Ben-Gurion University of the Negev

The Faculty of Engineering

The Department of Software and Information Engineering

Beyond Text Analysis: Leveraging Network Structures and Natural Language Processing to Identify Drivers of Vaccine Hesitancy on Hebrew Twitter

Thesis submitted in partial fulfillment of the requirements

for the Master of Sciences degree

Omer Neu

Under the supervision of **Dr Oren Tsur**

June 2024

Ben-Gurion University of the Negev

The Faculty of Engineering

The Department of Software and Information Engineering

Beyond Text Analysis: Leveraging Network Structures and Natural Language Processing to Identify Drivers of Vaccine Hesitancy on Hebrew Twitter

Thesis submitted in partial fulfillment of the requirements

for the Master of Sciences degree

Omer Neu

Under the supervision of **Dr Oren Tsur**

Signature of student: ________________ Date: ________________

Signature of supervisor: ________________ Date: ________________

Signature of chairperson of the

committee for graduate studies: ________________ Date: ________________

June 2024

Beyond Text Analysis: Leveraging Network Structures and Natural Language Processing to Identify Drivers of Vaccine Hesitancy on Hebrew Twitter

Omer Neu

Master of Sciences Thesis

Ben-Gurion University of the Negev

2024

Abstract

New variants of the COVID-19, such as the highly infectious NJ.1 that started spreading late in December 2023, are often immune to existing vaccination. In addition, efforts to protect the population by achieving herd immunity, slowing the emergence of new variants, and potentially eradicating the disease are hindered by vaccine hesitancy and vaccine refusal. Social networks provide such "anti-vaxxers" an efficient vehicle for promoting vaccine hesitancy. This work uses a unique dataset that covers 80%–90% of the Hebrew tweets to study vaccine hesitancy during the first three years of the COVID-19 pandemic. We fine-tune an array of large language models tailored for morphologically rich languages to identify tweets expressing vaccine hesitancy rather than legitimate vaccine and COVID-related concerns, achieving an F score of 0.75. We further use these large language models, along with graph embeddings and diffusion models, to accurately identify users that actively promote vaccine hesitancy, achieving an F score of 0.87 on the user level in a challenging setting. We compare classification results achieved by the different approaches and discuss their advantages, their limitations, and the ways they allow approximate vaccine hesitancy trends in the wider network. The results indicate that, while text-based classifiers outperform graph-based approaches, they suffer peculiar false-positive errors, such as classifying medical experts as anti-vaxxers.

Acknowledgements

I sincerely thank Dr. Oren Tsur for the long support and mentoring for this research, Professor Meir Kalech for help in the last steps, Dr. Avraham Israeli for the assistance, and, last but not least, my mother Ronit for everything.

Contents

List of Figures **vi**

List of Tables **viii**

1 Introduction **1**

1.1 Vaccine Hesitancy 2

1.2 The Research 3

2 Related Work **5**

2.1 Vaccination and Vaccine Hesitancy in Israel 5

2.2 Social Media and Vaccine Hesitancy 7

3 Data **10**

3.1 Raw Datasets 10

3.2 Gold Labels and Matching Protocol 11

3.3 Data Annotation 12

3.4 The Social Graph 14

4 Computational Approach **16**

4.1 Tweet-Level Classification 16

4.2 User-Level Classification 16

5 Results **20**

5.1 Tweet Classification 20

5.2 User Classification 21

5.3 Network-Based Prediction 24

5.4 Social Graph 25

5.5 Error Analysis 28

6 Discussion and Conclusions **31**

6.1 Limitations 31

6.2 Ethical Considerations 33

6.3 Conclusions and Future Work 33

Appendix A **35**

Appendix B **38**

Bibliography **40**

List of Figures

2.1 Cumulative vaccination rate in Israel. 6

3.1 Distributions of three of the user's matching space parameters. 13

3.2 Illustration of node categories in the network. Seed nodes are in solid color (blue and red), added nodes are color-hashed and categorized into grey-shaded categories. 14

5.1 HeRo tweet classification of vaccine-hesitancy group and their twins. 21

5.2 HeRo precision and recall for δ_r of vaccine-hesitant and -compliant user classes. 22

5.3 Combined ego networks of the vaccine-hesitant group and their twins using the Fruchterman–Reingold layout. The big circle corresponds to the AVS and twin groups, the black diamond shape corresponds to the positive classified (HeRo) ego network. Blue points show twins and their ego network, red points shows AVS and their their ego network. 26

5.4 Confusion matrix of tweet classification for the HeRo model. . 28

B.1 Numbers correspond to tweet indices in Table 4.1. 39

List of Tables

3.1 Features and ranges for twin matching. Values of *Tweet Count*, *Follower Count*, and *Friend Count* should be in the range of 70%–130% with respect to an AVS account. *Date of Last Tweet* and *Registration Date* define a 90 day window. 11

4.1 A sample of tweets translated from Hebrew, their class, and the context explaining the (manual) annotation. Prediction was performed on the original Hebrew text using AlephBert-Gimel [1]. Original tweets (Hebrew) are available in Appendix A. Translation was done with Google Translate and refined manually as necessary. 19

5.1 Precision, recall, and F score for *tweet* classification. Each tweet was classified as either non-COVID (nc), COVID related (cr), or vaccine hesitant (vh). The macro average F score of the models is given in the rightmost column. 22

5.2 Macro averaged precision, recall, and F score for *user* classification using δ_n as a threshold where n is the minimum number of predicted vaccine-hesitant tweets to be classified as vaccine hesitant. 23

5.3 Macro averaged precision, recall, and F score for *user* classification using δ_r as a threshold, where r is the minimum ratio of predicted vaccine-hesitant tweets to non-vaccine-hesitant tweets of a user classified as vaccine hesitant. 23

5.4 Precision, recall, and F score resulting from applying our models to AVS and twins and with various minimum diffusion scores for the Degroot model. 24

5.5 User prediction of the vaccine-hesitant group and twin's ego network produced by a fine-tuned HeRo model. 25

5.6 Model prediction against our samples on the social graph . . . 27

5.7 Model results for known pro-vaccination activists on Twitter. Language model (HeRo), belief propagation (Degroot), and node2vec with multilayer preceptron (MLP). 30

A.1 36

A.2 37

Chapter 1

Introduction

Vaccination is one the major achievements of public health [2, 3]. Eradication of an infectious disease, near eradication, or the indirect protection of the population, commonly referred to as “herd immunity” are typically achieved through widespread vaccination campaigns that protect individuals who, for various reasons such as a compromised immune system, cannot be vaccinated. A worldwide vaccination campaign eradicated smallpox [4], and poliomyelitis (polio) was nearly eradicated but reemerged due to ineffective cessation protocols and the spread of vaccine hesitancy in various communities [5–7].

Vaccine hesitancy is the “delay in acceptance or refusal of vaccination despite the availability of vaccination services” [7]. The drivers of vaccine hesitancy vary for individuals and communities, reflecting an array of social, political, and religious factors, including distrust of governmental and medical institutions, cynicism toward the pharmaceutical industry, misguided beliefs linking vaccination and autism, and a strong belief in individual freedoms [6, 8].

The SARS-CoV-2 virus, commonly known as COVID-19, was first identified in China in December 2019 and quickly spread to Europe, the United States, and worldwide. Over 670 million cases were reported, with a death toll approaching seven million [9].[1] A number of vaccine platforms were approved by December 2020 and are credited with preventing over 19 million deaths in the first year of its introduction [10].

1.1 Vaccine Hesitancy

The unprecedented speed of development of the COVID-19 vaccines[2] and the relatively novel use of the mRNA molecule to achieve an immune response served as a fertile ground for the growth of vaccine hesitancy [12–15], in spite of the proven safety and efficiency of these vaccines [10, 16].

Some of the most ardent anti-vaccination proponents are popular media figures, celebrities, and online influencers such as (former) Fox News host Tucker Carlson, popular podcaster Joe Rogan, 2024 Presidential Candidate Robert F. Kennedy Jr., and NBA star Kyrie Irving[3]. Together, these public personalities have an audience of tens of millions[4]. Online activity and the exposure

[1] Last updated March 10, 2023; retrieved Dec. 31, 2023.

[2] The COVID vaccine was developed in less than a year, compared with about a decade for the polio vaccine and many decades for vaccinations against other diseases such as meningitis, whooping cough, and ebola [11].

[3] Full statements and dates are provided in the appendix.

[4] Tucker Carlson Tonight was one of the highest-rated programs in cable news, averaging 3.2 million viewers https://press.foxnews.com/2023/03/fox-news-channel-finishes-first-quarter-of-2023-as-top-network-in-all-of-cable-with-viewers-across-primetime-and-total-day. Rogan hosts the most popular podcast on Spotify with an estimated 11 million plus listeners per episode. https://time.com/6147548/spotify-joe-rogan-controversy-isnt-over/

to misinformation are associated with vaccine hesitancy. Fake news spreads further and faster than evidence-based news [17], and exposure to such misinformation significantly reduces the intent to vaccinate [12–14, 18, 19]. These findings were echoed by President Biden, who denounced social platforms for allowing the dissemination of COVID related misinformation:
"the only pandemic we have is among the unvaccinated, and that—and they [social platforms] are killing people." [5]

1.2 The Research

This research investigates vaccine hesitancy within the Hebrew-speaking population of Israel by analyzing the social media discourse on X (formerly Twitter) during the COVID-19 vaccination period. The research addresses the following questions:

1. How can state-of-the-art Hebrew language models be fine-tuned to distinguish between tweets promoting vaccine hesitancy and legitimate concerns about COVID-19 and vaccines?

2. How can a language model that identifies tweets promoting vaccine hesitancy be connected to a model that distinguishes a *user* that promotes vaccine hesitancy?

3. Can the social network structure on X be leveraged to identify and

Time magazine (retrieved Jan. 14, 2024).

[5] https://www.nytimes.com/2021/07/16/us/politics/biden-facebook-social-media-covid.html ”Biden Denounces Social Media for Virus Disinformation” (NYT, July 16, 2021; retrieved June 30, 2024)

classify vaccine-hesitant users?

