*big data and cognitive computing*

MDPI

*Article*

# Q8VaxStance: Dataset Labeling System for Stance Detection towards Vaccines in Kuwaiti Dialect

**Hana Alostad [1,*], Shoug Dawiek [1] and Hasan Davulcu [2]**

[1] Computer Science Department, Gulf University for Science and Technology, Mubarak Al-Abdullah 32093, Kuwait; dawiek.s@gust.edu.kw
[2] Computer Science Department, School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA; hdavulcu@asu.edu
* Correspondence: alostad.h@gust.edu.kw

**Abstract:** The Kuwaiti dialect is a particular dialect of Arabic spoken in Kuwait; it differs significantly from standard Arabic and the dialects of neighboring countries in the same region. Few research papers with a focus on the Kuwaiti dialect have been published in the field of NLP. In this study, we created Kuwaiti dialect language resources using Q8VaxStance, a vaccine stance labeling system for a large dataset of tweets. This dataset fills this gap and provides a valuable resource for researchers studying vaccine hesitancy in Kuwait. Furthermore, it contributes to the Arabic natural language processing field by providing a dataset for developing and evaluating machine learning models for stance detection in the Kuwaiti dialect. The proposed vaccine stance labeling system combines the benefits of weak supervised learning and zero-shot learning; for this purpose, we implemented 52 experiments on 42,815 unlabeled tweets extracted between December 2020 and July 2022. The results of the experiments show that using keyword detection in conjunction with zero-shot model labeling functions is significantly better than using only keyword detection labeling functions or just zero-shot model labeling functions. Furthermore, for the total number of generated labels, the difference between using the Arabic language in both the labels and prompt or a mix of Arabic labels and an English prompt is statistically significant, indicating that it generates more labels than when using English in both the labels and prompt. The best accuracy achieved in our experiments in terms of the Macro-F1 values was found when using keyword and hashtag detection labeling functions in conjunction with zero-shot model labeling functions, specifically in experiments KHZSLF-EE4 and KHZSLF-EA1, with values of 0.83 and 0.83, respectively. Experiment KHZSLF-EE4 was able to label 42,270 tweets, while experiment KHZSLF-EA1 was able to label 42,764 tweets. Finally, the average value of annotation agreement between the generated labels and human labels ranges between 0.61 and 0.64, which is considered a good level of agreement.

**Keywords:** Arabic NLP; Kuwaiti dialect; dataset labeling; stance detection; weak supervised learning; zero-shot learning



**Citation:** Alostad, H.; Dawiek, S.; Davulcu, H. Q8VaxStance: Dataset Labeling System for Stance Detection towards Vaccines in Kuwaiti Dialect. *Big Data Cogn. Comput.* **2023**, *7*, 151. https://doi.org/10.3390/bdcc7030151

Academic Editor: Domenico Ursino

Received: 10 August 2023
Revised: 5 September 2023
Accepted: 13 September 2023
Published: 15 September 2023



## 1. Introduction

COVID-19 vaccinations were essential in Kuwait for controlling the spread of the virus and protecting public health. However, there have been concerns about vaccine hesitancy and misinformation in the country [1–3], which may impact vaccination rates and the effectiveness of vaccination efforts for other types of vaccines in the future.

This trend is concerning, as vaccines are essential for preventing the spread of infectious diseases and protecting public health [4,5]. Detecting and addressing opposing stances towards vaccination on social media are essential public health efforts. Public health officials need to have access to this information to target interventions and address misinformation. In addition, they must present accurate and evidence-based information about vaccines to the public to combat vaccine hesitancy and protect the health of individuals and communities.

This research aims to label a large dataset of tweets written in the Kuwaiti dialect. The tweets are classified pragmatically depending on their attitude towards vaccines in order to track negative views on social media. This research is an integral part of a more comprehensive attempt to understand the elements that cause vaccine hesitancy and to create practical approaches for addressing it. Furthermore, by analyzing social media data we can better understand the methods of spreading misinformation and vaccine-related conspiracy theories and their consequences on public opinion. Ultimately, this knowledge can help public health officials to propose initiatives to secure the health of individuals and communities.

The main contribution of this research is creating the first dataset of tweets labeled regarding stance towards vaccines in the Kuwaiti dialect (42,764 labeled tweets). This dataset is a valuable resource for researchers studying vaccine hesitancy and its impact on public health. Additionally, this research implements the first Kuwaiti dialect annotation system for vaccine stance detection (Q8VaxStance) by using weak supervised learning and applying prompt engineering to zero-shot models as labeling functions to programmatically annotate the dataset regarding stance towards vaccines in the Kuwaiti dialect. The use of zero-shot models as labeling functions and weak supervised learning frameworks enables us to programmatically annotate a large dataset with minimal assistance from subject matter experts and minimal need for manually labeling a large dataset; thus, it enables us to save time and money, as recruiting expert annotators is an expensive and time-consuming task.

Finally, considering the limited availability of linguistic resources for the Kuwaiti dialect, this research tries to fill this gap in the field of natural language processing by providing a dataset to develop and evaluate machine learning models for stance detection in the Kuwaiti dialect. The following are the research questions of our study:

1. How can we create a labeling system to annotate a large dataset of Kuwaiti dialect tweets for stance detection towards vaccines with or without help from subject matter experts (SMEs)?
2. What experimental setup produces the best performance for the proposed labeling system?

