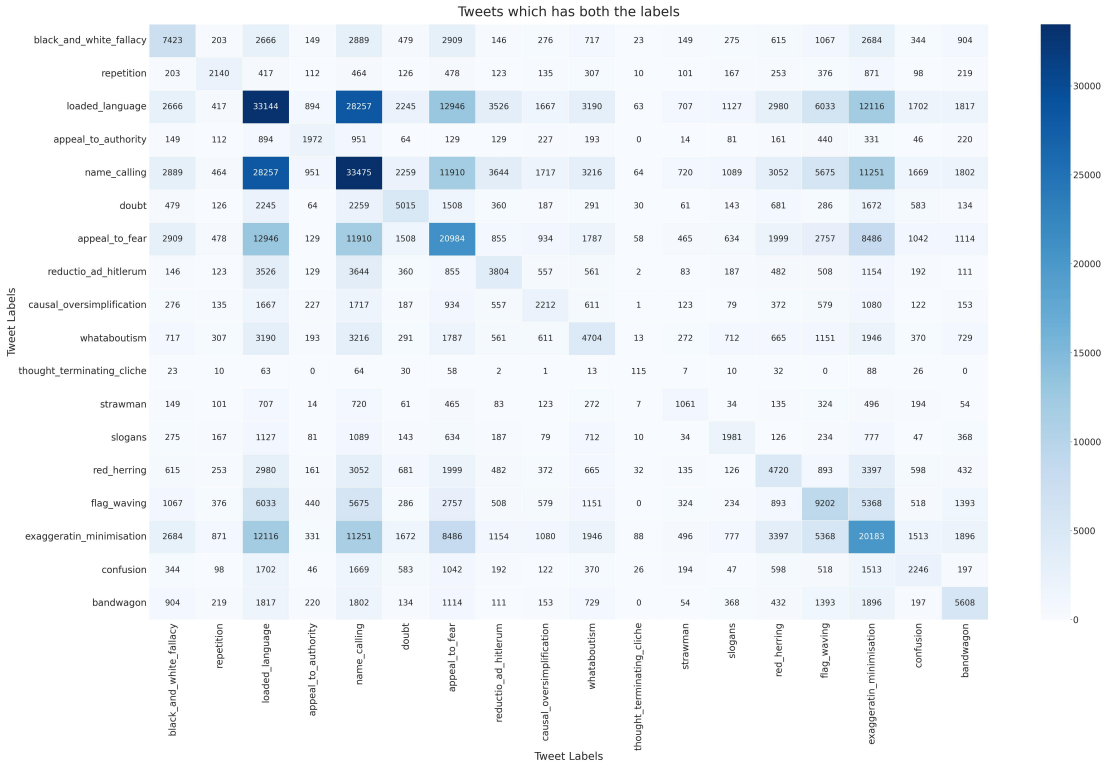

Fig. 6. Frequency of tweets containing both labels.

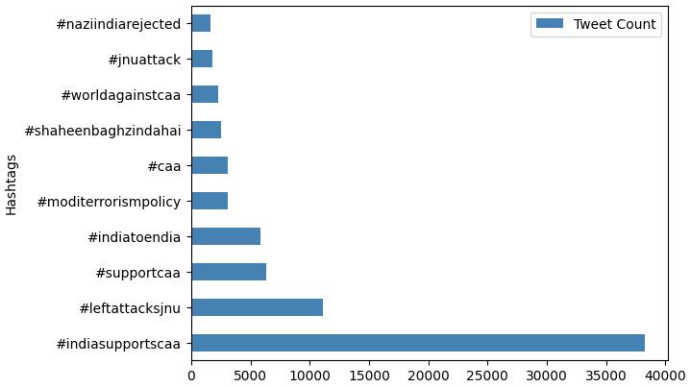

Fig. 7. Hashtag Frequency.

V. Annotation Setup

This section describes the workflow for the annotation structure that we developed as a result of our extensive research and discussion. We also present details about each level of the pipeline and the platforms that are employed. We built the annotation pipeline after careful consideration and debate, followed by iterative improvement based on the observations.