Chapter 2

Related Work

2.1 Vaccination and Vaccine Hesitancy in Israel

As mandated by the state health insurance law, all Israeli residents are registered with an approved health maintenance organization. While vaccination decisions are a matter of individual choice, vaccinations are encouraged and offered free of charge in community clinics and schools (with parental consent). As a result, 98% of Israeli children are vaccinated by the protocol recommended by the Ministry of Health (DTaP-Hib-IPV,MMRV, PCV, Hepatitis A/B, Rota) [20]. The national medical coverage sponsored by the Israeli state contributed to the success of the campaign for COVID vaccinations [21], with a vaccination rate (first dose) exceeding 87% among the

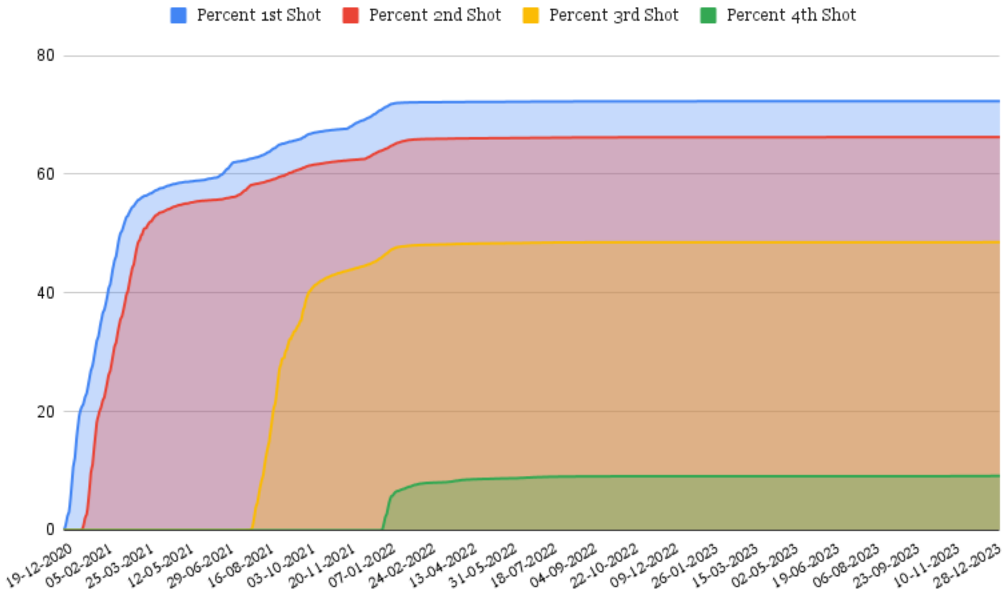


Figure 2.1: Cumulative vaccination rate in Israel.

population aged 15 years and older. [1] Figure 2.1 present the accumulative vaccination percentage by dosage.

In spite of the high vaccination rate, some individuals refuse to vaccinate and promote vaccine hesitancy. The main factors associated with vaccine hesitancy in Israel are similar to those found in other countries: level of education, religious belief, ethical stance, safety concerns, and misgivings regarding its effectiveness [22–26]. Reservations regarding the COVID vaccine stem from similar causes [12, 27].

[1] Data from the COVID Dashboard https://datadashboard.health.gov.il/portal/dashboard/corona of the Ministry of Health. Accessed Dec. 31, 2023.

2.2 Social Media and Vaccine Hesitancy

Social media platforms serve both as a vehicle for disseminating anti-vaccination ideology and as a tool for qualitatively, quantitatively, and computationally studying the phenomenon of vaccine hesitancy [28–33]. Information exposure on Twitter explains differences in vaccination rates between individuals of similar demographics [34]. Vaccine refusal on Facebook was found to be promoted as a civil right and used for political mobilization [35]. Specifically, rejection of the COVID vaccine among American Twitter users correlated with media use [36]. The narratives promoted on Facebook pages of deceased victims of COVID-19 was found to be overwhelming political and anti-government [37]. The global and multilingual effects of vaccine misinformation on Twitter was studied by Lenti et al. [38], who surveyed 200 000 Facebook users to study the different drivers of vaccination refusal and the barriers to higher vaccine coverage [15].

The drivers of vaccine hesitancy on Twitter were analyzed during the measles outbreaks that preceded the COVID-19 pandemic. Misinformation about vaccines causing autism had already been unearthed using social network analysis and semantic network of co-occurrence of words [39]. In addition, narrative analysis [40] was used to map the themes of the vaccination debate on Dutch Twitter [41] using the conversational context (i.e., reply chain) and showed that the Dutch vaccine hesitancy community describes themselves as conservative. In other work, Broniatowski et al. [42] gathered URLs from pro- and anti-vaccination groups on Facebook to analyze the monetization strategies of the websites linked to by those groups. Vaccine hesitancy was also classified prepandemic on Twitter using a dataset based on keywords [43–45].

At the outbreak of COVID-19 in the US, a bipartisan political consensus existed regarding the measures that should be taken, whereas extremists on social networks were already divided. This exemplified the false polarization thesis of social networks, as discussed by Bail [46]. Later, the partisan divide appeared in a Pew research survey [47], which showed that, in December 2020, Republicans were less willing to get the COVID-19 vaccine than Democrats. By comparing "partisan" media consumption (CNN, MSNBC versus FoxNews, Newsmax) Green et al. [36] found that both Democrats and Republican with strong partisan identities tended to believe more in COVID-19 misinformation than those with weak partisan identities [48]. In addition, a partisan divide appeared during the measles outbreak that preceded the COVID-19 pandemic; when a few democratic states were more vaccine hesitant than republican states [49]. These results suggests that a partisan divide regarding vaccine hesitancy may occur in Israel as well.

Manual annotation of the themes of vaccine hesitancy tweets in Turkey [50] showed that the main themes were "poor scientific process" and "conspiracy." Lenti et al. [38] used COVID-19 keywords and community detection on the retweet network and co-sharing network (a network where edges represent sharing the same URL address) on Twitter. They manually annotated tweets within the communities and found that the number of vaccine-hesitant people on Twitter increased during the COVID-19 pandemic. They also found that vaccine-hesitant people were less isolated on Twitter than they were before the COVID-19 pandemic. Research on deceased anti-vaxxers used semantic clustering on their posts to analyze the shifting narrative anti-vaxxers [37]. A sophisticated model used on American Twitter included a morality analysis, topic modeling, and vaccine-hesitancy detection to display the motivations for vaccine hesitancy [51].

In Israel, the majority of the population applies pediatric vaccinations according to the recommended protocol, although a minority go against the standard vaccination program (a self-reporting study [52] shows ≈9% deviate from the pediatric vaccination protocol). For the annual influenza vaccine only ≈15% of the population take the vaccination, whereas that number almost doubled during COVID-19 [53]. A small study in Israel [54] ($n = 70$) showed that 15.7% of the participating parents did not vaccinate their children according to the pediatric protocol. The study also found that a parent in a Facebook group related to parenting (local parenting groups) is more likely to vaccinate their children following the pediatric protocol.

Chapter 3

Data

3.1 Raw Datasets

We collected 80%–90% of the public Hebrew tweets from January 2019 to March 2023.[1] The raw dataset contains ≈254 million tweets posted by ≈3 million users. However, many users are relatively inactive or not "natives" in the Hebrew sphere. Only ≈1.1 million users tweeted more than three tweets in the data (a total of ≈196 million tweets) and only ≈95 000 users tweeted more than 100 tweets (≈185 million tweets). About 20% of the tweets in the data are retweets.

[1] We used Twitter's streaming API, tracking a list of Hebrew stopwords. We established the 80%–90% margin by extrapolating the coverage achieved by the term tracked over the Hebrew tweets collected using the Decahose, providing a random sample of 10% of the public stream.

3.2 Gold Labels and Matching Protocol

We manually identified 384 users that actively and openly promoted anti-vaccination.[2] We refer to this group of users as the anti-vaxx-seed (AVS).

The vast majority of the Israeli population does not reject vaccination (see Section 2). We therefore assume that, when sampled at random, most Israeli Twitter users are not vaccine hesitant. However, as reported in Section 3.1, randomly sampled accounts are likely to be inauthentic or only minimally active. To create a dataset that supports a nontrivial classification task, we use stratified sampling, which requires that the sampled accounts share a number of features with AVS accounts: total number of tweets, number or friends [3] and followers, the date of most recent activity, and the account age. Table 3.1 provides the features and similarity ranges.

Parameter	Matching space
Tweets (T)	$0.7T_H \leq T \leq 1.3T_H$
Followers (FL)	$0.7FL_H \leq FL \leq 1.3FL_H$
Friends (FR)	$0.7FR_H \leq FR \leq 1.3FR_H$
Last tweet date (LT)	$LT_H - 90_d \leq LT \leq LT_H + 90_d$
Registration date (RD)	$RD_H - 90_d \leq RD \leq RD_H + 90_d$

Table 3.1: Features and ranges for twin matching. Values of *Tweet Count*, *Follower Count*, and *Friend Count* should be in the range of 70%–130% with respect to an AVS account. *Date of Last Tweet* and *Registration Date* define a 90 day window.

We refer to accounts u and v as *twins* if they have similar characteristics (as defined above). We sampled four twins for each $u \in$ AVS and created a second dataset denoted "twins." Figure 3.1 shows the difference in the distributions

[2] Either general vaccine refusal, COVID-specific vaccine refusal and hesitancy, or denying the severity of the COVID pandemic ("seasonal flu with good PR").

[3] The number of friends is the number of people a user is following.

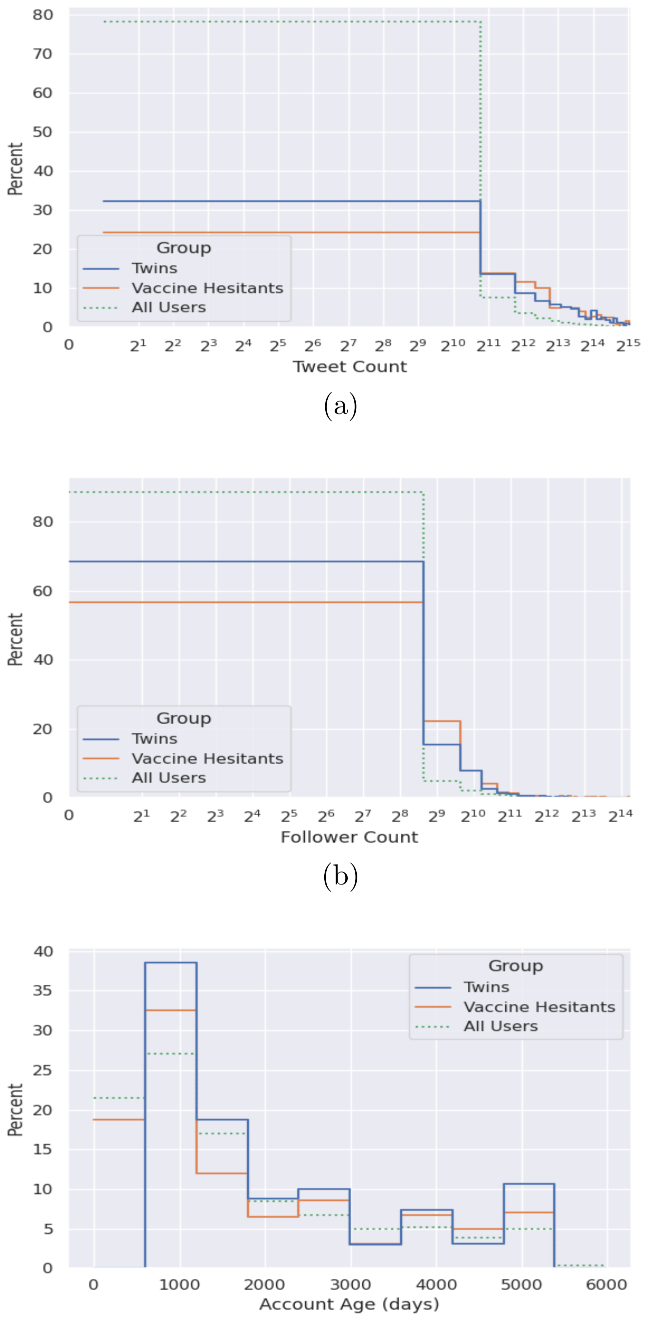
of *Tweets Count*, *Follower Count*, and *Account Age* for the general population (raw data), AVS, and twin. The AVS and twin accounts follow a significantly different distribution than the general-population accounts.