This paper is organized as follows. In the Background section, we review the relevant literature on vaccine hesitancy and stance detection towards the COVID-19 vaccine, natural language processing (NLP) research involving the Kuwaiti dialect, and dataset annotation approaches in NLP. In the Methodology section we describe the dataset collection and preparation process. Next, we explain the process of labeling the dataset manually and describe the steps and architecture of the proposed Q8VaxStance labeling system. Next, in the Experimental Results and Discussion section, we present the results of our performance evaluation based on the Q8VaxStance labeling system experiments. Finally, in the Conclusion section, we summarize the study's main findings and propose several directions for future work.

## 2. Background

### *2.1. Vaccine Hesitancy and Stance Detection Using Social Network Analysis and Natural Language Processing*

The COVID-19 pandemic significantly affected the overall stance towards vaccines, as it increased negative attitudes towards vaccines in Kuwait and around the globe [1–3,6]. This should raise a red flag and alert policymakers and governments to take action.

Many researchers have studied this topic; for example, the researchers of [7] used multi-task aspect-based sentiment analysis (ABSA) and social features for stance detection in tweets based on BiGRU–BERT deep learning models. It combines aspect-based sentiment information with features based on textual and contextual information that does not emerge directly from Twitter texts. Another contribution to this topic is found in [8], where the researchers presented a dataset of Twitter posts with a strong anti-vaccine stance to be used in studying anti-vaccine misinformation on social media and to enable a better understanding of vaccine hesitancy. In [9], the researchers collected and annotated 15,000 tweets as misinformation or general vaccine tweets. The paper's best classification

performance resulted from using the BERT language model, with a 0.98 F1 score on the test set. The study presented in [10] analyzed COVID-19 vaccine tweets and tested their association with vaccination rates in 192 countries worldwide. The authors compared COVID-19 vaccine tweets by country in terms of (1) the number of related tweets per million Twitter users, (2) the proportion of tweets mentioning adverse events (death, side effects, and blood clots), (3) the appearance of negative sentiments as compared to positive sentiments, and (4) the appearance of fear, sadness, or anger as compared to joy. Finally, in contrast to the above research papers, which focused on negative stances, the researchers in [5] investigated and focused on the trend in positive attitudes towards vaccines across ten countries.

### 2.2. Natural Language Processing (NLP) of Kuwaiti Dialect

There has been an increased interest in developing natural language processing (NLP) models for the Arabic language. Arabic is a widely spoken and written language with a significant presence in the online world. Researchers in the Arabic world have started to focus on creating resources and language models for the Arabic language; examples of Arabic language models include AraBERT [11], ARBERT, MARBET [12], and CAMeLBERT [13], all of which focus on Modern Standard Arabic (MSA). In addition, there are models that cover Arabic dialects for specific countries.

We have found that there is a gap in the field of natural language processing for the Kuwaiti dialect; there is limited availability of linguistic resources for this dialect, with only a few published research papers in the field of NLP focusing on it [14–17].

In [14], the authors used a traditional machine learning approach by applying decision tree and SVM algorithms to classify opinions expressed in microblogging posts in the Kuwaiti dialect. They used a dataset of Kuwaiti Twitter posts annotated manually by three native Kuwaiti dialect speakers, enabling the researchers to achieve average values of precision and recall of 76% and 61%, respectively, with the SVM algorithm.

Another research study on the Kuwaiti dialect was conducted by the authors of [15]; in this paper, the researchers presented an approach to analyze the content of tweets by merging a text mining strategy with the spatial information in order to assess the topics of interest. In this way, they provided a deeper understanding of the topics people think about, when they think about them, and where they tweet about them. The results showed that the four most popular topics of interest in Kuwait were religion, emotion, education, and policy. In addition, they found that on Fridays people posted more about religion and that on weekends they tweeted more often about emotional expressions. Moreover, people posted more about policy and education on weekdays rather than on weekends.

The most recently published research papers studying the Kuwaiti dialect are [16,17]. In [16], we proposed a weak supervised approach to construct a large labeled corpus for sentiment analysis of tweets written in the Kuwaiti dialect. The proposed automated labeling system achieved a high level of annotation agreement between the automated labeling system and human-annotated labels, with 93% pairwise percent agreement and a 0.87 Cohen's kappa coefficient. Furthermore, we evaluated the dataset using multiple traditional machine learning classifiers and advanced deep learning language models to test its performance. The best reported accuracy was 89% when the resulting labeled dataset was trained with the ARBERT model. The labeling system architecture of Q8VaxStance is different from the labeling proposed system in [16]; first, in Q8VaxStance the main labeling task is stance detection. In addition, we experimented with different types of labeling functions (zero-shot models, keyword detection) and used prompt engineering. In [16], on the other hand, the main task was sentiment classification, not stance; moreover, we used only one simple fixed prompt, with all labeling functions as zero-shot models, and did not experiment with the keyword detection labeling functions. Finally, the dataset used in [16] differs from Q8VaxStance regarding the time frame, the type of extracted events, and the size. Thus, although the two proposed systems are both based on weak supervision, they are different and not comparable.

