2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Tweeted Fact vs Fiction: Identifying Vaccine Misinformation and Analyzing Dissent

Shreya Ghosh* and Prasenjit Mitra*†
*College of IST, The Pennsylvania State Univerity, USA. Email: {shreya,pmitra}@psu.edu
†L3S Research Center, Leibniz University, Hannover, Germany

***Abstract*—In this paper, we develop an end-to-end knowledge extraction and management framework for COVID-19 vaccination misinformation. This framework automatically extracts information consistent and inconsistent with scientific evidence regarding vaccination. Additionally, using novel natural language processing methods (including triple-attention based sarcasm detection and utilizing topic-based similarity scoring, agglomerative clustering, and word embedding vectors for misinformation category identification and counter-fact summarization in a semi-supervised way from web-based sources), we explore public opinion towards vaccination resistance. Our knowledge extraction pipeline constructs knowledge-bases automatically, categorizes vaccine dissenting tweets into 15 misinformation categories automatically, and effectively analyzes discourses in those tweets. Our contributions are as follows: (i) the proposed knowledge extraction framework does not require huge amounts of labelled tweets of different categories (our method uses only 50-labelled tweets for each of 15 misinformation categories, in stark contrast to existing approaches that typically rely on 10,000 or more labelled tweets), and (ii) our module outperformed baselines by a significant margin of $\approx 8\%$ to $\approx 14\%$ (F1 score) in the classification tasks using Twitter dataset.**

I. Introduction

Social media platform such as, Twitter has played a crucial role in connecting people, sharing and voicing their opinions regarding their safety and medical needs without the limitation of one-way communication (i.e, television, radio). Twitter also acts as platform to propagate information that is not based on scientific consensus [1]. In this paper, we (i) characterize misinformation about vaccination on social media, and (ii) devise knowledge extraction techniques to identify vaccine dissenting discourse and users involved in such dissent, as well as users who change their stance based on such discourse.

Significance, Challenges, and Contributions.

Supervised learning can be used to automatically detect misinformation and true information consistent with scientific consensus from tweet discourse [2], [3]. Unfortunately, the availability of labelled data to train a supervised learning model is often insufficient. There is also temporal and location diversity along with other contexts, namely, external influence, political propaganda to name only a few, that impacts the public opinion in a significant way and the topic of discourse changes over time. Therefore, a fixed set of labels ("topics") of tweets does not seem realistic. We present a systematic knowledge extraction framework, which provides an overview of opinions expressed in tweets by analyzing the content (vaccine dissent and misinformation) and analyzing the linguistic and semantic characteristics of tweets leveraging novel machine learning methods at different temporal scales. Analyzing heterogeneous data sources and extracting implicit information becomes more challenging when such data-instances are dynamic (as topics of discourse change based on varied influences) and voluminous. *Specifically, our problem is to classify vaccine dissenting tweets into different classes based on the reasoning given to support them (See Table I)*. To achieve that, we need to identify public stance ("against", "in favour" and "neutral") and sentiment ("negative", "positive" and "neutral") towards vaccination, followed by analysing vaccine dissenting tweets ("against" stance category and "negative" sentiment) to identify misinformation topics. However, efficient identification of public opinion in terms of stance (expressed in misinformation tweets) and sentiment is not straightforward, since there is no defined contextualization process to deal with inherent ambiguities of opinions due to humor, irony and conversation context. Human conversations often consist of sarcasm and irony that is not easily detected by automated methods and that makes the problem more complex. The objectives and contributions of the paper (Fig. 1 depicts the overall framework presented in this work) are:

- **Knowledge-extraction framework**: Our work pioneers the development of an automated system that extracts knowledge from online sources to create a comprehensive knowledge base on vaccination-related misinformation and accurate information. We develop a misinformation identification technique with very limited labelled tweets to categorize tweets into different sub-classes of misinformation efficiently proposing a novel triple-attention based sarcasm detection module.
- **Vaccine dissenting discourse analysis**: We present an in-depth discourse analysis using a three-tier knowledge mining module to understand the characteristics of vaccine dissenting users and their tweets as well as their conversational features. These modules have shown promising

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627307

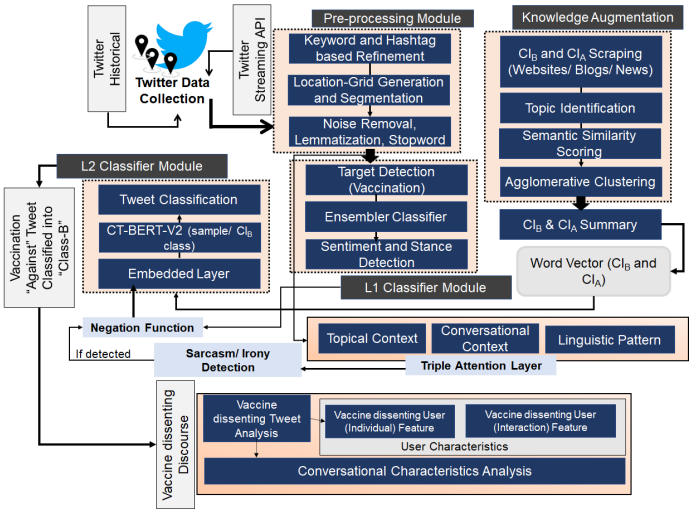

Fig. 1: Overall working modules of proposed knowledge extraction framework for vaccination misinformation and vaccine dissenting discourse (Cl_B: misinformation, Cl_A: fact)

accuracy in identifying the characteristics of vaccine dissenting discourse, e.g., when more users engage in vaccine dissenting discussion stating misinformation, and disapprove vaccination in Twitter.