3.3 Data Annotation

COVID denial and an anti-vaccination stance can be conveyed without explicitly using words such as "COVID" or "vaccine." . Conversely, users can discuss novel vaccines, raise legitimate concerns, or report side effects without rejecting vaccination. See Table 4.1 for examples. Thus, tweets in our data could fall under one of the following three classes: (i) unrelated to COVID, (ii) COVID related but not promoting anti-vaccination, and (iii) promoting anti-vaccination.

A sample of 4000 tweets posted between January and June 2021 by AVS and twin users was manually labeled under the three classes given above. These tweets were sampled in a stratified manner corresponding to the number of tweets posted on each account. We defined a list of COVID-related keywords (e.g., COVID, virus, vaccine, Pfizer) and sampled both AVS and twin accounts for tweets containing these keywords *and* tweets not matching these keywords in a 1 : 1 ratio. The first six months of 2021 were sampled because this period includes the introduction of the COVID vaccine (January 2021) and the administration of the first booster shot (see Fig. 2.1), which means that the safety and effectiveness of the vaccine was being debated.

Unlike tweets, users were classified into two classes: (i) anti-vaxxers promot-



(a)

(b)

(c)

Figure 3.1: Distributions of three of the user's matching space parameters.

ing anti-vaccination sentiment (AV) and (ii) vaccine-compatible (VC). By definition, all users in AVS were labeled AV. Based on the reported vaccination rate, we classified all users in the twin set as VC, although a small subset of these uses may have been mislabeled. This approach produces various effects on the classification results obtained by textual classifiers (large language models) and network-based classifiers (node2vec, diffusion processes). We discuss these effects in the Results chapter.

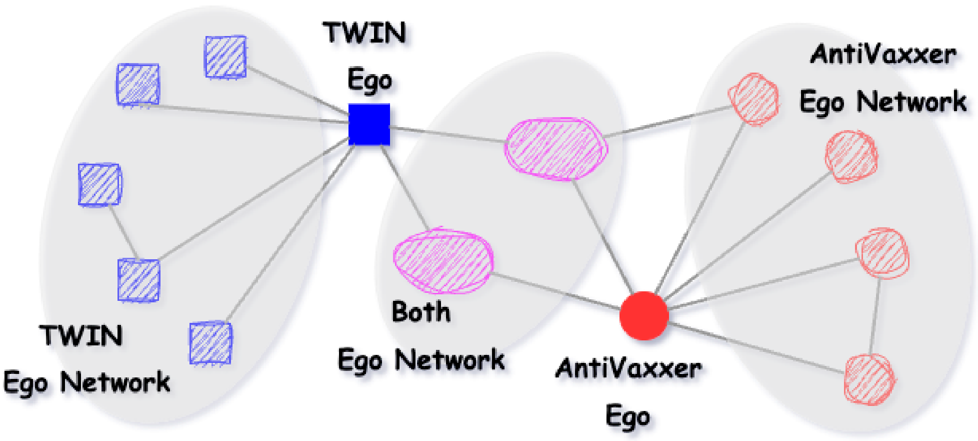


Figure 3.2: Illustration of node categories in the network. Seed nodes are in solid color (blue and red), added nodes are color-hashed and categorized into grey-shaded categories.

3.4 The Social Graph

One unique aspect of our dataset is its comprehensive coverage of the Hebrew tweets throughout the COVID years. This coverage allows us to recover the network of engagement between the users, especially between AVS and their twins. To this end we created a social network in the following manner: For each $u \in AVS \bigcap TWINS$ we first retrieve her ego network G^u. A directed edge $\overrightarrow{uv}$ is established if u retweeted v at least $m = 2$ times, and a different edge direction is established if v retweeted u at least $m = 2$ times. In the second stage we recover the edges between all the nodes in $\{G^u\}_{u \in AVS \bigcap TWINS}$. For simplicity, we only use the largest weakly connected component, denoted G^U. For analytical purposes we categorize the nodes in G^U into five categories, as illustrated in Fig. 3.2: the AVS nodes and twin nodes (egos), nodes added for being in an ego network of an anti-vaxx ego or in a twin ego (but not in an ego of both), and nodes added for being in the ego network of at least one anti-vaxx ego *and* at least one twin ego.

Chapter 4

Computational Approach

4.1 Tweet-Level Classification

Processing Hebrew texts is challenging, mainly due to the rich morphology, the ambiguity caused by the omission of diacritics ("niqqud"), and the relatively limited volume of data. The following models were developed to analyze the ambiguity and the morphological complexity: AlephBert [55], AlephBErtGimmel [1], HeBert [56], and HeRo [57]. We fine-tuned these four models on a random sample of 85% annotated tweets.

4.2 User-Level Classification

Classifying users as anti-vaxxers can be challenging due to varying degrees of opaqueness, obsessiveness, type of rejection (all vaccines, only COVID vaccines, COVID denial, etc.), and the natural mixture of topics in the users stream. We explore both text- and network-based approaches for user classification.

Consider an accurate model (or oracle) M for the classification of tweets: $M(t) = 1$ if the text t is classified as anti-vaccination, and $M(t) = 0$ otherwise. Given a set of tweets $T^u = \{t^u_1, \ldots, t^u_n\}$ posted by user u and some threshold value δ, we define the classification function as follows:

$$\theta(u) = \begin{cases} C^+, & \phi(u) \geq \delta \\ C^-, & \text{otherwise.} \end{cases} \tag{4.1}$$

We consider two types of thresholds: fixed and relative. The fixed threshold is a scalar $\delta \in \mathbb{N}^+$ and

$$\phi(u) = \sum_{i=1}^{|T^u|} M(t^u_i).$$

For the relative threshold, $\delta \in [0, 1]$ and

$$\phi(u) = \frac{\sum_{i=1}^{|T^u|} M(t^u_i)}{|T^u|}.$$

For convenience we denote a fixed threshold δ_n and a relative threshold δ_r. That is, assume $\delta_n = 10$ and $\delta_r = 0.2$ and consider a user u for which $T^u = \{t^u_1, \ldots, t^u_{100}\}$. Furthermore consider that $M(t^u_i) = 1$ for $i \leq 10$, and $M(t^u_i) = 0$ for $i > 10$. Using the fixed threshold u is classified as anti-vaccination because $\phi(T^u) = 10 \geq \delta_n$. However, using the relative threshold, u will be not classified as anti-vaccination because $\phi(T^u) = 0.1 < \delta_r$.

The hyperparameters $\delta_{n,r}$ are learned by applying a grid search to a development set.

Assuming we are provided not only with a set of users and their tweets but also with the social relations between the users, encoded as a social graph G, we can leverage the structure to classify the tweets. Moreover, recall that previous work demonstrated the importance of the network in the dissemination of fake news or in the development of communities of like-minded individuals (see Chapter 2). We therefore consider two inherently different network-based approaches:

1. *Classification by node embedding.* Node2vec is a powerful algorithm for learning node embeddings based on a biased random walk [58]. The learned representations enable high-performance node classification tasks over large networks [59]. Given a social graph G, we first learn the node2vec embeddings for all nodes in the graph and then use these representation as feature vectors in a simple classifier. Specifically, we insert a standard multilayer perceptron network into ϕ in Eq. (4.1) and set $\delta = 0.5$, reflecting the C^+ class likelihood.

2. *Belief propagation.* Online hate mongers could be detected by using a diffusion model based on DeGroot's process [60, 61]. A modified version of this process, in which the score of some subset of nodes (the seed) is fixed, performs better in a setting very similar to the one at hand [62]. Since the AVS users are manually identified as anti-vaxxers we follow Israeli and Tsur, fixing the "diffusion score" $DS(u) = 1 \forall\ u \in AVS$ and applying the diffusion process to compute the diffusion score of all other nodes. After the diffusion process converges, the classification is

decided by using $\phi(u) = DS(u)$ in Eq. (4.1). We follow previous work and set $\delta = 0.75$ [i.e., a node for which $DS(u)$ is in the forth quartile] as the threshold for classification as an anti-vaxxer (C^+). W also tested smaller values of δ.

	Text	Label	Prediction	Comment
1	Received after contacting the management. And now she is waiting for the second vaccine, because the first one, as mentioned, does not really protect yet. In my opinion it is an unparalleled audacity to send a teacher to the battlefield before we have been immunized. The sheep's silence of Ran Erez, chairman of the secondary teachers' organization, is puzzling to me.	Covid-Related	Covid-Related	'The sheep's silence' is a translation to 'The Silence of the Lambs'. The tweet is about the delays of vaccinating teachers while teaching in front on yet-to-be vaccinated pupils.
2	You are the one who talked during the past year about a thousand dying every winter from flu-like diseases. So now you update the number to 6000?	Vaccine-Hesitancy	Covid-Related	The lack of context hinder the task of classifying tweets, it could be a "just asking questions" tactic toward a public health official or toward anti-vaxxer (this tweet was replay to a Covid-Skeptic) ⌈☞ **A reply to Dr Lass** –*omer* ⌋
3	I totally continue with the mask. Not ready for another round of this shit.	Covid-Related	Covid-Related	
4	An emergency call from a brave scientist, an expert in virology and vaccines who desperately turns to medical leaders and policy makers: "You will cause a catastrophe on a large scale due to the 'vaccinate all' policy. A brave call that resonates in the shocking darkness of conformism, cowardice, and just plain ignorance of doctors who give a hand to a reckless and dangerous experiment!!"	Vaccine Hesitancy	Vaccine Hesitancy	
5	Are some of the hospitalized in serious condition, after two vaccinations? known?	Covid-Related	Covid-Related	The intentions of the text are indeterminate, it could be genuine concern or "just asking questions" tactic to raise skepticism of the vaccines.
6	Chibuta barely played this year—it will take him a long time to get back into playing shape. What's more, most of their players who went to Europe came back much worse. Ali Muhammad has not recovered since he got Corona (this also takes time) Hamed, in my opinion, will not come, and here we are also talking about an older player who I am not sure can give more than a good season	Non-Covid	Non-Covid	Sport related tweet mentioning a COVID-19 ill player.
7	Am I in favor of forced vaccinations? I am???? No way? Simply unvaccinated will not go to shopping malls and restaurants and concerts, will not meet their families (not yours, Mr. Zeliger) and sit at home under full curfew.That's how I am, enlightened and liberal. A true democrat.	Vaccine Hesitancy	Vaccine Hesitancy	A highly sarcastic reply that ignores the risk of unvaccinated population to the public without the nuance toward those can't be vaccinated.
8	The spread of the virus of anarchy is a fact	Non-Covid	Covid-Related	
9	And your incitement against ultra-Orthodox during the entire Corona period, why did they make you a partner? A friend asks	Non-Covid	Covid-Related	Using the COVID-19 as a time period.
10	Warning!!!! Vaccines will destroy the genetic makeup of your mascara	Covid-Related	Vaccine-Hesitancy	A known activist against anti-vaccination. Tweet accompanied with edited picture mocking anti-vaccination.