Contrary to previous papers that collected and used a dataset from Twitter in their experiments, the researchers in [17] collected and analyzed a corpus of WhatsApp group chats involving mixed-gender Kuwaiti participants. This pilot study aimed to obtain insights into features to be used later for developing a gender classification system for the Kuwaiti dialect. The study's results showed no significant differences between men and women in the number of turns, length of turns, and number of emojis. However, the study showed that men and women differed in their use of lengthened words and in the emojis that they used [17].

Based on the above review, there is an opportunity for researchers in the field of NLP to in filling the gap with respect to the Kuwaiti dialect, which remains underrepresented and not widely covered in this academic field.

### 2.3. Dataset Labeling Approaches

Data labeling is a challenging task for any NLP project; with the advances in deep learning and transfer learning algorithms, there is an increasing need to label large datasets. On the other hand, labeling large datasets is a time-consuming task, and subject matter experts (SMEs) generally do not have time to label these datasets, as they already have their own tasks to focus on. Obtaining labels annotated by experts can be expensive and time-consuming, while labels from crowdsourced labelers often contain mistakes that can affect the performance of supervised machine learning models [18]. Lastly, privacy may be an issue for certain projects, in which case the task of labeling the dataset cannot be outsourced or assigned to SMEs. Many academic researchers have proposed solutions allowing more data to be labeled with or without the limited help of human annotators. The following are among the approaches that can be used to annotate datasets for machine learning with limited or no help from annotators. The first approach is to use an active learning system, in which a human annotator makes queries in the form of unlabeled instances to achieve high accuracy of labeling with fewer training labels by allowing a model to choose the data to be annotated and ultimately used for learning [19]. The second approach is semi-supervised learning, a machine learning approach that combines small labeled and unlabeled samples to train models. It uses unsupervised algorithms to leverage the unlabeled data to improve the model's performance by utilizing the additional information present in the unlabeled samples [20]. In data annotation, weak supervised learning refers to creating labeled training data efficiently using various sources containing heuristics and knowledge bases without relying on fully annotated data. It allows for creating a large set of noisy labeled training data programmatically using various sources [21].

The Snorkel framework is an open-source weak supervised learning framework. Researchers at the Stanford AI Lab proposed this project, which started in 2015; it is the oldest and most stable among the available weak supervised learning software frameworks. The steps of the Snorkel system are as follows [22]:

1. SMEs write labeling functions (LFs) that express weak supervision sources such as distant supervision, patterns, and heuristics.
2. Snorkel applies the LFs on unlabeled data and learns a generative model to combine the LF outputs into probabilistic labels.
3. Snorkel uses these labels to train a discriminative classification model such as a deep neural network.

In one paper that utilized Snorkel [22], its weak supervised learning performance was tested in several ways. First, the authors compared productivity when teaching SMEs to use Snorkel versus spending the equivalent amount of time hand-labeling data. The result was that when they used the Snorkel framework they were able to build models 2.8 times faster and with 45.5% better predictive performance on average.

The second performance evaluation in [22] was based on projects in collaboration with Stanford, the U.S. Department of Veterans Affairs, and the U.S. Food and Drug Administration; in this evaluation, they found that Snorkel led to an average 132% improvement over baseline techniques. In addition to the above examples, the Snorkel framework has

been utilized in many domains. It was used in a study for pain recognition in postoperative patients [23] and to extract observed spatial relations from radiology reports [24] In another study, Snorkel was employed as a weak supervision approach to leverage domain resources and expertise in order to improve clinical natural language processing [24]. The previous examples of Snorkel framework usage demonstrate its effectiveness in different domains, as it enables efficient and effective labeling of datasets and reduces the need for extensive manual annotation by combining weak supervision sources and leveraging domain-specific knowledge.

The third dataset annotation approach is transfer learning. This machine learning technique leverages the knowledge gained from a source domain to improve the learning process in a target domain. Using transfer learning overcomes the challenges of limited annotations, computational limitations, and model generalization with limited data [25].

Zero-shot (ZS) learning is based on transfer learning; it is suitable when no labeled data are provided [26]. The ZS model can predict the class of the unlabeled sample using natural language inference (NLI), even if the model was not trained on those classes. ZS models leverage the semantic similarity between labels and the text context [27]. In natural language inference (NLI) learning, the text is treated as the premise. Next, the hypothesis and the expected labels are used to set the ZS model, where the hypothesis/prompt usually uses the following format: "this example is about {label}". When running the ZS model with the values of the labels, premise, and hypothesis, it returns the entailment score or a confidence level that tells whether or not the premise is related to that label.

To use a ZS models with variant dialects of Arabic, it should support Arabic or multiple languages. Based on [28], which applied the XLM-RoBERTA (XLM-R) model to the cross-lingual natural language inference (XNLI) task for the Arabic language, XLM-R outperformed other models such as mBERT on various cross-lingual benchmarks, including cross-lingual natural language inference. Furthermore, XLM-R was trained using one hundred languages, including Arabic and many other low-resource languages, and it has demonstrated its effectiveness in zero-shot transfer and resource-constrained settings. It enables effective cross-lingual zero-shot transference in natural language processing tasks, reducing the need for extensive labeled data in different languages [29].