- Our proposed framework outperforms baselines by a significant margin ($\approx$ 14%) in identifying misinformation in Twitter with limited labelled data and effective analyses of vaccine dissenting discourses.

II. Proposed Framework

A. Collection and Labelling of Tweet Data

We utilized the *Twitter Streaming API v2 (Academic Research)* to collect tweets ($\approx$ 130M) spanning October 2020 to January 2022 using a *keyword*-based search.[1] We annotated our dataset (using MTurk) to assign three labels to each tweet: *topic of tweet (tweet_To)*, *sentiment (tweet_S)*, and *stance (tweet_St)* for evaluation purposes. The *tweet_To* can belong to any of the sixteen categories, $M1 - M15$ or "Other" (refer to Table I). The *tweet_S* label comprises three categories: "positive", "negative", and "neutral"; while *tweet_St* also consists of three categories, specifically, "in-favor", "against", and "neutral". The geo-location (latitude, longitude) of a tweet is converted to a specific location-string (country, state, city, etc.) using *reverse geo-coding* and the *Google Place API*[2]. A *timeline (timeL)* of a topic represents a chronological sequence of user engagement (tweet, retweet, comment) count for that topic. For instance, such topics may include *vaccine unsafe* or *vaccine affecting fertility*, with labels representing stance (*against*, *in-favor*, and *neutral*) and sentiment (*negative*, *positive*, and *neutral*). The timeL illustrates the trend of user engagement on the topic across different stance and sentiment categories over a specific time period.

[1]The keyword list and annotation details are provided in the supplemental document (here) due to page limitations.

[2]https://developers.google.com/maps/documentation/places/web-service/overview

ID	Subclasses of misinformation	Ratio of tweets (Sarcasm)
M1	vaccine-unsafe-die	0.22 (0.27)
M2	vaccine-substance-development	0.26 (0.18)
M3	vaccine-natural-immunity	0.38 (0.12)
M4	vaccine-makes_me_sick	0.47 (0.11)
M5	vaccine-pregancy-fertility	0.29 (0.37)
M6	vaccine-side-effect	0.43 (0.20)
M7	vaccine-alter-DNA	0.21 (0.62)
M8	vaccine-microchip-tracking	0.26 (0.507)
M9	vaccine-not_recommended	0.41 (0.13)
M10	vaccine-unnecessary	0.52 (0.11)
M11	vaccine-trust_issue	0.37 (0.13)
M12	vaccine-child-infant	0.40 (0.10)
M13	vaccine-gain-big_companies	0.39 (0.12)
M14	mask-regulation-not-required	0.36 (0.16)
M15	vaccine-not_for_me	0.34 (0.13)

TABLE I: Misinformation sub-classes extracted by our framework (detailed in supplemental document). One tweet can be categorized into multiple labels resulting total number of tweets $\geq$ 1. The ratio of tweets in each category presenting sarcasm is denoted within ()

Initially, we employ a POS tagger (Flair[3]) to tag each word in $tweet\_text$. Next, we perform *Lemmatization* to convert words to their base forms using the *WordNetLemmatizer* function from the NLTK Python library.

B. L1 classifier:Vaccination (In favour/ Against/ Neutral)

Our first module (L1 classifier) attempts to classify tweets into three categories: "in favour", "against" and "neutral". The **Stance detection** module is implemented by ensembling transformer-based pre-trained encoders, namely, $BERT_{LARGE}$, $BERTweet$ [4] and $COVID-Twitter-BERT$ [5]. *COVID-Twitter-BERT* is pre-trained on 97M tweets related to COVID-19. BERTweet is trained using 850M tweets and achieves state-of-the-art benchmarks on both SemEval 2017 [6] sentiment analysis and SemEval 2018 irony detection [7] shared tasks. We selected two BERTweet models (BERTweet-base and BERTweet-covid19-base-cased) and fine-tuned for three downstream tasks: stance detection, sentiment detection and emotion-detection. Hinton, Vinyals, and Dean proposed a *student-teacher* architecture [8] to transfer knowledge from a large teacher model to a small student model by capturing the behaviors of the teacher model. We utilize a *knowledge distillation method* [8] where the teacher model is a self-voted BERT[4], and represented as:

$$L(x,y) = CrE(BERT(x,\vartheta),y) + \chi MSE(BERT(x,\vartheta), \frac{1}{T}\sum_{i=1}^{T} BERT(x,\vartheta_{t-i})) \quad (1)$$

where BERT(x, ϑ) is the student model, χ is the weight parameter to balance the importance of two loss functions, namely, mean squared error (MSE) and cross-entropy (CrE). However, we propose a different distillation strategy (two-stage fine-tuned strategy) for stance classification. In the first

[3]https://huggingface.co/flair/pos-english

[4]Fine tuning multiple BERT with random seeds, and output is the prediction with high probability.

stage, teacher model (pre-trained BERTweet-base on SemEval stance detection dataset) produces stance classes on data (vaccination), which is used as labelled samples to train student models (COVID-Twitter-BERT and BERTweet-covid19-base-cased). In the next stage, ground truth label data (vaccination) is used to fine-tune the student models to achieve better performance as well reducing overall computational cost.