Table 4.1: A sample of tweets translated from Hebrew, their class, and the context explaining the (manual) annotation. Prediction was performed on the original Hebrew text using AlephBertGimel [1]. Original tweets (Hebrew) are available in Appendix A. Translation was done with Google Translate and refined manually as necessary.

Chapter 5

Results

5.1 Tweet Classification

With a dataset of 4000 tweets manually annotated into three classes, we trained baseline algorithms and fine-tuned a transformer model to classify tweets. Table 5.1 shows that AlephBertGimmel achieves the best overall results for classifying tweets (F score = 0.75). Logistic regression on the bag-of-words vector representation achieves an F score of 0.5.

We used HeRo to predict unannotated tweets by vaccine-hesitant people and their twins. The results show that, for both vaccine-hesitant and vaccine-compatible people, the majority of the tweets were not about COVID-19. This finding is consistent with our previous findings, which were based on our annotated data (see Fig. 5.1).

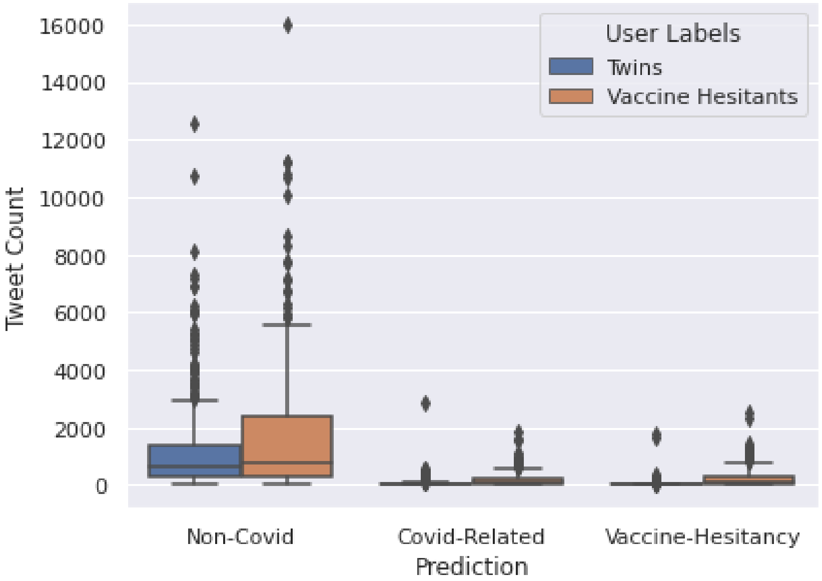


Figure 5.1: HeRo tweet classification of vaccine-hesitancy group and their twins.

5.2 User Classification

Using the fine-tuned transformer models on classifying tweets we built *user* classifier models by classifying user tweets during the initial COVID-19 vaccination period and evaluating the thresholds or ratio of vaccine-hesitancy tweets to classify a user as either vaccine hesitant or vaccine compliant. Applying the user classification procedure to over 4000 unlabeled tweets from the vaccine-hesitant group allowed us to evaluate the classification accuracy based on the *user label* as opposed to the tweet label (i.e., AVS or twin). In Fig. 5.2, a higher δ_r causes the model to lose recall on the vaccine-hesitant group because most people in the vaccine hesitant group dedicate less than 5% of their tweets to vaccines, even during the initial COVID-19 vaccination period.

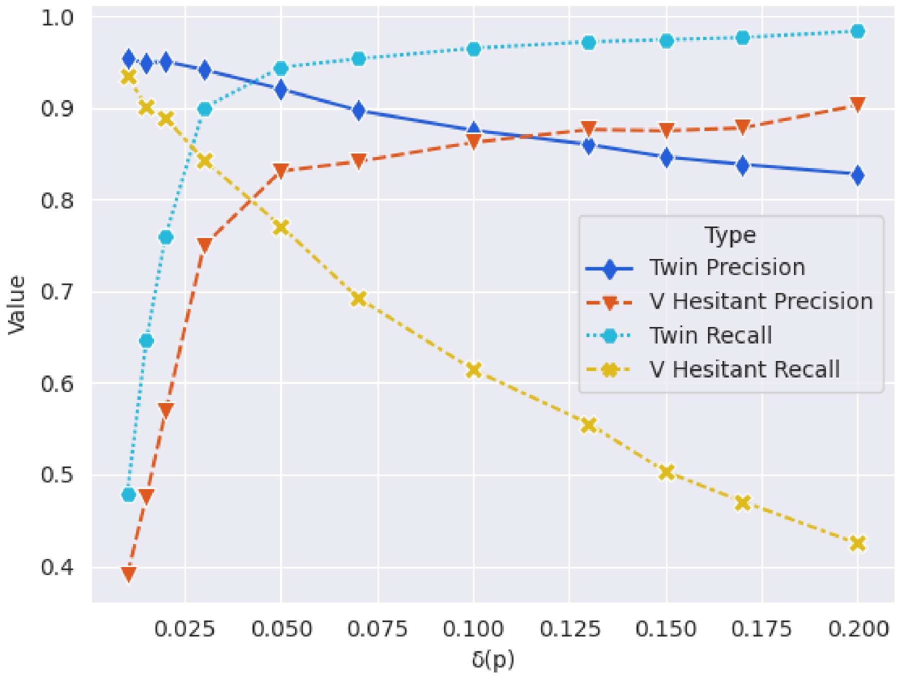


Figure 5.2: HeRo precision and recall for δ_r of vaccine-hesitant and -compliant user classes.

	P_{nc}	R_{nc}	FS_{nc}	P_{cr}	R_{cr}	FS_{cr}	P_{vh}	R_{vh}	FS_{vh}	FS_{avg}
Logistic+BOW	0.89	0.97	0.92	0.28	0.16	0.20	0.54	0.29	0.38	0.50
Logis+Tf-Idf	0.86	0.99	0.92	0.43	0.10	0.16	0.83	0.23	0.36	0.48
SVM+BOW	0.93	0.58	0.71	0.09	0.74	0.16	0.56	0.08	0.14	0.33
SVM+Tf-Idf	0.00	0.00	0.00	0.06	1.00	0.11	0.00	0.00	0.00	0.04
AlephBert	0.95	0.85	0.89	0.59	0.71	0.64	0.59	0.59	0.59	0.71
AlephBertG	0.96	0.88	0.92	0.65	0.70	0.68	0.63	0.71	0.67	**0.75**
HeBert	0.92	0.83	0.88	0.51	0.63	0.56	0.57	0.53	0.55	0.66
HeRo	0.93	0.84	0.88	0.56	0.64	0.60	0.59	0.63	0.61	0.70

Table 5.1: Precision, recall, and F score for *tweet* classification. Each tweet was classified as either non-COVID (nc), COVID related (cr), or vaccine hesitant (vh). The macro average F score of the models is given in the rightmost column.

Tables 5.2 and 5.3 show the results when δ_n is the minimal number and when δ_r is the minimal ratio, respectively, of explicit vaccine-hesitant tweets

	$\delta_n = 1$			$\delta_n = 5$			$\delta_n = 10$			$\delta_n = 20$			$\delta_n = 50$		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
AlephBert	0.65	0.60	0.41	0.69	0.73	0.63	0.73	0.79	0.73	0.78	0.81	0.79	0.87	0.77	0.80
AlephBertG	0.66	0.64	0.47	0.71	0.77	0.70	0.76	0.81	0.77	0.82	0.83	**0.83**	0.87	0.76	0.80
HeBert	0.63	0.55	0.33	0.69	0.73	0.61	0.71	0.77	0.69	0.76	0.81	0.78	0.84	0.76	0.79
HeRo	0.64	0.57	0.35	0.67	0.71	0.59	0.71	0.77	0.69	0.77	0.81	0.78	0.86	0.78	0.81

Table 5.2: Macro averaged precision, recall, and F score for *user* classification using δ_n as a threshold where n is the minimum number of predicted vaccine-hesitant tweets to be classified as vaccine hesitant.

	$\delta_r = 0.01$			$\delta_r = 0.02$			$\delta_r = 0.05$			$\delta_r = 0.1$			$\delta_r = 0.15$		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
AlephBert	0.71	0.76	0.66	0.78	0.84	0.79	0.89	0.86	**0.87**	0.87	0.78	0.81	0.86	0.73	0.76
AlephBertG	0.75	0.82	0.75	0.82	0.85	0.83	0.87	0.84	0.85	0.87	0.79	0.82	0.86	0.74	0.77
HeBert	0.69	0.72	0.60	0.76	0.83	0.77	0.87	0.85	0.86	0.88	0.79	0.82	0.86	0.73	0.76
Hero	0.67	0.71	0.59	0.76	0.82	0.77	0.88	0.86	**0.87**	0.87	0.79	0.82	0.86	0.74	0.77

Table 5.3: Macro averaged precision, recall, and F score for *user* classification using δ_r as a threshold, where r is the minimum ratio of predicted vaccine-hesitant tweets to non-vaccine-hesitant tweets of a user classified as vaccine hesitant.

that cause a user to be classified as vaccine hesitant. With $\delta_r = 0.5$, both AlephBert and HeRo produced the highest F score of 0.87.

5.3 Network-Based Prediction

	P	R	F
AlephBertG ($\delta_n = 20$)	0.82	0.83	0.83
HeRo ($\delta_r = 0.05$)	0.88	0.86	*0.87*
node2vec+MLP	0.76	0.76	0.76
Degroot 0.75	0.81	0.52	0.51
Degroot 0.5	0.87	0.62	0.66
Degroot 0.3	0.86	0.70	0.74
Degroot 0.2	0.87	0.73	0.78
Degroot 0.1	0.86	0.76	0.8
Degroot 0.05	0.84	0.79	*0.81*
Degroot 0.03	0.82	0.79	0.8

Table 5.4: Precision, recall, and F score resulting from applying our models to AVS and twins and with various minimum diffusion scores for the Degroot model.