Another choice is using multilingual mDeBERTa, a state-of-the-art (SOTA) model, in XNLI tasks. It is the best performing multilingual base-sized transformer model, achieving a 79.8% ZS cross-lingual accuracy for XNLI and a 3.6% improvement over XLM-R Base [30].

## 3. Methodology

### *3.1. Dataset Collection*

To collect the dataset containing tweets related to the COVID-19 pandemic in Kuwait, we implemented the following steps:

1. We manually searched the Twitter platform and collected specific keywords and hashtags associated with Kuwaiti people's attitudes towards the vaccine.
2. We used an online tool, Communalytic [31], along with the Twitter academic API to extract tweets, and we used the collected keywords and hashtags from the previous step to search for historical tweets. The time frame of collection was from the start of the vaccination campaign in Kuwait to the end of all precautions against COVID-19 (December 2020 to July 2022).

### *3.2. Dataset Preparation*

To prepare our dataset and make sure that it only contained tweets from Kuwait, we filtered out tweets that did not have one of the following keywords in the user_location field: Koweït, Q8, kw, kwt, kuwait, وطن النهار ، كويتي ، كويتيه ، الكويت, and KU. We programmatically removed unrelated tweets by excluding all posts not written in the Arabic language or containing keywords related to Arabic spam posts. Next, we cleaned the text of the tweets by removing digits, special characters, URLs, emojis, mentions, tashkīl (diacritics), and punctuation. We did not remove hashtags, as based on our observations of

the dataset hashtags are heavily used to express the stance towards vaccination; instead, we only removed the hash # and underscore _ characters between the hashtag keywords, which allowed the hashtags to be processed as regular text. After the dataset preparation and cleaning process, the total number of extracted unlabeled tweets was 42,815.

### *3.3. Dataset Labeling*

To validate our proposed labeling system, we needed a manually labeled dataset. Two native Kuwaiti dialect speakers from the research team hand-labeled the dataset using an online tool called NLP Annotation Lab [32]. The annotators were able to label 878 tweets out of 2000 extracted tweets that were different from the original dataset and classify them as either anti-vaccine or pro-vaccine. Finally, the two annotators manually checked the labeled dataset for disagreements, revised the labels, and approved the final labels. The distribution of the manually labeled tweets used to validate the Q8VaxStance labeling system was 350 anti-vaccine tweets and 528 pro-vaccine tweets.

### *3.4. Q8VaxStance Labeling System*

Our first research question aimed to investigate whether a weak supervised learning approach combined with the prompt engineering of zero-shot models could label a large dataset of tweets for stance detection towards vaccines with limited help from SMEs. To obtain an answer to our first research question, we performed the following steps:

1. We selected the weak supervised learning framework to use in our experiments. After examining several Python packages and frameworks that support weak supervised learning for natural language processing, we decided to use the Snorkel open-sourced software framework [33] based on the good results we were able to establish in [16] for the sentiment classification of the Kuwaiti dialect.
2. We set up 52 experiments, as described in Table 1; for each experiment, we created the labeling functions that determine the stance towards vaccines. Figure 1 illustrates the general Q8VaxStance labeling system architecture used in the KHZSLF experiment setup; the system architecture for the KHLF and ZSLF experiments is similar, with a few labeling functions being excluded depending on the specific experimental setup.
3. We applied the labeling functions on 42,815 unlabeled tweets and trained the model using the Snorkel package to predict the dataset labels. As a first experiment, we created labeling functions to label the dataset based on the presence of specific pro-vaccine and anti-vaccine keywords and hashtags in the tweet texts. In this experiment, we used the same keywords and hashtags that were used before to obtain the dataset from Twitter.
4. We conducted several experiments to compare the performance of using only zero-shot (ZS) learning-based labeling functions versus combining keyword-based labeling functions with zero-shot learning-based labeling functions. We implemented the inference code provided by the ZS models' creators using the huggingface website. The following pretrained zero-shot models were used in the ZS labeling functions:
   (a) joeddav/xlm-roberta-large-xnli [34].
   (b) MoritzLaurer/mDeBERTa-v3-base-mnli-xnli [35].
   (c) vicgalle/xlm-roberta-large-xnli-anli [36].
5. We applied prompt engineering to check the effect of using different prompts and labels on the labeling system performance, then determined the best labels and prompt combinations that produced the best performance when using the zero-shot learning-based labeling function. To apply prompt engineering, we varied the text of labels and prompts; in addition, we tested different combinations consisting of English labels and prompts, Arabic labels and prompts, and mixed language labels and prompts to check the effect of the language used in the labels and prompts on system performance. Tables 2 and 3 contain a list of the labels and prompts used in our experiments.