Another important feature useful for stance detection is the structure of the social networking platform, i.e, social connections and interactions among the users, who voice out their opinion. The above-mentioned distillation method leveraging BERT models attempts to classify stance based on linguistic patterns. However, network features give us strong cues about a person's stance and help us to understand the alignment of a user towards a topic. Connected users influence each other. This work uses two network features: (i) interaction network, where retweets, replies, or any direct mentions are analyzed, and (ii) *preference network* that captures tweets, and comments liked by the users in past seven days. We have considered past seven days data as users' preferences may change over time. Both these features help in stance detection as it captures the users' perceptions and preferences (See fifth row of Table V). We introduce two novel techniques to enhance stance detection: graph convolutional networks (GCNs) to capture network features and a hierarchical attention mechanism to model tweet content.

Graph Convolutional Networks (GCNs) for Network Features: To effectively capture network features, we employ GCNs [9] to model the interaction network and preference network. GCNs provide a powerful way to incorporate local network information into the stance classification task. Let $G = (V, E)$ represent the network graph, where V is the set of nodes (users) and E is the set of edges (interactions or preferences). We first compute the adjacency matrix $A \in \mathbb{R}^{N \times N}$ of the graph G, where N is the number of nodes. Then, we normalize the adjacency matrix as $\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is the diagonal degree matrix of A. The graph convolution operation is defined as:

$$H^{(l+1)} = \sigma(\hat{A} H^{(l)} W^{(l)}) \tag{2}$$

where $H^{(l)} \in \mathbb{R}^{N \times F_l}$ is the feature matrix at the l-th layer, $W^{(l)}$ is the weight matrix, and $\sigma(\cdot)$ is an activation function (e.g., ReLU). By stacking multiple GCN layers, we can model higher-order network structures.

Hierarchical Attention Mechanism for Tweet Content: To better model the content of tweets, we propose a hierarchical attention mechanism. We first encode each word in a tweet using the pre-trained contextualized word embedding (ensemble of three encoders as discussed above). We then apply a bidirectional LSTM (Bi-LSTM) to capture contextual information:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t), \overleftarrow{h_t} = \overleftarrow{LSTM}(x_t) \tag{3}$$

We concatenate the forward and backward hidden states to obtain the final representation $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Next, we compute a word-level attention score a_t:

$$a_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}, \quad e_t = v_a^\top \tanh(W_a h_t + b_a) \tag{4}$$

where v_a, W_a, and b_a are trainable parameters. We obtain the tweet representation s by taking the weighted sum of word representations: $s = \sum_{t=1}^{T} a_t h_t$. Finally, we concatenate the GCN-based network feature representation and the hierarchical attention-based tweet representation to form a joint feature vector (z). This feature vector is then fed into a fully connected layer followed by a softmax function to classify the stance of the tweet into one of the three categories: "in favor", "against", and "neutral".

We train the model by minimizing the cross-entropy loss between the predicted probabilities and the ground truth stance labels. During training, we also employ a multi-task learning strategy to jointly optimize the model for stance detection, sentiment detection, and emotion detection, sharing the same underlying tweet and network feature representations. This enables the model to learn more informative representations, leading to improved performance in the stance classification task. We performed a user study to evaluate our system. Let us provide an example "I don't trust the COVID-19 vaccine; it was developed too quickly. "#antivax #vaccinesafety" In this case, our proposed method would first use the transformer-based pre-trained encoders (e.g., BERT, BERTweet, and COVID-Twitter-BERT) to obtain a textual representation of the tweet, capturing the semantic information of the tweet content. Next, our hierarchical attention mechanism weighs the importance of different parts of the tweet. In this example, it recognizes that the phrases "don't trust", "COVID-19 vaccine", and "developed too quickly" are more important for determining the stance. Additionally, the method incorporates Graph Convolutional Networks (GCNs) to capture network features from the user's interaction network and preference network. For example, if the user frequently interacts with others who share negative opinions about vaccines and often likes tweets containing misinformation about vaccine safety, these network features would provide strong cues about the user's stance against vaccination. After obtaining both the tweet representation and network features, the model concatenates these features to form a joint feature vector. This joint feature vector is then used to classify the tweet's stance as "in favor", "against", or "neutral." In this example, the model classifies the tweet as "against" due to the negative sentiment expressed in the text, the presence of the hashtag "#antivax", and the user's network features, which indicate a history of engaging with anti-vaccine content.

Sentiment detection We propose a fusion-model for sentiment analysis of COVID-19 vaccination-related tweets. The first layer of the model consists of four classification models: SVM, CNN, BiLSTM, and COVID-Twitter-BERT. Our goal in using two types of classifiers (classical and deep learning) is to enable the system to classify a wide range of test samples effectively. Some studies demonstrate that data samples in a

low confidence decision region of one classifier may be present in a high confidence decision region of another classifier [10].

We adopt a classical support vector machine combined with Bayesian probabilities [11] that employs a Naive Bayes log-count ratio to represent the word count feature of the model. The model implementation consists of three embedding layers and a sigmoid activation layer. The first embedding layer utilizes Naïve Bayes log-count ratios, while the second layer stores the learned coefficients (by SVM). The third layer incorporates context-specific knowledge, such as emoticons, emojis, and punctuation, to enhance the model's sentiment detection capabilities. A dot product operation generates the final prediction.