On this network we used belief propagation from Degroot and classification with node embedding (see Section 4.2). Applying the network-based model to the ego-network showed that the AVS and their twins are correctly classified when ignoring the unlabeled nodes on the ego networks. For the Degroot model we split the AVS into training and test sets with the test set containing 40% of the AVS in the ego network. Table 5.4 summarizes the best textual models with the network-based models shown for comparison. The best diffusion score for the Degroot model applied to the AVS and twins on the ego network is for $\delta = 0.05$.

5.4 Social Graph

Figure 5.3 plots the combined ego network in the Fruchterman–Reingold layout [63], which is used for informative displays of social networks [64, 65]. The figure shows the ego network of the AVS and twin groups. The AVS ego clusters across the center and south east portion of the network, and their positively classified nodes are mostly clustered in the south-east. We ran the fined-tuned HeRo, Degroot, and node2vec+MLP models on the ego-network to determine how many points in the AVS cluster, the twin cluster, or both clusters (cf. Fig. 3.2) are positively classified as vaccine hesitant.

Model		$AVS_{textego}$	$Twins_{textego}$	$Both_{textego}$
HeRo	C^+	563	231	131
	C^-	1358	2916	1854
DeGroot 0.75	C^+	425	0	9
	C^-	1496	3101	1976
DeGroot 0.05	C^+	952	17	176
	C^-	969	3084	1809
node2vec	C^+	864	206	543
	C^-	1057	2895	1442

Table 5.5: User prediction of the vaccine-hesitant group and twin's ego network produced by a fine-tuned HeRo model.

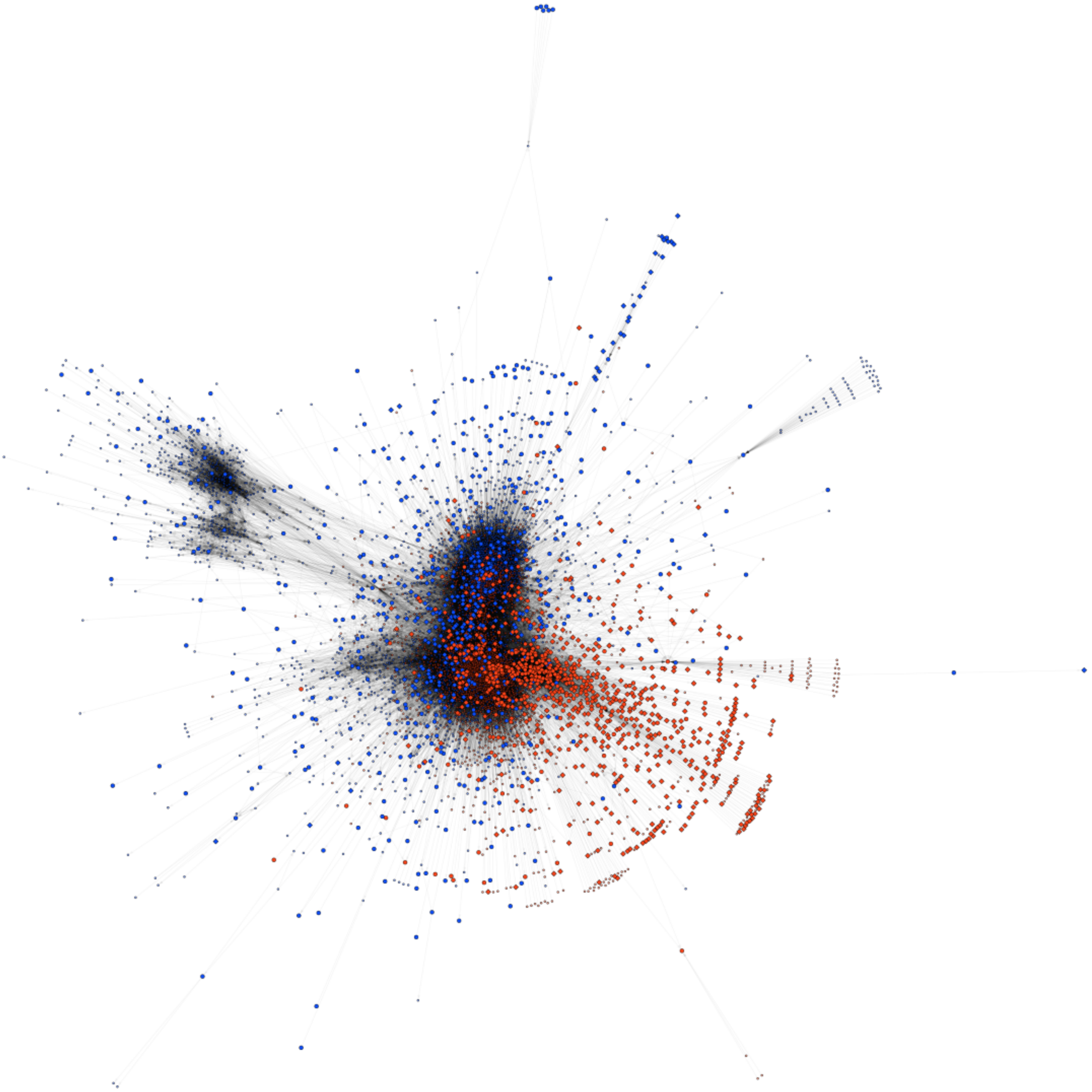


Figure 5.3: Combined ego networks of the vaccine-hesitant group and their twins using the Fruchterman–Reingold layout. The big circle corresponds to the AVS and twin groups, the black diamond shape corresponds to the positive classified (HeRo) ego network. Blue points show twins and their ego network, red points shows AVS and their their ego network.

By leveraging the social graph of our models, we can reach a wider user base than when relying solely on AVS and their twins. This broader reach is reflected in Table 5.5, where our model suggests a higher prevalence of vaccine hesitancy among users in the AVS ego network compared with users in the twin or the overlapping ego networks. To validate this finding, we sampled users from the social graph and compared the model's predictions with our evaluation results.

To obtain a balanced sample of positive and negative predictions, we employed stratified sampling based on the output of each model. For each user in the sample, we retrieved tweets predicted by the HeRo model as positive ($C+$). These $C+$ tweets served as the primary source for evaluating the user's stance on vaccination within the social graph sample. If the $C+$ tweets provided insufficient evidence, we then analyzed the user's remaining tweets from the COVID-19 vaccination period.

Model		$C+_{True}$	$C-_{True}$
HeRo	C^+	12	12
	C^-	2	18
DeGroot 0.75	C^+	20	0
	C^-	34	82
DeGroot 0.05	C^+	50	0
	C^-	4	64
node2vec	C^+	6	22
	C^-	0	15

Table 5.6: Model prediction against our samples on the social graph

5.5 Error Analysis

Classifying tweets into three categories of which two are adjacent but inherently different (i.e., classifying a tweet as COVID-related or vaccine hesitant proved to be a challenge, as seen in Fig. 5.4, which shows non-COVID-related tweets classified as COVID-related. This situation occurs primarily because of word occurrence (e.g., when COVID-19 or viruses are mentioned as a metaphor, such as in row nine of Table 4.1). The lack of context for a tweet can cause a false classification or a false annotation. For example, the annotation of the tweet in row two of Table 4.1 was determined to be false upon considering the tweet's context (the tweet was a reply to a COVID-19 skeptic famously known for calling COVID-19 "like a flu"). However, the HeRo model still correctly classified the tweet without seeing the context. Such an "error" could also happen in the reverse direction.

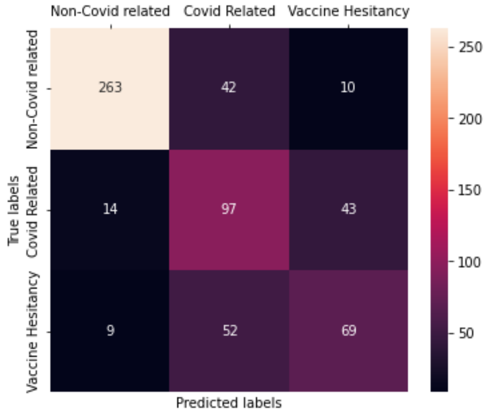


Figure 5.4: Confusion matrix of tweet classification for the HeRo model.

The network-based approaches node2vec and belief propagation were less accurate when tested on the AVS and their twins than when the users were expanded to the social graph (ego network) of the AVS and the twins. In the latter case, belief propagation produced the highest accuracy in the social graph user sample (Table 5.6). The lower accuracy of the language model when applied to the social graph may be explained by the limited scope of the training data.

The effectiveness of the language model may be limited by the scope of the training data. If the model is fine-tuned primarily on tweets promoting vaccine hesitancy that follow a specific narrative or rhetorical style, it might struggle to accurately classify tweets that deviate from that style. In essence, the model might overly focus on the specific rhetoric used in the training data, leading to misidentification of vaccine hesitancy expressed in different ways.

While sampling users on the social network graph, we encountered an interesting case where the language model misclassified as vaccine hesitant the head of the National Expert Cabinet for Dealing with the Corona Crisis, Professor Ran Balicer (@RanBalicer on the platform). Professor Balicer frequently tweeted about vaccinations and related ongoing research, which likely explains why most of his tweets revolve around COVID-19 and its vaccines, which were under development at the time. Unlike the language model, the network models did not classify Professor Balicer as vaccine hesitant.

The unexpected misclassification of Professor Balicer by the language model prompted us to investigate how the model performed on other COVID-19 experts and vocal vaccine proponents active on the platform (whom we shall refer to as "vaccine activists"). As shown in Table 5.7, only one of the five selected vaccine activists was not classified as vaccine hesitant by the HeRo language model. In contrast, the network models (Degroot and MLP) successfully classified all five activists as vaccine compliant.

Name	User Name	Known on Twitter for	HeRo	Degroot	MLP
Ran Balicer	@RanBalicer	National Expert Cabinet for Dealing with the COVID-19 Crisis (often on television)	$C+$	$C-$	$C-$
Moshe Bar Siman Tov	@moshebst	CEO of Ministry of Health at the inception of the pandemic	$C-$	$C-$	$C-$
Eran Segal	@segal_eran	Expert commentator from Weizmann Institute of Science	$C+$	$C-$	$C-$
Eldad SitBon	@LittleMoiz	Analyzing the pandemic on Twitter	$C+$	$C-$	$C-$
Unknown	@Anat_Holy	Popular anti-alternative-medicine figure; known for mocking anti-vaccination	$C+$	$C-$	$C-$

Table 5.7: Model results for known pro-vaccination activists on Twitter. Language model (HeRo), belief propagation (Degroot), and node2vec with multilayer preceptron (MLP).