**Table 1.** Experimental setup for the labeling functions (LFs) used in Q8VaxStance.

| Experiment | Keywords and Hashtags LFs | Zero-Shot Models LFs | English Prompt | Arabic Prompt | English Labels | Arabic Labels | Count |
|---|---|---|---|---|---|---|---|
| KHLF | ✓ | | | | | | 1 |
| KHZSLF-EE | ✓ | ✓ | ✓ | | ✓ | | 6 |
| KHZSLF-EA | ✓ | ✓ | ✓ | | | ✓ | 9 |
| KHZSLF-AA | ✓ | ✓ | | ✓ | | ✓ | 9 |
| ZSLF-EE | | ✓ | ✓ | | ✓ | | 6 |
| ZSLF-EA | | ✓ | ✓ | | | ✓ | 9 |
| ZSLF-AA | | ✓ | | ✓ | | ✓ | 9 |
| ZSLF-AA-AE-EE | | ✓ | ✓ | ✓ | ✓ | ✓ | 3 |
| | | | | | | Total Experiments | 52 |

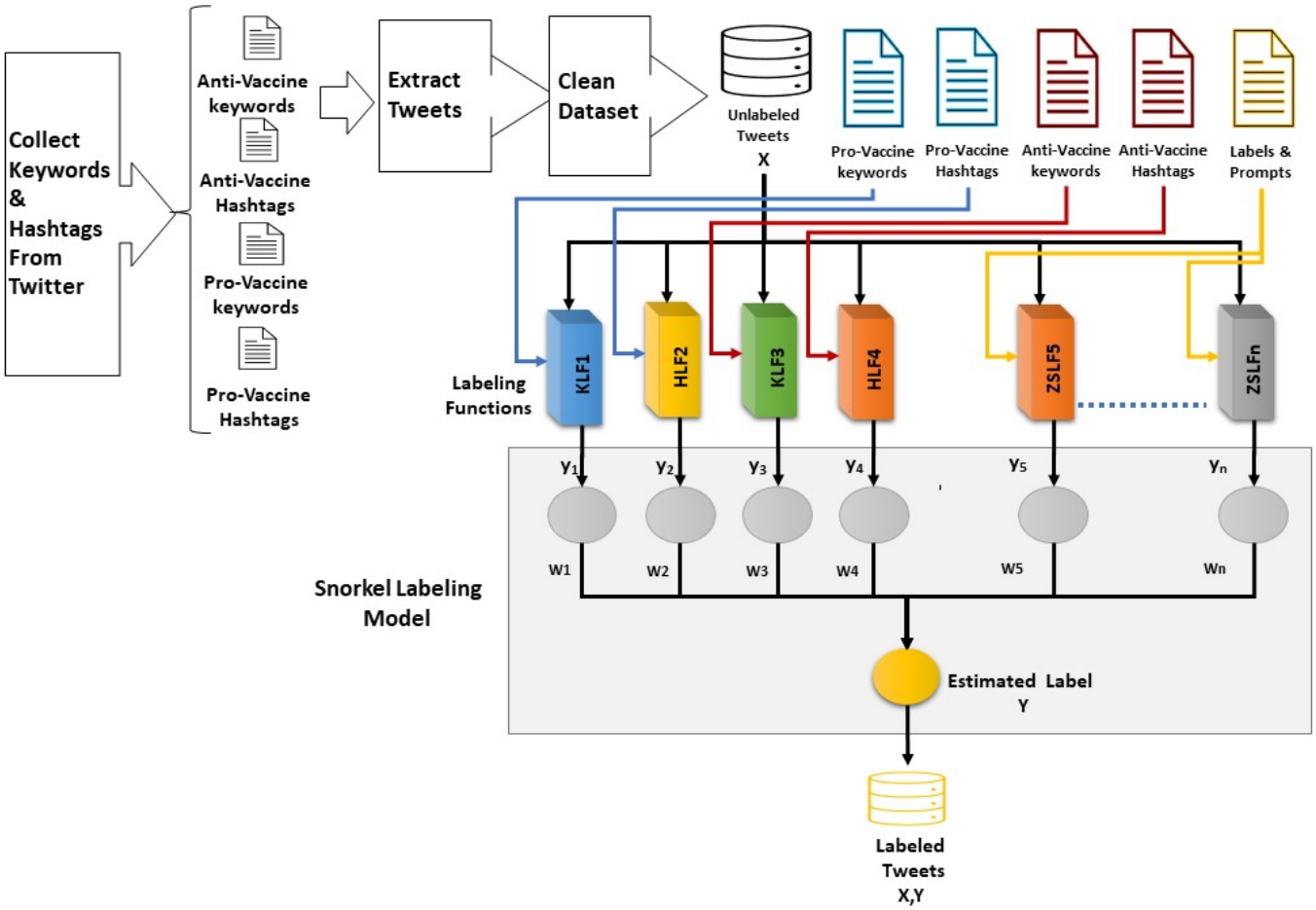


**Figure 1.** Q8VaxStance labeling system architecture used in the KHZSLF experiments.

**Table 2.** List of labels used in zero-shot model labeling functions.

| Labels | Language |
|---|---|
| pro-vaccine, anti-vaccine | English |
| in favor vaccine, against vaccine | English |
| مع التطعيم ، ضد التطعيم | Arabic |
| معارض التطعيم , مؤيد التطعيم | Arabic |
| لا للتطعيم , نعم للتطعيم | Arabic |