The outcome of this step would be a weakly labeled dataset annotated with 18 distinct propaganda techniques namely Black-and-white Fallacy, Repetition, Loaded Language, Appeal to authority, Name calling/Labeling, Doubt, Appeal to fear/prejudice, Reductio ad hitlerum, Causal Oversimplification, Whataboutism, Thought-terminating cliche, Straw Man, Slogans, Red Herring, Flag-waving, Exaggeration/Minimisation, Confusion, Bandwagon.

A. Weak Supervision

The primary goal of our whole research is to eliminate and streamline the human labor involved in labeling. We came across a technique called Weak Supervision that allows labeling without human intervention. In Weak Supervision, we first look for the pattern corresponding to a label. We code labeling functions for the appropriate pattern that returns 1 if the pattern is detected in the input and 0 otherwise. In this manner, we create labeling functions for each label. After we've developed the labeling functions, we pass our unlabeled dataset through the pipeline to get the probabilities of all the labels on each of the data instances. Later, we established a probability threshold for generating numerous labels for an instance. Snorkel is a Python package that allows us to accomplish the aforementioned processes. We were able to write labeling functions for nine out of eighteen labels in our study. The following is the code logic for each of the labeling functions after sending tweet data as input to the labeling function:

Black And White Fallacy: The programming logic for recognizing this specific propaganda technique begins with extracting the tweet's keywords. The program will next search for antonyms for each keyword and determine whether they overlap with other keywords or their synonyms. If a match is

found, it belongs to that label, and the labeling function will return the label applied to it or -1 to indicate abstention. For this code, we used the nltk library and existing Wordnets.

Algorithm 1 Black And White Fallacy

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
list = Extract the keywords from tweet
for *i in list* **do**
  for *j in synonyms(i)* **do**
    if *antonym(j) is present in list* **then**
      return 1
    end
  end
end
return -1

Appeal to authority: Each data instance was checked for authoritarian names or quotes made by influential people. This was done using the Stanford NER tagger which identifies various organizations and names used in the text. If the instance had a name/organization entity, we returned 1 else -1.

Algorithm 2 Appeal to authority

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
initialization;
Tokenize the data
Extract Tags from token
if *Tag == 'PERSON'* **then**
  return 1
else
  return -1
end

Name calling/Labeling: For this propaganda, we searched the text for negative words. We web-scraped and manually created a negative word corpus which was used to compare with the text's content. If the instance contained any negative word, the function returns 1 or else -1.

Algorithm 3 Name Calling

Data: Tweet Data, Negative Corpus
Result: Returns 1 if the tweet belongs to the label else -1
initialization;
Tokenize the data
if *data in Negative Corpus* **then**
  return 1
else
  return -1
end

Doubt: For this propaganda technique, we employed an uncertainty estimator, a pre-trained algorithm that forecasts the likelihood of ambiguity or doubt in the tweet. In accordance with the performance, we establish a threshold. If the value that is generated is above the threshold, it belongs to that label otherwise, the labelling function will return -1.

Algorithm 4 Doubt

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
data = certainity_estimator(tweet)
if *data.certain* $\leq$ *0.5* **then**
  return 1
end
else
  return -1
end

Repetition: To identify Repetition in the text, first, the text was tokenized then from which, we extracted n-grams (n=1,2,3,4). The count of each n-gram was calculated along with its synonyms to accurately recognize repetition propaganda. If the count is greater than a set threshold, the labeling function would return 1 or else -1.

Algorithm 5 Repetition

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
initialization;
for $iteration = 2, \ldots, 5$ **do**
  Extract n-grams;
  Count the occurrence of each n-gram;
  if *count* $\geq$ *2* **then**
    return 1
  else
    return -1
  end
end

Appeal to fear/prejudice: Here we utilized a pre-existing Python package called text2emotion, which analyzes the emotions in the text and converts them into a dictionary with the emotion as the key and the probability of that emotion in the text as the value. The label generated by the labeling function is then found using a threshold.

Algorithm 6 Appeal to Fear/Prejudice

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
data = get_emotion(tweet)
if *data['Fear']* $\geq$ *0.5* **then**
  return 1
end
else
  return -1
end

Flag-waving: Each data instance was checked for occurrences of words that play on strong national feelings, positive

or negative. The text was compared against a manually created and web-scraped corpus of words. If the instance had any matching word, we returned 1 else -1.