This result highlights a vulnerability of the language model when applied to individuals on the opposite end of the vaccine-hesitancy spectrum. Vaccine activists frequently post about and advocate for vaccination, while also sharing their expertise. A possible explanation for the misclassification is that the language model struggles to differentiate between the language used by these activists and the language promoting vaccine hesitancy. This overlap might lead the model to misclassify the experts' tweets as vaccine hesitant.

Chapter 6

Discussion and Conclusions

This study demonstrates the potential of fine-tuned transformer models for user classification tasks, specifically for Hebrew tweets with Hebrew transformer models pretrained on vaccine hesitancy. This study could help policy makers and public health official understand the core components of vaccine hesitation in Israel.

6.1 Limitations

1. *Limited nuance in vaccine hesitancy.* This study focused on identifying users who actively promoted vaccine hesitancy or expressed strong doubts about public health recommendations regarding vaccines. This approach may not have fully captured the spectrum of vaccine hesitancy, which can include individuals with concerns about specific aspects of vaccines or individuals who are easily swayed by misinforma-

tion. Additionally, the approach may not distinguish between those who completely deny the existence of the virus and those who have doubts about vaccination but nevertheless acknowledge the existence of COVID-19.

2. *Misclassifications of public health experts and pro-vaccine activists.* The language model might misclassify public health experts who frequently discuss COVID-19 and vaccines as promoting vaccine hesitancy, despite their tweets aiming to educate and advocate for vaccination. A possible explanation for the misclassifications is the composition of the twin group, which is the user sample representing vaccine-compliant users. This group may not adequately capture the variety of language used by public health experts and pro-vaccine activists who frequently discuss COVID-19 and vaccines.

3. *Limited labeling for negative example users (twins).* We identified a set of anti-vaxxers on social media and automatically generated a control group (twins) based on these users. While the scarcity of anti-vaxx content within the training data from the twins justifies the assumption that most twins are unlikely to be anti-vaxxers, this approach relies on an implicit negative label.

4. *Single annotator for gold standard labels.* We relied on the author's annotations to identify vaccine-hesitant tweets (positive examples) and differentiate them from vaccine-compliant tweets. This strategy may have introduced subjectivity and bias into the training data, which could affect the accuracy and generalizability of the language model.

6.2 Ethical Considerations

Analyzing and modeling vaccine hesitancy on Twitter is of great importance. However, classifying *users* as vaccine hesitant based on an algorithm may result in falsely identifying users as vaccine-hesitant, which may result in suspension of their account or other measures. Although we always opted for a conservative approach and focused on aggregated measures characterizing the trends of a *platform*, we note that user labeling should be carefully used, ideally involving a "man-in-the-loop."

6.3 Conclusions and Future Work

This study investigated the effectiveness of language and network models for identifying vaccine hesitancy on social media platforms catering to Hebrew-speaking users in Israel. When using our models on the AVS group and their matched twin group, the language model performed better than both the baseline approach and the network model approach, where the network is the sharing network of the AVS+twins ego network. After annotating users *on the ego network* as vaccine hesitant or vaccine compliant, the belief propagation network model performed better than the language model, with the caveat that prominent vaccine experts who frequently discussed COVID-19 and vaccines were misclassified as vaccine hesitant. This highlights a potential vulnerability of language models when applied to individuals who actively advocate for vaccination but may use language that overlaps with the vaccine-hesitancy rhetoric.

Our ongoing research includes the use of the network features such as sharing or replying to social media posts. We use these networks to create a vector representation of a user network. Merging network features and the language models should improve our results.

A second research route is to use the sharing network to evaluate the correlation between political affiliation and vaccine hesitancy. This approach could involve using both manual annotation of the accounts of political officials or using the algorithm family known as “community detection,” which partitions news into highly connected “communities.” This method could partition the network into political segments without the risk of bias from manual annotations and could segment the vaccine-hesitancy community into subsections with differing degrees of hesitancy. Finally, the community approach could be useful for identifying users at risk of anti-vaccine sentiment and for developing targeted interventions to reduce anti-vaccine sentiment.

Appendix A

Table A.1 presents a sample of anti-vaccination comments made by former Fox News host Tucker Carlson, Robert F. Kennedy Jr. (2024 presidential candidate and the founder and chairman of the Children Health Defence), Podcaster Joe Rogan, and NBA star Kyrie Irving. All statements were retrieved on Jan. 14, 2024.

Name	Quote	Source (URL)	Date
Tucker Carlson	“Bill Gates has gained extraordinary powers over what you can and cannot do to your own body. Bill Gates would like you to take the coronavirus vaccine.”	Tucker Carlson Tonight (bitly.cx/ wNFcD)	Feb. 22, 2022
Tucker Carlson	“He [Anthoy Fauci] lied about herd immunity in order to sell more vaccines, which also didn’t work, which weren’t even actually vaccines, but they did hurt a lot of people, tens of thousands.”	Tucker Carlson Tonight (bitly.cx/ 4DVj)	Aug. 22, 2022
Tucker Carlson	“The point of mandatory vaccination is to identify the sincere Christians in the ranks, the free thinkers, the men with high testosterone levels, and anyone else who doesn’t love Joe Biden, and make them leave immediately. It’s a takeover of the U.S. military.”	Tucker Carlson Tonight (bitly.cx/ owgX)	Sept. 21, 2021
Robert F. Kennedy Jr.	“It is criminal medical malpractice to give a child one of these vaccines.”	AP News (bitly.cx/ 0Jnq)	Dec. 15, 2021
Robert F. Kennedy Jr.	“CDC’s convenient new metric will allow the Medical Cartel to stay in the Covid death business as long as it likes while enjoying all the attendant benefits of power and control,even if COVID-19 disappears on its own as did SARS and all previous coronavirus pandemics.”	Instagram - archived (archive.is/ uhYwu)	April 13, 2020

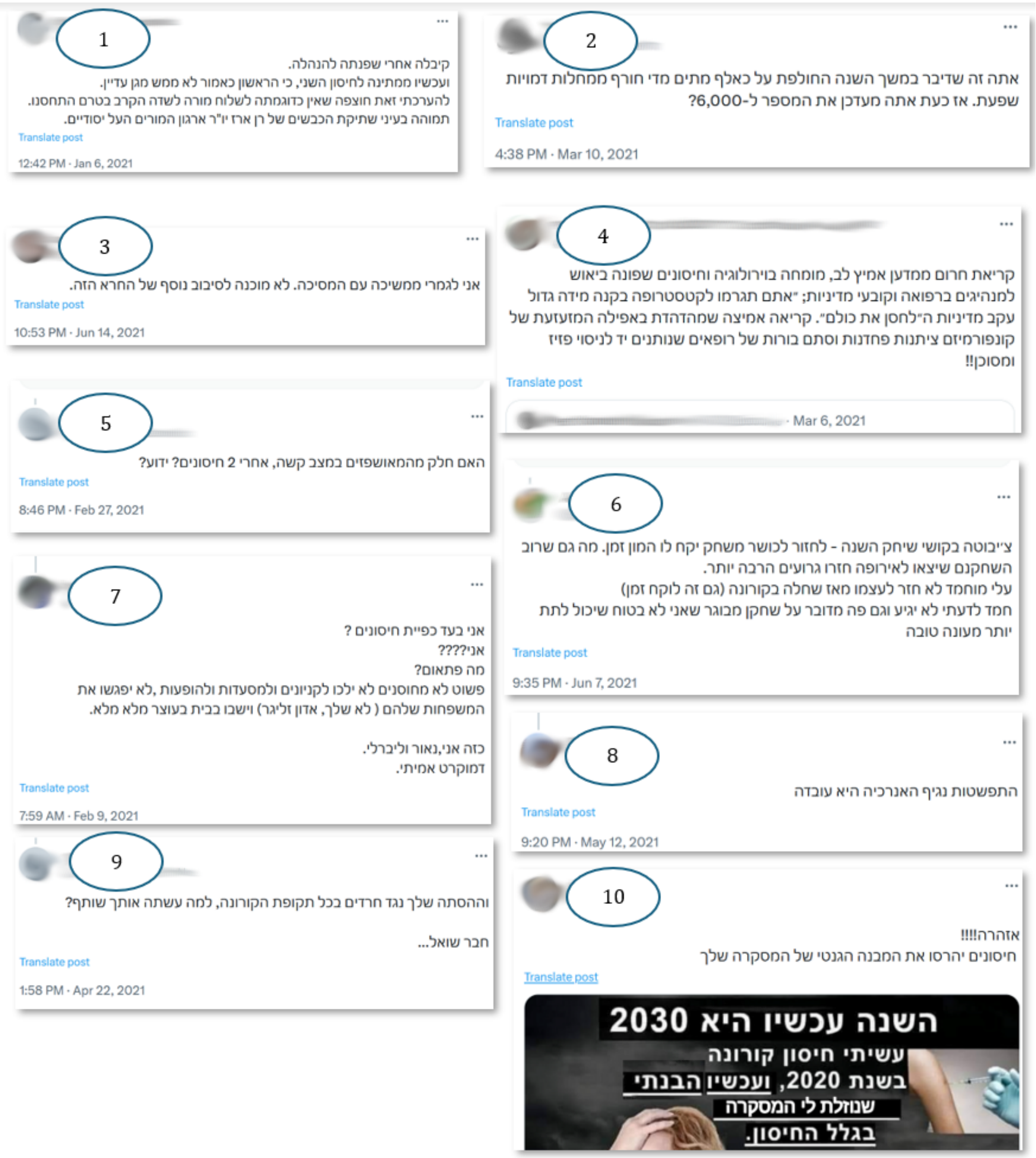
Table A.1

Name	Quote	Source (URL)	Date
Joe Rogan	“If you’re, like, 21 years old, and you say to me, should I get vaccinated? I’ll go, ’No’ . . . . If you’re a healthy person, and you’re exercising all the time, and you’re young, and you’re eating well, like, I don’t think you need to worry about this.”	Reuters (bitly.cx/ j5Vt5)	April 23, 2021
Kyrie Irving	“That’s the role I play, but I never wanted to give up my passion, my love, my dream just over this mandate.”	Forbes (bitly.cx/ J9VlJ)	Nov. 14, 2021
Kyrie Irving	“I am staying grounded in what I believe in. It is as simple as that. It is not about being anti-vax or about being on one side or the other. It is just really about being true to what feels good for me . . . . If I am going to be demonized for having more questions and taking my time to make a decision with my life, that is just what it is . . . . I know the consequences of the decisions that I make with my life. I am not here to sugarcoat any of that.”	ESPN (bitly.cx/ FtKUo)	Oct. 14, 2021

Table A.2

Appendix B

Figure B.1 presents the original Hebrew tweets, corresponding to the translations in Table 4.1.