We apply a 1-D convolution with f filters to the input word-embedding matrix S. To extract n-gram features, we use different kernel sizes (c) on the word-embedding matrix at various granularities (individual sentence and tweet). The feature map is generated by sliding the filter across the entire text, producing output $FM_k \in \mathbb{R}^{(m-c+1)\times f}$. Max pooling is then applied to obtain a fixed-size vector, which is concatenated to form the final representation. The network's hidden layer is fully connected, and three softmax cells handle classification. The training hyperparameters include an ReLu activation function, an embedding dimension of 50, 150 filters, a kernel size of 4, a dropout rate of 0.2, and 150 neurons in the hidden layer. We use categorical cross-entropy as the loss function, followed by a dropout layer.

BiLSTM serves as another classifier in our fusion-based model, capturing both preceding and future context. We then employ an attention layer to measure the importance of various feature vectors. Our attention function f_{att} uses a dot product, and the representation is defined as follows: $r^{att} = \sum_{t=1}^{T} \frac{\exp(f_{att}(h_t, s_t))}{\sum_{i=1}^{T} \exp(f_{att}(h_t, s_t))} h_t$. The decoder input layer is replaced by the weighted representation (r_{att}). Lastly, a softmax layer produces the output labels (sentiment). The network is trained to minimize the cross-entropy loss of ground truth labels and predicted labels. The training parameters include an embedding dimension of 200, a dropout rate of 0.2, three neurons in the output layer, and an ReLu activation function.

Our final model is COVID-Twitter-BERT. We utilize the pre-trained COVID-Twitter-BERT (CT-BERT) v2 model from Hugging Face, trained on 160M tweets from January to July 2020. To train the meta learner, we must fuse the four base learners (SVM, CNN, BiLSTM, CT-BERT). We implement stacked generalization as the fusion method to assign different weights to the base learners' outputs. The fusion process is as follows: (i) The training dataset (TD) is divided into N equal folds; (ii) Each base learner is applied to all folds except one (TD^{-j}), generating a temporary prediction vector; (iii) A new training dataset (TD') is created by incorporating the temporary predictions, which is then used to train the meta learner. It is essential to note that for better efficacy, base learners should have lower classification errors. We select base learners with this consideration in mind. Finally, an iterative gradient boosting algorithm is employed to produce the ultimate fusion outcome.

C. L2 classifier: Tweet misinformation classes

L2 classifier categorizes the "against" and "negative" sentiment tweets into sixteen misinformation classes (M1-M15 and Other, See Table I) by following sub-modules:

1) Building knowledge-base from trusted sources: In this section, we introduce an automated knowledge extraction and augmentation method designed to reduce the reliance on a large volume of labeled tweets across various misinformation categories. It is important to note that the types of misinformation may change over time, rendering supervised training based on labeled tweets infeasible. To address this, we build our knowledge base by automatically scraping trusted sources (detailed in the supplemental document).

- We develop a sophisticated web crawler to extract information from websites, blogs, and news articles that specifically discuss COVID-19 vaccination misinformation and facts. Our knowledge base also incorporates diverse opinions, including vaccine dissenting and pro-vaccination webpages. To implement the crawler script, we utilize the *beautiful-soup4* Python library[5] for parsing HTML and XML data. The script searches for keywords such as "Misinformation", "Myths", "Truth", and "Fact" identifying blocks of text between two occurrences of these words. For parsing PDFs (as some sources contain PDFs), we employ the *Pytesseract*[6] Python library, an OCR tool. The crawler script automatically scrapes and generates "misinformation" and corresponding fact/true information dataframes without manual intervention, allowing for the flexible addition of sources during the development process.
- To date, we have scraped 80 sources, including webpages, blogs, and news articles, collecting 488 misinformation instances and their corresponding counter facts. However, due to the varied sources, the knowledge base contains repetitive data, rendering it redundant. To address this issue, we subsequently develop a semantic scoring mechanism and clustering approach to extract unique misinformation categories.

2) Clustering Techniques: We employ advanced clustering techniques to group similar types of misinformation based on semantic similarity scores. We adapt the pre-trained t-BERT 2020 model [12] for sentence embedding. A specialized variant of t-BERT (topic-informed BERT-based architecture) is utilized for pairwise semantic similarity detection, in which we incorporate two categories (misinformation and fact) into the architecture to infer similarity between both "misinformation" and "fact" within each class (M1-M15, see Table I). To cluster similar misinformation into the same categories, we implement Agglomerative Clustering on the similarity score matrix values. Each class contains similar misinformation and their corresponding scientifically-consensus-based counter facts. Through this method, we automatically extract 15 misinformation classes, as represented in Table I. We utilize T-BertSum [13] to summarize the corresponding facts for