Algorithm 7 Flag Waving

Data: Tweet Data, Word Corpus
Result: Returns 1 if the tweet belongs to the label else -1
initialization;
Tokenize the data
Extract Keywords
if *data in Word Corpus* **then**
  return 1
else
  return -1
end

Loaded Language: For this technique we used the Vader Sentiment library, which is a sentiment analysis tool that outputs the positive, neutral, negative, and compound scores of a sentence that has been passed. It is a tool with a lexicon and rules adjusted to the sentiments expressed in social media. Vader makes use of a variety of sentiment lexicons that are often classified as either positive or negative depending on their orientation of semantics. So based on the scores, whether highly positive or negative, it is labeled as 1 or -1.

Algorithm 8 Loaded Language

Data: Tweet Data
Result: Returns 1 if the tweet belongs to the label else -1
initialization;
Find the negative, positive, neutral, and compound polarity scores using the Vader sentiment analysis tool
Find the extremely negative polarity scores
if *neu* $\geq$ *0.8* **then**
  return -1
end
if *pos* $\geq$ *0.5 or neg* $\geq$ *0.5* **then**
  return 1
end
if *compound* $\geq$ *0.5 or compound* $\leq$ *-0.5* **then**
  return 1
end
else
  return -1
end

Reductio ad hitlerum: For this propaganda technique, the logic used is the same as for the loaded language. The Vader sentiment analysis tool is used for getting a dictionary of negative, neutral, positive and compound polarity scores of a sentence. Based on the scores if it is highly negative then the labelling function returns 1 otherwise -1.

B. Weak Classifier

Devising coding logic for the remaining nine labels in our model is a tedious and NP-hard problem. To address this, we developed a Weak Classifier. Weak Classifier is a supervised model that is trained on a very little quantity of data because there is not a lot of data available for our investigation. We used SEMIEVAL 2020 data instances as our training data. We have experimented a variety of binary and multi-labeling supervised models and discovered that developing a supervised binary classifier for each of the eight labels provided the highest accuracy.

Algorithm 9 Reductio Ad Hitlerum

Data: Tweet data
Result: Returns 1 if the tweet belongs to the label else -1

STEP 1: Find the negative, positive, neutral and compound polarity scores using the vaderSentiment analysis tool

STEP 2: Find the extreme positive and extreme negative polarity scores
if *neg* $\geq$ *0.5 or compound* $\leq$ *-0.4* **then**
  return 1
end
else
  return -1
end

VI. Experiments and Evaluation

After applying both the Snorkel and the Weak classifier model on the tweets, their outputs were combined to form an 18-class multi-labeled dataset. The Snorkel model on its own was tested against criteria like polarity, which identifies the unique labels that the labeling functions outputs; coverage, which is the total fraction of the dataset that is labeled by the model; overlap, which calculates how much each labeling function agree and conflict, which calculates how much each labeling function disagree.

The snorkel model was run in various batches due to the large size of the dataset. As shown in Table I, Name Calling and Loaded Language had the highest coverage as they are the most common forms of propaganda found in social media content.

TABLE I
Snorkel Metrics

Label	Coverage	Overlap	Conflict
Black&White Fallacy	0.2945	0.2845	0.2845
Repetition	0.1095	0.1040	0.1040
Loaded Language	0.6880	0.5735	0.5735
Appeal to Authority	0.2060	0.1765	0.1765
Name Calling	0.5185	0.4985	0.4987
Doubt	0.1165	0.1075	0.1075
Reduction Ad Hitlerium	0.1680	0.1670	0.1670
Fear	0.2305	0.2045	0.2045

To further evaluate the model, the dataset was divided into training and validation sets. A pre-trained BERT model, cased