1

קיבלה אחרי שפנתה להנהלה.
ועכשיו ממתינה לחיסון השני, כי הראשון כאמור לא ממש מגן עדיין.
להערכתי זאת חוצפה שאין כדוגמתה לשלוח מורה לשדה הקרב בטרם התחסנו.
תמוהה בעיני שתיקת הכבשים של רן ארז יו"ר ארגון המורים העל יסודיים.

Translate post

12:42 PM · Jan 6, 2021

2

אתה זה שדיבר במשך השנה החולפת על כאלף מתים מדי חורף ממחלות דמויות שפעת. אז כעת אתה מעדכן את המספר ל-6,000?

Translate post

4:38 PM · Mar 10, 2021

3

אני לגמרי ממשיכה עם המסיכה. לא מוכנה לסיבוב נוסף של החרא הזה.

Translate post

10:53 PM · Jun 14, 2021

4

קריאת חרום ממדען אמיץ לב, מומחה בוירולוגיה וחיסונים שפונה ביאוש למנהיגים ברפואה וקובעי מדיניות; "אתם תגרמו לקטסטרופה בקנה מידה גדול עקב מדיניות ה"לחסן את כולם". קריאה אמיצה שמהדהדת באפילה המזעזעת של קונפורמיזם ציתנות פחדנות וסתם בורות של רופאים שנותנים יד לניסוי פזיז ומסוכן!!

Translate post

· Mar 6, 2021

5

האם חלק מהמאושפזים במצב קשה, אחרי 2 חיסונים? ידוע?

Translate post

8:46 PM · Feb 27, 2021

6

צ׳יבוטה בקושי שיחק השנה - לחזור לכושר משחק יקח לו המון זמן. מה גם שרוב השחקנם שיצאו לאירופה חזרו גרועים הרבה יותר.
עלי מוחמד לא חזר לעצמו מאז שחלה בקורונה (גם זה לוקח זמן)
חמד לדעתי לא יגיע וגם פה מדובר על שחקן מבוגר שאני לא בטוח שיכול לתת יותר מעונה טובה

Translate post

9:35 PM · Jun 7, 2021

7

אני בעד כפיית חיסונים ?
אני????
מה פתאום?
פשוט לא מחוסנים לא ילכו לקניונים ולמסעדות ולהופעות ,לא יפגשו את המשפחות שלהם (לא שלך, אדון זליגר) וישבו בבית בעוצר מלא מלא.

כזה אני,נאור וליברלי.
דמוקרט אמיתי.

Translate post

7:59 AM · Feb 9, 2021

8

התפשטות נגיף האנרכיה היא עובדה

Translate post

9:20 PM · May 12, 2021

9

וההסתה שלך נגד חרדים בכל תקופת הקורונה, למה עשתה אותך שותף?

חבר שואל...

Translate post

1:58 PM · Apr 22, 2021

10

אזהרה!!!!
חיסונים יהרסו את המבנה הגנטי של המסקרה שלך

Translate post

Figure B.1: Numbers correspond to tweet indices in Table 4.1.

Bibliography

[1] Guetta, Eylon, Shmidman, Avi, Shmidman, Shaltiel, Shmidman, Cheyn Shmuel, Guedalia, Joshua, Koppel, Moshe, Bareket, Dan, Seker, Amit, and Tsarfaty, Reut. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all, 2022. URL https://arxiv.org/abs/2211.15199.

[2] Hinman, Alan. Eradication of vaccine-preventable diseases. *Annual review of public health*, 20(1):211–229, 1999.

[3] Greenwood, Brian. The contribution of vaccination to global health: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1645):20130433, 2014.

[4] Henderson, Donald A. The eradication of smallpox–an overview of the past, present, and future. *Vaccine*, 29:D7–D9, 2011.

[5] Minor, Philip D. Polio eradication, cessation of vaccination and re-emergence of disease. *Nature Reviews Microbiology*, 2(6):473–482, 2004.

[6] Taylor, Luke E, Swerdfeger, Amy L, and Eslick, Guy D. Vaccines are not

associated with autism: an evidence-based meta-analysis of case-control and cohort studies. *Vaccine*, 32(29):3623–3629, 2014.

[7] MacDonald, Noni E et al. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164, 2015.

[8] Rossen, Isabel, Hurlstone, Mark J, Dunlop, Patrick D, and Lawrence, Carmen. Accepters, fence sitters, or rejecters: Moral profiles of vaccination attitudes. *Social science & medicine*, 224:23–27, 2019.

[9] Dong, Ensheng, Du, Hongru, and Gardner, Lauren. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[10] Watson, Oliver J, Barnsley, Gregory, Toor, Jaspreet, Hogan, Alexandra B, Winskill, Peter, and Ghani, Azra C. Global impact of the first year of covid-19 vaccination: a mathematical modelling study. *The Lancet Infectious Diseases*, 22(9):1293–1302, 2022.

[11] Mallapaty, Smriti, Callaway, Ewen, Kozlov, Max, Ledford, Heidi, Pickrell, John, Van Noorden, Richard, et al. How covid vaccines shaped 2021 in eight powerful charts. *Nature*, 600(7890):580–583, 2021.

[12] Dror, Amiel A, Eisenbach, Netanel, Taiber, Shahar, Morozov, Nicole G, Mizrachi, Matti, Zigron, Asaf, Srouji, Samer, and Sela, Eyal. Vaccine hesitancy: the next challenge in the fight against covid-19. *European journal of epidemiology*, 35:775–779, 2020.

[13] Troiano, Gianmarco and Nardi, Alessandra. Vaccine hesitancy in the era of covid-19. *Public health*, 194:245–251, 2021.

[14] Soares, Patricia, Rocha, João Victor, Moniz, Marta, Gama, Ana, Laires, Pedro Almeida, Pedro, Ana Rita, Dias, Sónia, Leite, Andreia, and Nunes, Carla. Factors associated with covid-19 vaccine hesitancy. *Vaccines*, 9(3):300, 2021.

[15] Bergen, Nicole, Kirkby, Katherine, Fuertes, Cecilia Vidal, Schlotheuber, Anne, Menning, Lisa, Mac Feely, Stephen, O'Brien, Katherine, and Hosseinpoor, Ahmad Reza. Global state of education-related inequality in covid-19 vaccine coverage, structural barriers, vaccine hesitancy, and vaccine refusal: findings from the global covid-19 trends and impact survey. *The Lancet Global Health*, 11(2):e207–e217, 2023.

[16] Dagan, Noa, Barda, Noam, Kepten, Eldad, Miron, Oren, Perchik, Shay, Katz, Mark A, Hernán, Miguel A, Lipsitch, Marc, Reis, Ben, and Balicer, Ran D. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 384(15):1412–1423, 2021.

[17] Vosoughi, Soroush, Roy, Deb, and Aral, Sinan. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[18] Loomba, Sahil, de Figueiredo, Alexandre, Piatek, Simon J, de Graaf, Kristen, and Larson, Heidi J. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021. ISSN 2397-3374.

[19] Ognyanova, Katherine, Lazer, David, Baum, Matthew, Druckman, James, Green, Jon, Perlis, Roy H, Santillana, Mauricio, Simonson, Matthew D, Lin, Jennifer, and Uslu, Ata. The covid states project #60: Covid-19 vaccine misinformation: From uncertainty to resistance, August 2021. URL `osf.io/xtjad`.

[20] WHO, A. World health statistics 2016: monitoring health for the sdgs sustainable development goals. *World Health Organization*, 2022.

[21] Rosen, B., Waitzberg, R., and Israeli, A. Israel's rapid rollout of vaccinations for covid-19. *Isr J of Health Policy Res*, 10(6), 2021.

[22] Velan, Baruch, Kaplan, Giora, Ziv, Arnona, Boyko, Valentina, and Lerner-Geva, Liat. Major motives in non-acceptance of a/h1n1 flu vaccination: the weight of rational assessment. *Vaccine*, 29(6):1173–1179, 2011.

[23] Muhsen, Khitam, El-Hai, Reem Abed, Amit-Aharon, Anat, Nehama, Haim, Gondia, Mervat, Davidovitch, Nadav, Goren, Sophy, and Cohen, Dani. Risk factors of underutilization of childhood immunizations in ultraorthodox jewish communities in israel despite high access to health care services. *Vaccine*, 30(12):2109–2115, 2012.

[24] Kaliner, E (1), Moran-Gilad, J, Grotto, I, Somekh, E, Kopel, E, Gdalevich, M, Shimron, E, Amikam, Y, Leventhal, A, Lev, B, et al. Silent reintroduction of wild-type poliovirus to israel, 2013–risk communication challenges in an argumentative atmosphere. *Eurosurveillance*, 19 (7), 2014.

[25] Velan, Baruch. Vaccine hesitancy as self-determination: an israeli perspective. *Israel Journal of Health Policy Research*, 5(1):1–6, 2016.

[26] Reiss, Dorit Rubinstein and Karako-Eyal, Nili. Informed consent to vaccination: theoretical, legal, and empirical insights. *American Journal of Law & Medicine*, 45(4):357–419, 2019.

[27] Rosen, Bruce, Waitzberg, Ruth, Israeli, Avi, Hartal, Michael, and Davidovitch, Nadav. Addressing vaccine hesitancy and access barriers to achieve persistent progress in israel's covid-19 vaccination program. *Israel journal of health policy research*, 10(1):1–20, 2021.

[28] Kata, Anna. Anti-vaccine activists, web 2.0, and the postmodern paradigm–an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25):3778–3789, 2012.

[29] Betsch, Cornelia, Brewer, Noel T, Brocard, Pauline, Davies, Patrick, Gaissmaier, Wolfgang, Haase, Niels, Leask, Julie, Renkewitz, Frank, Renner, Britta, Reyna, Valerie F, et al. Opportunities and challenges of web 2.0 for vaccination decisions. *Vaccine*, 30(25):3727–3733, 2012.

[30] Dredze, Mark, Broniatowski, David A, Smith, Michael C, and Hilyard, Karen M. Understanding vaccine refusal: why we need social media now. *American journal of preventive medicine*, 50(4):550–552, 2016.

[31] Hoffman, Beth L, Felter, Elizabeth M, Chu, Kar-Hai, Shensa, Ariel, Hermann, Chad, Wolynn, Todd, Williams, Daria, and Primack, Brian A. It's not all about autism: The emerging landscape of anti-vaccination sentiment on facebook. *Vaccine*, 37(16):2216–2223, 2019.

[32] Jamison, Amelia M, Broniatowski, David A, Dredze, Mark, Sangraula, Anu, Smith, Michael C, and Quinn, Sandra C. Not just conspiracy theories: Vaccine opponents and proponents add to the covid-19 'infodemic'on twitter. *Harvard Kennedy School Misinformation Review*, 1, 2020.