[5]https://beautiful-soup-4.readthedocs.io/en/latest/

[6]https://pypi.org/project/pytesseract/

Observation & Insights
Initiated tweet (Against) → Conversation thread (majority Against)
21% of the dataset — Users support stating "negative" sentiment about their own vaccination experience
Initiated tweet (Against) → Conversation thread (majority In Favor)
26% of the dataset — Number of users participating in the thread more compared to the number of replies posted by one user — Users posting "positive" sentiment (vaccination experience) and current COVID trend
Initiated tweet (Misinformation) → Conversation thread (majority fact)
11% of the dataset — Major trend observed: Against and misinformation ($\approx 11\%$) → Sarcasm (Neutral+In favor) ($\approx 46\%$) → In favor ($\approx 43\%$)
Initiated tweet (Fact or true information) → Conversation thread (majority misinformation)
42% of the dataset — Number of tweets from specific users is more compared to the number of users in the thread — Several misinformation classes are discussed in the thread — Mentioned "external URLs" or providing references supporting misinformation for each tweet of fact — Majority of tweets ($\approx 76.7\%$) contain "mentions" of other users — Majority of tweets ($\approx 95.6\%$) mentioning misinformation topics show negative" sentiment about vaccination — Majority of the topics ($\approx 83\%$) include child vaccine, controversial substance, vaccine makes you sick"
Asking for information type tweets → Majority ($\approx$ 87%) replied with "negative" sentiment and misinformation topics of vaccination

TABLE II: Conversation sequence analysis

each misinformation class within the knowledge base. This summary can be used as a de-escalation strategy to counteract misinformation propagation and provide accurate information to vaccine-resistant individuals.

3) Tweet Classification of misinformation categories: We have constructed the word embedding vectors of misinformation and fact classes derived from the previous step. Each misinformation class has a word embedding vector obtained from fine-tuning *BERTopic* [14] embedding layer, namely em_vec. Contrary to document embedding using BERTopic, we feed all text data of each misinformation[7] in the pre-trained language model and extract topic-representations. We skip the second step of BERTopic which clusters the embeddings of the conventional document embedding, as our input is already clustered based on domain-specific (COVID vaccination) knowledge. CT-BERT V2 is used for L2 classifier, where we added two layers (layer 0, layer 1). Layer 0 of CT-BERT V2 is trained using em_vec which helps to augment coherent topic representations for each misinformation. Additional embedding layer (Layer 1) is deployed using labelled tweets (#50) of each category of misinformation (M1-M15) which helps in further fine-tuning the L2 classifier. We analyzed incorrect test samples from L2 classifier, and observed classification errors due to different factors as mentioned below:

- *Sarcasm/ irony (contributing ≈ 73% of the error samples)*: For example " I got my microchip …I mean my first dose of the Covid vaccine today. Have I turned into a zombie or vampire" [Model predicted it as "against" and misinformation class M8] "hope the covid vaccine alters my dna and I get to join the x men" [Model predicted it as "against" and misinformation class M7] "A new strain, more contagious …yet the same rushed vaccine will save you? Hurry up and get in line for your shot!!!" [Model predicted it as "in favour"]
- *Asking for information (contributing ≈ 21% of error samples):* User is requesting for more information for deciding regarding vaccination shot. For example: "I am cancer survivor. Is it unsafe for me to get the vaccine? Whether I am higher risk of developing serious sideeffects from the shot?" [Model predicted as "against" category and misinformation class M1] "My kids are turning 8 soon. Will more dosage mean better longer lasting immunity or severe sideeffects? Child COVID vaccine battle heats up in Sacramento. Is mandating it for all kids premature?" [Model predicted as "against" category and misinformation class M12]
- *Incomplete/ Out-of-context (Contributing ≈ 6% of the error samples):* This category includes either out-of-context tweet samples or incomplete tweets where proposed model fails to detect the context of the tweet. For example: "If the vaccine is to help with depopulation, what does the actual virus help with?" [Model predicted as "neutral"]

We propose a triple-attention based model for identifying sarcasm and refining the categories by considering above-mentioned error classes and enhancing the accuracy. It may be noted that existing approaches fail to identify such scenarios effectively: (a) Supervised technique where sarcasm detection model is trained using common texts from wiki and sarcastic similie does not work for our scenario due to discourse domain shift to COVID-19 and vaccination topics. (b) Hashtag based refinement does not work as the tweet samples do not contain specific hashtags such as $\#sarcasm$, $\#sarcastic$, or sentiment based $\#sad$, $\#excited$. (c) Rule based approach is not suitable either due to the requirement of large sarcasm-labeled corpus (on COVID vaccination). Our aim is to identify such sarcasm or irony from Twitter discourse with limited labelled data (COVID-19 vaccination). The triple-attention based layers are mentioned as follows:

- Layer 1: Topical Context - Some topics are more prone to sarcasm than others. For example, tweets about controversial topics like microchips, DNA changes, etc. are more likely to draw sarcasm than tweets about vaccine side effects. Here, we implemented *LDA* for topic modelling controversial topics and classifying sentiment of Tweets into "Positive", "Negative", "Sarcastic". This layer has a fully connected self-attention layer.
- Layer 2: Conversational Context - It refers to text in the con-

[7]After clustering, each misinformation class has several similar items obtained from different web-sources

versation of which the target tweet is a part. We considered "Re-tweet (original tweet stance analysis)" and "Replies in the thread" to understand the conversation context of the tweet. Target tweet and previous tweet in the conversation thread are analysed along with comments in thread structure. Further, a sequence labelling (positive, negative, sarcastic) of the tweets in the sequence is done to predict sarcasm in every text unit in the sequence.