**Table 3.** List of prompts used in zero-shot model labeling functions
.

| Prompts | Language |
|---|---|
| the attitude towards COVID-19 vaccination is {} | English |
| the stance towards COVID-19 vaccination is {} | English |
| the opinion towards COVID-19 vaccination is {} | English |
| الرأي في هذه التغريده {} | Arabic |
| الموقف في هذه التغريده تجاه التطعيم {} | Arabic |
| التوجه في هذه التغريده {} | Arabic |

Our second research question aimed to evaluate the performance of the Q8VaxStance system on labeling a large dataset for stance detection towards vaccines. To be able to address this question, we tested the human-labeled dataset using the model we trained using the Snorkel package and the 42,815 unlabeled samples; then, we compared the accuracy, macro-F1 score, and total number of generated labels for each experiment. The details of the experimental results are presented in the next section. Finally, we used ANOVA and Tukey's HSD statistical tests to compare the experiments in order to determine whether they were statistically significant, as well as to discover the main factors affecting the experimental performance and the labeling functions' ability to generate more labels.

## 4. Experimental Results and Discussion

To execute our experiments, we followed the steps presented in Figure 1. We started with tweet extraction using the Twitter academic API; after pre-processing and cleansing, the total number of extracted unlabeled tweets was 42,815. Then, we applied Snorkel labeling functions on the tweets based on each experimental setup, as shown in Table 1 and Figure 1. Next, we used the Snorkel framework to train the labeling model to predict the labels based on the weights of labeling functions. When we trained the SnorkelLabel model, we set the number of epochs and seed values to 100 inside the fit method, and we applied the trained model on the human-annotated dataset to carry out the performance evaluation.

The results of the individual groups of experiments are illustrated in Tables 4–7; comparing the results, it can be observed that the experiments using mixed keywords and zero-shot models for the labeling functions provide very close performance values, with the average accuracy value ranging from 0.80 to 0.82 and the average Macro-F1 score from 0.80 to 0.82. The annotation agreement between the generated labels and labels from the human SMEs detected using the Cohen's kappa score ranged between 0.61 to 0.64, while the annotation agreement values are not in a perfect agreement (the value should be closer to 1). Nonetheless, these values are considered a good level of agreement compared to random chance.

The best accuracy, Macro-F1, and Cohen's kappa score values were achieved in experiments KHZSLF-EE4 and KHZSLF-EA1, with nearly the same accuracy and Macro-F1 values of 0.83 and 0.83, respectively. Likewise, the Cohen's kappa score achieved in these experiments was 0.66 and 0.67. Moreover, the best accuracy for the experiments in the groups using Arabic labels and templates was in experiments KHZSLF-AA8 and KHZSLF-AA9, with accuracy, Macro-F1, Cohen's kappa score values of 0.83, 0.82, and 0.65 respectively.

Next, the results were analyzed to detect which experiments generated a more balanced distribution of the generated dataset labels and which experiments abstained and could not generate many labels. The results show that, on average, the experimental groups KHZSLF-AA, ZSLF-AA, and KHZSLF-EA created nearly balanced datasets. In contrast, experiments KHZSLF-EE, ZSLF-EE, and ZSLF-EA created imbalanced datasets.

We observed that most experiments using only zero-shot models as labeling functions generated more labels than the others. However, although they produced more labels, the average accuracy and Macro-F1 values were lower than in the experiments using mixed keywords and zero-shot models as labeling functions. Furthermore, the average Cohen's kappa score for the experiments using only zero-shot models as labeling functions was between 0.55 and 0.59, indicating a moderate agreement between system-generated

labels and human-generated labels. The details of the results for generated labels in each experiment group are illustrated in Tables 8–11.

**Table 4.** LF-EE experiment results.

| Experiment | Accuracy | Macro-F1 | Cohen Kappa | Experiment | Accuracy | Macro-F1 | Cohen Kappa |
|---|---|---|---|---|---|---|---|
| KHZSLF-EE1 | 0.815 | 0.810 | 0.618 | ZSLF-EE1 | 0.795 | 0.785 | 0.570 |
| KHZSLF-EE2 | 0.802 | 0.798 | 0.598 | **ZSLF-EE2** | **0.803** | **0.789** | **0.579** |
| KHZSLF-EE3 | 0.824 | 0.820 | 0.638 | ZSLF-EE3 | 0.795 | 0.780 | 0.561 |
| **KHZSLF-EE4** | **0.839** | **0.834** | **0.668** | ZSLF-EE4 | 0.775 | 0.766 | 0.533 |
| KHZSLF-EE5 | 0.822 | 0.817 | 0.633 | ZSLF-EE5 | 0.779 | 0.765 | 0.532 |
| KHZSLF-EE6 | 0.825 | 0.821 | 0.640 | ZSLF-EE6 | 0.784 | 0.768 | 0.538 |
| Average | 0.820 | 0.820 | 0.633 | Average | 0.790 | 0.780 | 0.552 |

Bold text represents the best performance achieved by experiments in each group.