[33] Jamison, Amelia, Broniatowski, David A, Smith, Michael C, Parikh, Kajal S, Malik, Adeena, Dredze, Mark, and Quinn, Sandra C. Adapting and extending a typology to identify vaccine misinformation on twitter. *American Journal of Public Health*, 110(S3):S331–S339, 2020.

[34] Dunn, Adam G, Surian, Didi, Leask, Julie, Dey, Aditi, Mandl, Kenneth D, and Coiera, Enrico. Mapping information exposure on social media to explain differences in hpv vaccine coverage in the united states. *Vaccine*, 35(23):3033–3040, 2017.

[35] Broniatowski, David A, Jamison, Amelia M, Johnson, Neil F, Velasquez, Nicolás, Leahy, Rhys, Restrepo, Nicholas Johnson, Dredze, Mark, and Quinn, Sandra C. Facebook pages, the "disneyland" measles outbreak, and promotion of vaccine refusal as a civil right, 2009–2019. *American journal of public health*, 110(S3):S312–S318, 2020.

[36] Green, Jon, Druckman, James N, Baum, Matthew A, Ognyanova, Katherine, Simonson, Matthew D, Perlis, Roy H, and Lazer, David. Media use and vaccine resistance. *PNAS nexus*, 2(5):pgad146, 2023.

[37] Habib, Hussam and Nithyanand, Rishab. The morbid realities of social media: An investigation into the narratives shared by the deceased vic-

tims of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 303–314, 2023.

[38] Lenti, Jacopo, Mejova, Yelena, Kalimeri, Kyriaki, Panisson, André, Paolotti, Daniela, Tizzani, Michele, and Starnini, Michele. Global misinformation spillovers in the vaccination debate before and during the covid-19 pandemic: Multilingual twitter study. *JMIR infodemiology*, 3: e44714, 2023.

[39] Featherstone, Jieyu D, Ruiz, Jeanette B, Barnett, George A, and Millam, Benjamin J. Exploring childhood vaccination themes and public opinions on twitter: A semantic network analysis. *Telematics and Informatics*, 54:101474, 2020.

[40] Alleyne, Brian. *Narrative networks: Storied approaches in a digital age.* Sage, 2014.

[41] Lutkenhaus, Roel O, Jansz, Jeroen, and Bouman, Martine PA. Mapping the dutch vaccination debate on twitter: Identifying communities, narratives, and interactions. *Vaccine: X*, 1:100019, 2019.

[42] Broniatowski, David A, Greene, Kevin T, Pisharody, Nilima, Rogers, Daniel J, and Shapiro, Jacob N. Measuring the monetization strategies of websites with application to pro-and anti-vaccine communities. *Scientific Reports*, 13(1):15964, 2023.

[43] Mahajan, Rutuja, Romine, William, Miller, Michele, and Banerjee, Tanvi. Analyzing public outlook towards vaccination using twitter. In

2019 IEEE international conference on big data (big data), pages 2763–2772. IEEE, 2019.

[44] Yuan, Xiaoyi, Schuchard, Ross J, and Crooks, Andrew T. Examining emergent communities and social bots within the polarized online vaccination debate in twitter. *Social media+ society*, 5(3):2056305119865465, 2019.

[45] Cossard, Alessandro, Morales, Gianmarco De Francisci, Kalimeri, Kyriaki, Mejova, Yelena, Paolotti, Daniela, and Starnini, Michele. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140, 2020.

[46] Bail, Christopher. *Breaking the social media prism : how to make our platforms less polarizing.* Princeton University Press, Princeton, 2021. ISBN 9780691203423.

[47] Funk, Cary and Tyson, Alec. Intent to get a covid-19 vaccine rises to 60% as confidence in research and development process increases, 2020. URL `https://www.pewresearch.org/science/wp-content/uploads/sites/16/2020/12/PS_2020.12.03_covid19-vaccine-intent_REPORT.pdf`.

[48] Druckman, James N, Ognyanova, Katherine, Baum, Matthew A, Lazer, David, Perlis, Roy H, Volpe, John Della, Santillana, Mauricio, Chwe, Hanyu, Quintana, Alexi, and Simonson, Matthew. The role of race, re-

ligion, and partisanship in misperceptions about covid-19. *Group Processes & Intergroup Relations*, 24(4):638–657, 2021.

[49] Tomeny, Theodore S, Vargo, Christopher J, and El-Toukhy, Sherine. Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. *Social science & medicine*, 191:168–175, 2017.

[50] Küçükali, Hüseyin, Ataç, Ömer, Palteki, Ayşe Seval, Tokaç, Ayşe Zülal, and Hayran, Osman. Vaccine hesitancy and anti-vaccination attitudes during the start of covid-19 vaccination program: a content analysis on twitter data. *Vaccines*, 10(2):161, 2022.

[51] Pacheco, María Leonor, Islam, Tunazzina, Mahajan, Monal, Shor, Andrey, Yin, Ming, Ungar, Lyle, and Goldwasser, Dan. A holistic framework for analyzing the covid-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, 2022.

[52] Velan, Baruch, Boyko, Valentina, Lerner-Geva, Liat, Ziv, Arnona, Yagar, Yaakov, and Kaplan, Giora. Individualism, acceptance and differentiation as attitude traits in the public's response to vaccination. *Human vaccines & immunotherapeutics*, 8(9):1272–1282, 2012. ISSN 2164-5515.

[53] for Disease Control Ministry of Health, ICDC-Israel Center. Monitoring report of respiratory viruses, 2020. URL

`https://www.gov.il/BlobFolder/reports/corona-flu-05122020/he/files_weekly-flu-corona_corona-flu-05122020.pdf`.

[54] Einschenk, Moriel, Voloh Dahan, Polina, and Gandelman, Elinor. The correlation between social media groups affiliation and stance regarding vaccines in israel (hebrew). *The nurse in Israel (Herew)*, 2019.

[55] Seker, Amit, Bandel, Elron, Bareket, Dan, Brusilovsky, Idan, Greenfeld, Refael Shaked, and Tsarfaty, Reut. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*, 2021.

[56] Chriqui, Avihay and Yahav, Inbal. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 2022.

[57] Shalumov, Vitaly and Haskey, Harel. Hero: Roberta and longformer hebrew language models. *arXiv:2304.11077*, 2023.

[58] Grover, Aditya and Leskovec, Jure. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[59] Pozek, Mislav, Sikic, Lucija, Afric, Petar, Kurdija, Adrian S, Vladimir, Klemo, Delac, Goran, and Silic, Marin. Performance of common classifiers on node2vec network representations. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 925–930. IEEE, 2019.

[60] Ribeiro, Manoel, Calais, Pedro, Santos, Yuri, Almeida, Virgílio, and Meira Jr, Wagner. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[61] Mathew, Binny, Dutt, Ritam, Goyal, Pawan, and Mukherjee, Animesh. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182, 2019.

[62] Israeli, Abraham and Tsur, Oren. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121, 2022.

[63] Fruchterman, Thomas MJ and Reingold, Edward M. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11): 1129–1164, 1991.

[64] Bi, Lin, Wang, Yue, Zhao, Jian-ping, Qi, Hui, and Zhang, Ying. Social network information visualization based on fruchterman reingold layout algorithm. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, pages 270–273. IEEE, 2018.

[65] Fogarty, Emily A. Visualizing the relationship between geographic and social media network space. *GeoJournal*, 86:2483–2500, 2021.

תוכן עניינים

1.1 הססנות לחיסונים 2

1.2 המחקר 3

2.1 התחסנות בארץ 5

2.2 הססנות לחיסונים בראי המדיה החברתיות 7

3.1 מערך נתונים גולמי 10

3.2 גולד לייבלס ופרוטוקול שידוכי ביקורת 11

3.3 תיוג נתונים 12

3.4 הגרף החברתי 14

4.1 סיווג ברמת הציוץ 16

4.2 סיווג ברמת המשתמש 16

5.1 סיווג ציוצים 20

5.2 סיווג משתמשים 21

5.3 ניבוי על סמך רשת 24

5.4 רשת חברתית 25

5.5 ניתוח תקלות 28

6.1 מגבלות המחקר 31

6.2 שיקולים אתיים 33

6.3 מסקנות ומחקר עתידי 33

לא רק ניתוח טקסטים: ניתוח רשתות ועיבוד שפה טבעית לזיהוי מתנגדי חיסונים בטוויטר העברי

עומר נוי

דעבודת גמר לתואר מוסמך להנדסה

אוניברסיטת בן־גוריון בנגב

2024

תקציר

כמו ה־NJ.1 המדבק מאוד שהחל להתפשט בסוף דצמבר 2023, אשר לעיתים קרובות עמידים לחיסונים הקיימים, מתנגדי החיסונים פוגעים במאמצים להגן על האוכלוסייה באמצעות השגת חסינות עדר, האטת הופעתם של וריאנטים חדשים ומיגור אפשרי של המחלה. רשתות חברתיות מספקות למתנגדי החיסונים כלי יעיל להפצת מסרים הפוגעים בבריאות הציבור.

במחקר זה אנו משתמשים במאגר נתונים ייחודי ־ המכסה 80־90% מהציוצים העבריים ־ כדי לחקור התנגדות חיסונים במהלך שלוש השנים הראשונות של מגפת הקורונה. אנו מבצעים כיוון דק (FINE־TUNE) למגוון רחב של מודלים שפה גדולים (LLMs) המותאמים לשפות עשירות מורפולוגיה, כדי לזהות ציוצים המבטאים התנגדות לחיסונים, ולא חששות לגיטימיים בנוגע לחיסונים ולקורונה, תוך השגת ציון F SCORE של 0.75. בהמשך, אנו משתמשים ב־LLMs

אלה, יחד עם וקטוריזציה של רשת השיתופים ומודלים של דיפוזיה, כדי לזהות באופן מדויק משתמשים המקדמים באופן פעיל התנגדות להתחסן, תוך השגת ציון F SCORE של 0.87 ברמת המשתמש בסביבה מאתגרת. אנו משווים את תוצאות הסיווג שהושגו על ידי גישות שונות ומדברים על היתרונות שלהן, המגבלות שלהן והדרכים שבהן הן מאפשרות הערכה של מגמות ההתגדות להתחסן ברשת הרחבה יותר. אנו מוצאים שבעוד מסווגים מבוססי טקסט עולים על גישות מבוססות גרף, הם סובלים מטעויות חיוב כוזב ייחודיות - סיווג של מומחים רפואיים כפעילי נגד חיסונים.

אוניברסיטת בן־גוריון בנגב

הפקולטה להנדסה

המחלקה להנדסת מערכות תוכנה ומידע

לא רק ניתוח טקסטים: ניתוח רשתות ועיבוד שפה טבעית לזיהוי מתנגדי חיסונים בטוויטר העברי

חיבור זה מהווה חלק מהדרישות לקבלת התואר מוסמך להנדסה (M.Sc)

עומר נוי

בהנחיית **דוקטור אורן צור**

יוני 2024