- Linguistic Pattern Discovery: Sarcasm can be detected by the contrast between positive verbs and phrases indicating negative situations [e.g. "Oh sure! I support untested and unverified vaccine. Lord save the youth!"]. Here, we identify contexts that contain a positive sentiment contrasted with a negative situation [OR negative sentiment contrasted with a positive situation]. We have devised an iterative training step: Take "seed word" (e.g. support, save, rush) and sarcastic tweets and extracting phrases having contrasting polarity. This information is used to obtain embedding vector from different seeds.
- Features used: (i) Sentiment incongruities: The frequency with which a positive word is followed by a negative word and vice versa), (ii) Largest positive/negative subsequence: The length of the longest series of contiguous positive/negative words, and (iii) Pragmatic features: Existence of emoticons, laughter expressions, punctuation marks, ellipsis and capital words.

Using triple-attention layer, each of the tweets is classified as "sarcasm (yes)" or "sarcasm (no)". Next, a "negation function" is used on the output of L1 classifier, which means if sarcasm is detected and L1 predicted class is "against", then it is marked as "in favour", and vice-versa. If L1 classifier output is "neutral", then the sentiment of the tweet is verified, and assigned accordingly. Finally, "against" tweets are passed into L2 classifier for identifying misinformation classes. For identifying "asking for information" type of tweets (See section II-C3), we have used pragmatic feature of the tweets, namely, (a) identification of punctuation mark (?) and *wh-word*; (b) inspecting whether the polarity of the tweet as "neutral". The issue of the *incomplete/ out-of-context* tweets are resolved by "conversational context" layer of triple-attention model, and filtering tweets using "length" based constraints (we select tweets having atleast 50 length (characters) excluding URLs).

D. Analysis of Vaccine Dissenting Discourse

In our analysis of vaccine dissenting tweets, we investigate text patterns such as sentence types, determinants, special characters, and modifiers. We also consider text readability metrics such as word structure, average syllables per word, and sentence complexity. Additionally, we examine textual perception, informative opinions, and part-of-speech information. Capitalization features and word unigrams/bigrams are considered to cluster words used in similar contexts.

For vaccine dissenting user analysis, we focus on individual user features and communication-based features. Individual user features involve historical topics, profile information, historical sentiment, and interactional topics. Communication-based features consider the degree of interaction between two users and the rank of the addressee among the user's @-mention recipients.

To identify vaccine dissenting users, we employ Gradient Boosted Decision Trees (GBDT), an ensemble of decision trees that utilize the mentioned feature sets. The model is fitted in a forward step-wise manner to the current residuals of the decision nodes.

III. Performance Evaluation

The data was divided into training, validation, and test-sets following a 70:15:15 ratio. Stratified sampling was employed to maintain the distribution of classes consistent across the subsets, ensuring the model's exposure to all categories of misinformation. Notably, our dataset was significantly imbalanced, with certain misinformation categories being overrepresented. To mitigate this, we utilized the Synthetic Minority Over-sampling Technique (SMOTE) to balance the training set. The hyperparameters for our model were selected based on the best performance on the validation set.

Vaccination stance detection and misinformation classification Table IV demonstrates the effectiveness of our proposed model compared to the other classifiers considered. Among the traditional classifiers, SVM achieves the highest precision and F1-score. These results indicate that traditional machine learning classifiers can provide a baseline for sentiment analysis on vaccine dissenting tweets but may not be sufficient for capturing the complexity of the text data in this domain. Deep learning classifiers, such as LSTM and BiLSTM, show improved performance over traditional machine learning classifiers, with BiLSTM exhibiting higher precision and F1-score for both positive and negative sentiments. The pre-trained BERT models, $BERT_{BASE}$ and $BERT_{LARGE}$, further improve the performance of sentiment analysis on vaccine dissenting tweets. $BERT_{LARGE}$ outperforms $BERT_{BASE}$ in terms of precision and F1-score for both positive and negative sentiments. Our proposed model demonstrates the best performance among all classifiers, with a precision of 0.862 and an F1-score of 0.848 for positive sentiment, and a precision of 0.885 and an F1-score of 0.862 for negative sentiment. This improvement in performance can be attributed to the proposed fusion-based method using four classifiers. Additionally, our proposed model yields an improved accuracy of approximately 6% to 8% for emotion detection (anger, fear, joy, sadness) compared to the other classifiers.

Vaccination stance detection accuracy is reported along with an ablation study in Table V. Our model, which employs linguistic features alone, achieves competitive results compared to the state-of-the-art methods, such as Sentence BERT[15], Tahir et al.[16], and Feng et al.[17]. When integrating the network structure (L1+network) and topical information (L1+topical) into the L1 model, the results reveal a noticeable improvement in F1-score. This indicates that network and topic-related features provide valuable context that complements the linguistic features in identifying stances. The

TABLE III: Comparison of accuracy of L2 module with baselines for categorizing tweets into misinformation classes.

Model	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15
$BERT_{LARGE}$	78.2	74.6	80.5	82.7	78.03	75.81	82.08	80.24	84.10	81.04	80.10	80.09	82.17	80.94	78.16
COVID-Twitter-BERT	84.08	89.12	87.45	83.10	82.08	83.11	85.18	82.06	76.20	78.18	81.90	82.01	83.98	83.43	84.07
BERTweet-covid19-base-cased	83.02	85.11	80.83	83.18	81.23	83.02	84.16	83.07	86.15	83.04	81.05	82.7	81.44	81.73	82.19
Proposed Model	**92.48**	**90.04**	**88.02**	**89.16**	**88.04**	**89.90**	**92.01**	**90.05**	**87.18**	**89.03**	**91.45**	**88.02**	**87.94**	**91.80**	**88.43**

TABLE IV: Comparison on sentiment analysis classifier.