TABLE II
FARMER'S PROTEST TWEETS

Tweets	Label
A nation that destroys its soils destroys itself. Farmers are the lungs of our nation. REMEMBER it	Loaded Language, Whataboutism, Slogans, Exaggeration/Minimisation
No Farmer No Food	Misrepresentation of Someone's Position (Straw Man), Slogans, Presenting Irrelevant Data (Red Herring)
Stop Hating on Farmers. We are the Nation Builders. Just observe the period of CORONA, It is the Agriculture Sector that railed the National Economy on track. You people, Hate them... THINK. United we STAND and WIN	Name-calling/Labeling, Doubt, Thought-terminating cliché, Flag-waving
Farming is an occupation. It is not limited to a religion or a caste	Name-calling/Labeling, Appeal to fear/prejudice, Reductio ad hitlerum, Thought-terminating cliché
If you are neutral in situations of injustice, you have chosen the side of the oppressor. Speak up for farmers in India. Don't be an oppressor	Repetition, Black-and-white Fallacy, Loaded Language, Presenting Irrelevant Data (Red Herring), Confusion
To understand something, we need to question our understanding first	Loaded Language, Causal Oversimplification
Smile and Shine, Long Live Farmers Unity !!	Appeal to authority, Slogans
Farmers are asking for what's rightfully theirs. Repeal of farm bills. These bills were passed without a debate or any discussion with those who are impacted the most. so why can't they ask to repeal what they didn't want to begin with?	Whataboutism, Exaggeration/Minimisation, Bandwagon

English was applied as this encoding would be more accurate in understanding the context of the tweets. Furthermore, a test dataset was created manually with a portion of the tweets to evaluate our model.

As the dataset contains multiple labels, using absolute metrics like accuracy, precision, etc will not evaluate accurately. Hence, Example-based Accuracy, also known as Hamming Score was used. It calculated the average difference between the predicted and true labels of each data point, averaged over all classes.

The BERT model was applied to the combined output of the snorkel and weak classifier. The dataset was divided into training and test datasets in the ideal 80:20 ratio. We achieved a training accuracy of 94.5%. We also used the BERT model to predict the labels of the test dataset which yielded an accuracy of 92.7%.

As an additional validation, the trained model was also run to detect text-based propaganda techniques on 100,000 tweets of Farmer's Protest [30]–[32] in India (2020-2021) that has witnessed another social movement on Twitter with aggressive use of propaganda techniques both sides. This dataset is already available [33] at Kaggle. The trained model was able to show a very successful tweet-level propaganda detection and it was manually verified. This is shown in Table II.

VII. CONCLUSION AND FUTURE WORK

Supervised learning techniques are more effective when large annotated datasets are available. But labeling the data has become a bottleneck due to the cost associated with it in terms of time and human labor. In this work, we have demonstrated the feasibility of utilizing weak supervision to obtain results like supervised learning, which can help reduce the need for manual annotation and save time and resources. Using Weakly supervised learning techniques, we have presented a new tweet dataset for propaganda techniques on Twitter. This addresses a long-standing need for the emerging research area of propaganda detection on social media.

As recent propaganda research on Twitter uses the twin strategy of content-based techniques along with the detection of coordinated inauthentic behavior (CIB), we plan to investigate next the same dataset towards identifying the network patterns in the Twitter conversation graphs.

REFERENCES

[1] S. Roy, M. Mukherjee, P. Sinha, S. Das, S. Bandopadhyay, and A. Mukherjee, "Exploring the dynamics of protest against national register of citizens & citizenship amendment act through online social media: the indian experience," *arXiv preprint arXiv:2102.10531*, 2021.

[2] A. K. Kushwaha, S. Mandal, R. Pharswan, A. K. Kar, and P. V. Ilavarasan, "Studying online political behaviours as rituals: a study of social media behaviour regarding the caa," in *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, Proceedings, Part II*, pp. 315–326, Springer, 2020.

[3] D. B. Edingo, "Social media, public sphere, and counter publics: An exploratory analysis of the networked use of twitter during the protests against the citizenship amendment act in india," *The Journal of Social Media in Society*, vol. 10, no. 2, pp. 76–101, 2021.

[4] K. Neha, V. Agrawal, V. Kumar, T. Mohan, A. Chopra, A. B. Buduru, R. Sharma, and P. Kumaraguru, "A tale of two sides: Study of protesters and counter-protesters on# citizenshipamendmentact campaign on twitter," in *14th ACM Web Science Conference 2022*, pp. 279–289, 2022.

[5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[6] G. S. Jowett and V. O'donnell, *Propaganda & persuasion*. Sage publications, 2018.