**Table 5.** LF-AA experiment results.

| Experiment | Accuracy | Macro-F1 | Cohen Kappa | Experiment | Accuracy | Macro-F1 | Cohen Kappa |
|---|---|---|---|---|---|---|---|
| KHZSLF-AA1 | 0.820 | 0.810 | 0.621 | ZSLFAA1 | 0.776 | 0.760 | 0.536 |
| KHZSLF-AA2 | 0.809 | 0.804 | 0.609 | ZSLFAA2 | 0.780 | 0.777 | 0.558 |
| KHZSLF-AA3 | 0.826 | 0.820 | 0.641 | ZSLFAA3 | 0.795 | 0.783 | 0.568 |
| KHZSLF-AA4 | 0.810 | 0.801 | 0.602 | ZSLFAA4 | 0.775 | 0.770 | 0.541 |
| KHZSLF-AA5 | 0.790 | 0.786 | 0.573 | ZSLFAA5 | 0.792 | 0.788 | 0.578 |
| KHZSLF-AA6 | 0.815 | 0.811 | 0.623 | ZSLFAA6 | 0.790 | 0.784 | 0.570 |
| KHZSLF-AA7 | 0.808 | 0.797 | 0.596 | ZSLFAA7 | 0.810 | 0.803 | 0.606 |
| **KHZSLF-AA8** | **0.832** | **0.826** | **0.652** | **ZSLFAA8** | **0.824** | **0.818** | **0.636** |
| **KHZSLF-AA9** | **0.832** | **0.828** | **0.657** | ZSLFAA9 | 0.810 | 0.802 | 0.604 |
| Average | 0.816 | 0.809 | 0.619 | Average | 0.795 | 0.787 | 0.577 |

Bold text represents the best performance achieved by experiments in each group.

**Table 6.** LF-EA experiment results.

| Experiment | Accuracy | Macro-F1 | Cohen Kappa | Experiment | Accuracy | Macro-F1 | Cohen Kappa |
|---|---|---|---|---|---|---|---|
| **KHZSLF-EA1** | **0.839** | **0.836** | **0.673** | ZSLF-EA1 | 0.792 | 0.788 | 0.578 |
| KHZSLF-EA2 | 0.833 | 0.833 | 0.660 | ZSLF-EA2 | 0.808 | 0.802 | 0.605 |
| KHZSLF-EA3 | 0.832 | 0.828 | 0.659 | ZSLF-EA3 | 0.801 | 0.796 | 0.594 |
| KHZSLF-EA4 | 0.799 | 0.796 | 0.594 | ZSLF-EA4 | 0.787 | 0.781 | 0.562 |
| KHZSLF-EA5 | 0.823 | 0.819 | 0.639 | ZSLF-EA5 | 0.794 | 0.788 | 0.576 |
| KHZSLF-EA6 | 0.807 | 0.803 | 0.608 | ZSLF-EA6 | 0.784 | 0.777 | 0.556 |
| KHZSLF-EA7 | 0.825 | 0.821 | 0.644 | **ZSLF-EA7** | **0.809** | **0.803** | **0.606** |
| KHZSLF-EA8 | 0.837 | 0.833 | 0.667 | ZSLF-EA8 | 0.807 | 0.800 | 0.601 |
| KHZSLF-EA9 | 0.829 | 0.824 | 0.649 | ZSLF-EA9 | 0.801 | 0.796 | 0.593 |
| Average | 0.825 | 0.821 | 0.643 | Average | 0.798 | 0.792 | 0.58 |

Bold text represents the best performance achieved by experiments in each group.

**Table 7.** LF-AA-AE-EE experiment results.

| Experiment | Accuracy | Macro-F1 | Cohen Kappa |
|---|---|---|---|
| ZSLF-AA-AE-EE1 | 0.804 | 0.799 | 0.599 |
| **ZSLF-AA-AE-EE2** | **0.805** | **0.800** | **0.601** |
| ZSLF-AA-AE-EE3 | 0.802 | 0.798 | 0.596 |
| Average | 0.804 | 0.799 | 0.598 |

Bold text represents the best performance achieved by experiments in each group.

**Table 8.** Count of pro-vaccine labels vs. anti-vaccine labels: KHZSLF-AA experiments.

| Experiment | Pro | Anti | Total | Experiment | Pro | Anti | Total |
|---|---|---|---|---|---|---|---|
| KHZSLF-AA1 | 30,034 | 12,775 | 42,809 | ZSLF-AA1 | 25,439 | 17,318 | 42,757 |
| KHZSLF-AA2 | 21,503 | 21,312 | 42,815 | ZSLF-AA2 | 15,727 | 27,086 | 42,813 |
| KHZSLF-AA3 | 26,710 | 16,092 | 42,802 | ZSLF-AA3 | 30,191 | 12,609 | 42,800 |
| KHZSLF-AA4 | 19,543 | 23,253 | 42,796 | ZSLF-AA4 | 18,348 | 24,373 | 42,721 |
| KHZSLF-AA5 | 18,218 | 24,594 | 42,812 | ZSLF-AA5 | 20,431 | 22,380 | 42,811 |
| KHZSLF-AA6 | 18,494 | 23,535 | 42,029 | ZSLF-AA6 | 21,907 | 20,851 | 42,758 |
| KHZSLF-AA7 | 19,049 | 23,738 | 42,787 | ZSLF-AA7 | 20,209 | 22,606 | 42,815 |
| KHZSLF-AA8 | 18,929 | 23,838 | 42,767 | ZSLF-AA8 | 20,390 | 22,425 | 42,815 |
| KHZSLF-AA9 | 20,928 | 21,869 | 42,797 | ZSLF-AA9 | 18,276 | 24,539 | 42,815 |
| Average | 21,490 | 21,223 | 42,713 | Average | 21,213 | 21,576 | 42,789 |