Classifier	Positive		Negative	
	Precision	F1-score	Precision	F1-score
SVM	0.68(±0.002)	0.65(±0.006)	0.624(±0.012)	0.608(±0.005)
Random Forest	0.67(±0.005)	0.642(±0.002)	0.619(±0.011)	0.582(±0.002)
KNN	0.545(±0.005)	0.528(±0.011)	0.491(±0.011)	0.462(±0.004)
XG Boost	0.660(±0.004)	0.643(±0.011)	0.601(±0.004)	0.570(±0.006)
Gaussian Naïve Bayes	0.562(±0.006)	0.541(±0.010)	0.510(±0.014)	0.508(±0.006)
AdaBoost	0.631(±0.005)	0.618(±0.014)	0.584(±0.002)	0.540(±0.010)
Perceptron	0.668(±0.010)	0.640(±0.006)	0.603(±0.010)	0.577(±0.005)
LSTM	0.725(±0.005)	0.713(±0.005)	0.709(±0.010)	0.701(±0.019)
BiLSTM	0.759(±0.023)	0.748(±0.045)	0.712(±0.012)	0.708(±0.016)
$BERT_{BASE}$	0.825(±0.002)	0.79(±0.012)	0.809(±0.0062)	0.77(±0.003)
$BERT_{LARGE}$	0.836(±0.017)	0.810(±0.005)	0.81(±0.011)	0.792(±0.016)
Proposed model∗	0.862(±0.006)	0.848(±0.003)	0.885(±0.020)	0.862(±0.010)

∗ Improved accuracy ≈ 6% to ≈ 8% for emotion detection (anger, fear, joy, sadness)

Model	Against			In Favour		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Sentence BERT[15]	81.89	74.08	77.78	78.04	69.91	73.75
Tahir et al. [16]	76.05	71.12	73.50	76.11	71.91	73.95
Feng et al. [17]	81.07	76.09	78.50	75.91	73.02	74.43
L1 (linguistic)	0.742	0.816	0.77	0.818	0.850	0.833
L1+network	0.765	0.847	0.8039	0.826	0.851	0.8383
L1+topical	0.78	0.848	0.8125	0.828	0.853	0.840
L1+conversational	0.793	0.851	0.8209	0.846	0.852	0.848
L1+Linguistic (Sarcasm)⋄	0.801	0.845	0.822	0.853	0.861	0.856
L1+triple attention (all)	0.886	0.854	0.869	0.87	0.864	0.866
L1+FULL	**0.914**	**0.8537**	**0.882**	**0.881**	**0.872**	**0.876**

⋄ Improved accuracy ≈ 7% compared to sarcasm detection method [18]

TABLE V: Comparison of Stance detection and Ablation study on L1 classifier module

incorporation of conversational features (L1+conversational) leads to a significant enhancement in both precision and recall, demonstrating that conversational context is crucial for understanding the nuances of stance detection in social media data. The addition of linguistic sarcasm detection (L1+Linguistic (Sarcasm)) further refines the model's performance, as it captures the subtleties of sarcastic language that can impact the accurate detection of stances. The best performance is achieved when combining all the components in the L1+FULL model. This model significantly outperforms the baseline and state-of-the-art methods in terms of precision, recall, and F1-score for both "Against" and "In Favour" stances. Table III demonstrates our framework's superior performance in misinformation classification compared to other BERT models across 15 misinformation classes. Our model effectively identifies and categorizes vaccine-related misconceptions and concerns, helping understand public sentiment and mitigate misinformation spread.

Our analysis reveals noticeable differences in behavior between vaccine dissenting and non-dissenting users on social media, with dissenting users having longer tweets, posting more vaccination-related content, and receiving higher retweet counts. Our user classifier achieves an F1-score of 0.896. Conversation analysis (Table II) highlights trends such as negative sentiment dominating personal vaccination experience discussions and misinformation-initiated tweets transitioning to sarcasm and eventually pro-vaccination sentiment.

Echo chamber analysis: Vaccination misinformation Our analysis explores the echo chamber [19] effect in the context of vaccination misinformation on social media. We identified five distinct user communities: (1) initiators spreading misinformation on vaccine controversial substances and trust issues (31%); (2) a mix of dissenting and neutral users propagating misinformation on natural immunity, side effects, and personal unsuitability (8%); (3) users supporting vaccine rejection due to child side-effects and fertility/pregnancy concerns (13%); (4) users spreading misinformation on political influence, corporate gains, and negative vaccination mandate impacts (5%); and (5) a community supporting vaccination, primarily sharing sarcastic tweets about misinformation (43%). Pro-vaccination users tend to interact within their community, while misinformation users often mention users from opposing communities. Quote tweets are used in 76% of cases to embed misinformation and change the original tweet's stance. Cross-exchanged tweets between pro- and anti-vaccination communities are mainly used for attacks and negative portrayals.

IV. Related Work

The identification of vaccination misinformation has been a focus of research, with studies showing the rapid spread of panic and the need to detect and respond to public sentiments and rumors on social media [20]. Analyses of COVID-19 misinformation have highlighted key topics that evoke misinformation, such as city lockdowns, cures and preventive measures, school reopening, and foreign countries [21]. Studies have also shown the negative impact of misinformation on people's vaccination intentions, particularly when misinformation appears scientific [22].