[7] G. Da San Martino, A. Barron-Cedeno, and P. Nakov, "Findings of the nlp4if-2019 shared task on fine-grained propaganda detection," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pp. 162–170, 2019.

[8] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, and V. Basto-Fernandes, "Propaganda detection in text data based on nlp and machine learning," in *CEUR Workshop Proceedings*, vol. 2631, pp. 132–144, 2020.

[9] R. Patil, S. Singh, and S. Agarwal, "Bpgc at semeval-2020 task 11: Propaganda detection in news articles with multi-granularity knowledge sharing and linguistic features based ensemble learning," *arXiv preprint arXiv:2006.00593*, 2020.

[10] G. Da San Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeno, and P. Nakov, "Prta: A system to support the analysis of propaganda techniques in the news," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 287–293, 2020.

[11] A. Barrón-Cedeno, I. Jaradat, G. Da San Martino, and P. Nakov, "Proppy: Organizing the news based on their propagandistic content," *Information Processing & Management*, vol. 56, no. 5, pp. 1849–1864, 2019.

[12] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "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*, pp. 2931–2937, 2017.

[13] S. Yoosuf and Y. Yang, "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*, pp. 87–91, 2019.

[14] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," *arXiv preprint arXiv:1910.02517*, 2019.

[15] G. Caldarelli, R. De Nicola, F. Del Vigna, M. Petrocchi, and F. Saracco, "The role of bot squads in the political propaganda on twitter," *Communications Physics*, vol. 3, no. 1, pp. 1–15, 2020.

[16] S. Cresci, "A decade of social bot detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72–83, 2020.

[17] L. Nizzoli, M. Avvenuti, S. Cresci, and M. Tesconi, "Extremist propaganda tweet classification with deep learning in realistic scenarios," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 203–204, 2019.

[18] A. Tundis, G. Mukherjee, and M. Mühlhäuser, "Mixed-code text analysis for the detection of online hidden propaganda," in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–7, 2020.

[19] M. Rajmohan, R. Kamath, A. P. Reddy, and B. Das, "Emotion enhanced domain adaptation for propaganda detection in indian social media," in *Innovations in Computational Intelligence and Computer Vision*, pp. 273–282, Springer, 2022.

[20] L. Wang, X. Shen, G. de Melo, and G. Weikum, "Cross-domain learning for classifying propaganda in online contents," in *Conference for Truth and Trust Online 2020 (TTO)*, 2020.

[21] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[22] J. Robinson, S. Jegelka, and S. Sra, "Strength from weakness: Fast learning using weak supervision," in *International Conference on Machine Learning*, pp. 8127–8136, PMLR, 2020.

[23] A. Balsubramani and Y. Freund, "Scalable semi-supervised aggregation of classifiers," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[24] E. A. Platanios, M. Al-Shedivat, E. Xing, and T. Mitchell, "Learning from imperfect annotations," *arXiv preprint arXiv:2004.03473*, 2020.

[25] C. Arachie and B. Huang, "Adversarial label learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3183–3190, 2019.

[26] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," *Advances in neural information processing systems*, vol. 29, 2016.

[27] S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, *et al.*, "Snorkel drybell: A case study in deploying weak supervision at industrial scale," in *Proceedings of the 2019 International Conference on Management of Data*, pp. 362–375, 2019.

[28] A. J. Ratner, S. H. Bach, H. R. Ehrenberg, and C. Ré, "Snorkel: Fast training set generation for information extraction," in *Proceedings of the 2017 ACM international conference on management of data*, pp. 1683–1686, 2017.

[29] JustAnotherArchivist, "snscrape."

[30] D. Mishra, S. Z. Akbar, A. Arya, S. Dash, R. Grover, and J. Pal, "Rihanna versus bollywood: Twitter influencers and the indian farmers' protest," *arXiv preprint arXiv:2102.04031*, 2021.

[31] P. Waghre, "Radically networked societies: The case of the farmers' protests in india," *Indian Public Policy Review*, vol. 2, no. 3 (May-Jun), pp. 41–64, 2021.

[32] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of indian farmers' protest using twitter data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100019, 2021.

[33] P. Sharma, "Farmers protest tweets dataset(csv)."