Automated misinformation detection methods typically rely on supervised classifiers, but they require a substantial number of labeled samples. In contrast, the proposed framework aims to automate knowledge extraction and management by building a knowledge base from trusted web sources and identifying misinformation categories using that knowledge base.

Vaccination sentiment and stance analysis have been explored in various studies [23]. Socio-economic factors were found to play a significant role in shaping public opinion towards vaccination [24]. Stance detection models have been trained using Reddit posts and COVID-19-Stance data, achieving high accuracy in classifying sentiment and stance [25]. Public opinion and sentiment analysis have been conducted

on various topics, including COVID-19 lockdown policies and government responses [26]. Stance detection systems have also been proposed to infer stances on vaccination from social media data [27]. Additionally, sentiment analysis has been performed on COVID-19 related tweets, analyzing sentiment distribution across different countries [28]. In contrast to existing works, our proposed module aims to identify sarcasm, humor, and irony in vaccination-related Twitter data along with network and interaction features of tweets and posts.

V. Conclusion

In this work, we show how to develop knowledge base and augment the knowledge to classify tweets into different misinformation classes proposing knowledge extraction and tweet discourse analytics modules. The framework is useful for efficient stance analysis towards vaccination, misinformation detection and integration of external knowledge (scientific facts about vaccination from trusted source) to paint a comprehensive picture of information extracted from social media data such as tweets. Our automated data analytics framework helps understand public opinion regarding COVID-19 vaccination and related misinformation topics.

Acknowledgements

We acknowledge partial support by seed funding by the Center for Social Data Analytics (C-SoDA) at The Pennsylvania State University, USA and the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003.

References

[1] S. Ghosh and P. Mitra, "Catching lies in the act: A framework for early misinformation detection on social media," in *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 2023, pp. 1–12.

[2] F. Pierri, B. Perry, M. R. DeVerna, K.-C. Yang, A. Flammini, F. Menczer, and J. Bryden, "The impact of online misinformation on us covid-19 vaccinations," *arXiv preprint arXiv:2104.10635*, 2021.

[3] S. Ghosh and P. Mitra, "How early can we detect? detecting misinformation on social media using user profiling and network characteristics," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 9 2023.

[4] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 9–14.

[5] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[6] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[7] C. Van Hee, E. Lefever, and V. Hoste, "Semeval-2018 task 3: Irony detection in english tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 39–50.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[9] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *International conference on machine learning*. PMLR, 2019, pp. 6861–6871.

[10] Y. Yao, "Three-way decisions with probabilistic rough sets," *Information sciences*, vol. 180, no. 3, pp. 341–353, 2010.

[11] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.

[12] N. Peinelt, D. Nguyen, and M. Liakata, "tbert: Topic models and bert joining forces for semantic similarity detection," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7047–7055.

[13] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "T-bertsum: Topic-aware text summarization based on bert," *IEEE Transactions on Computational Social Systems*, 2021.

[14] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[16] A. Tahir, L. Cheng, P. Sheth, and H. Liu, "Improving vaccine stance detection by combining online and offline data," 2022.

[17] F. Xie, Z. Zhang, X. Zhao, H. Wang, J. Zou, L. Tian, B. Zhou, and Y. Tan, "Adversarial learning-based stance classifier for covid-19-related health policies," in *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part IV*. Springer, 2023, pp. 239–249.

[18] D. Bamman and N. Smith, "Contextualized sarcasm detection on twitter," in *proceedings of the international AAAI conference on web and social media*, vol. 9, no. 1, 2015, pp. 574–577.

[19] A. Cossard, G. D. F. Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini, "Falling into the echo chamber: the italian vaccination debate on twitter," in *Proceedings of the International AAAI conference on web and social media*, vol. 14, 2020, pp. 130–140.

[20] S. Ghosh, P. Mitra, and P. Nakov, "Clock against chaos: Dynamic assessment and temporal intervention for reducing misinformation propagation," 2022.

[21] Y. Leng, Y. Zhai, S. Sun, Y. Wu, J. Selzer, S. Strover, H. Zhang, A. Chen, and Y. Ding, "Misinformation during the covid-19 outbreak in china: Cultural, social and political entanglements," *IEEE Transactions on Big Data*, vol. 7, no. 1, pp. 69–80, 2021.

[22] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[23] S. Ghosh, P. Mitra, and B. L. Hausman, "Evade: Exploring vaccine dissenting discourse on twitter," in *epiDAMIK 5.0: The 5th International workshop on Epidemiology meets Data Mining and Knowledge discovery at KDD 2022*, 2022.

[24] H. Lyu, J. Wang, W. Wu, V. Duong, X. Zhang, T. D. Dye, and J. Luo, "Social media study of public opinions on potential covid-19 vaccines: informing dissent, disparities, and dissemination," *Intelligent medicine*, 2021.

[25] K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea, "Stance detection in covid-19 tweets," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1596–1611.

[26] L. Miao, M. Last, and M. Litvak, "Tracking social media during the covid-19 pandemic: The case study of lockdown in new york state," *Expert Systems with Applications*, vol. 187, p. 115797, 2022.

[27] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, "Stance analysis of twitter users: the case of the vaccination topic in italy," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 131–139, 2020.

[28] S. Yu, S. He, Z. Cai, I. Lee, M. Naseriparsa, and F. Xia, "Exploring public sentiment during covid-19: A cross country analysis," *IEEE Transactions on Computational Social Systems*, 2022.