**Table 9.** Count of pro-vaccine labels vs. anti-vaccine labels: KHZSLF-EE experiments.

| Experiment | Pro | Anti | Total | Experiment | Pro | Anti | Total |
|---|---|---|---|---|---|---|---|
| KHZSLF-EE1 | 20,552 | 21,702 | 42,254 | ZSLF-EE1 | 23,666 | 18,896 | 42,562 |
| KHZSLF-EE2 | 17,856 | 23,781 | 41,637 | ZSLF-EE2 | 26,931 | 15,621 | 42,552 |
| KHZSLF-EE3 | 20,925 | 21,262 | 42,187 | ZSLF-EE3 | 15,123 | 27,326 | 42,449 |
| KHZSLF-EE4 | 22,292 | 19,978 | 42,270 | ZSLF-EE4 | 21,195 | 16,743 | 37,938 |
| KHZSLF-EE5 | 19,938 | 22,115 | 42,053 | ZSLF-EE5 | 25,976 | 13,328 | 39,304 |
| KHZSLF-EE6 | 18,385 | 23,124 | 41,509 | ZSLF-EE6 | 24,551 | 12,668 | 37,219 |
| Average | 19,991 | 21,994 | 41,985 | Average | 22,907 | 17,430 | 40,337 |

**Table 10.** Count of pro-vaccine labels vs. anti-vaccine labels: KHZSLF-EA experiments.

| Experiment | Pro | Anti | Total | Experiment | Pro | Anti | Total |
|---|---|---|---|---|---|---|---|
| KHZSLF-EA1 | 23,102 | 19,662 | 42,764 | ZSLF-EA1 | 25,409 | 17,406 | 42,815 |
| KHZSLF-EA2 | 23,027 | 19,765 | 42,792 | ZSLF-EA2 | 25,101 | 17,714 | 42,815 |
| KHZSLF-EA3 | 20,647 | 22,117 | 42,764 | ZSLF-EA3 | 23,802 | 19,013 | 42,815 |
| KHZSLF-EA4 | 20,573 | 22,213 | 42,786 | ZSLF-EA4 | 20,156 | 22,659 | 42,815 |
| KHZSLF-EA5 | 21,577 | 21,177 | 42,754 | ZSLF-EA5 | 22,617 | 20,198 | 42,815 |
| KHZSLF-EA6 | 19,393 | 22,052 | 41,445 | ZSLF-EA6 | 22,913 | 19,902 | 42,815 |
| KHZSLF-EA7 | 22,189 | 20,594 | 42,783 | ZSLF-EA7 | 21,912 | 20,903 | 42,815 |
| KHZSLF-EA8 | 20,240 | 22,538 | 42,778 | ZSLF-EA8 | 21,102 | 21,713 | 42,815 |
| KHZSLF-EA9 | 23,212 | 19,580 | 42,792 | ZSLF-EA9 | 21,695 | 21,120 | 42,815 |
| Average | 21,551 | 21,078 | 42,629 | Average | 22,745 | 20,070 | 42,815 |

**Table 11.** Count of pro-vaccine labels vs. anti-vaccine labels: ZSLF-AA-AE-EE experiments.

| Experiment | Pro | Anti | Total |
|---|---|---|---|
| ZSLF-AA-AE-EE1 | 21,894 | 20,921 | 42,815 |
| ZSLF-AA-AE-EE2 | 21,650 | 21,165 | 42,815 |
| ZSLF-AA-AE-EE3 | 22,432 | 20,383 | 42,815 |
| Average | 21,992 | 20,823 | 42,815 |

Next, because the results of many experiments had very close performance values, we checked the statistical significance of the experiments in order to identify the experiments that performed better and detect the main factors affecting the performance of the experiments and the generated labels. To achieve this, we applied ANOVA and pairwise Tukey's HSD post hoc statistical tests. Table 12 illustrates the ANOVA test $p$-value results, while Tables 13 and 14 show the adjusted $p$-value results for each experiment group based on changing the type of labeling function and changing the language of labels and prompts used in the zero-shot models.

The following is a description of each experimental group:

- Changing the type of labeling function:

  - KHLF: keyword and hashtag detection used in labeling functions;
  - ZSLF: only zero-shot models used in labeling functions;
  - KHZSLF: both keyword and hashtag detection plus zero-shot models used in labeling functions.
- Changing the language of labels and prompts used in zero-shot models:
  - AA: Arabic labels and Arabic prompts;
  - EE: English labels and English prompts;
  - AE: Arabic labels and English prompts;
  - AAAEEE: mixed labeling function with mixed language labels and prompts;
  - NN: not using zero-shot models as labeling functions, i.e., using keyword and hashtag detection in labeling functions.

**Table 12.** *p*-value results of ANOVA test.

| *p*-Value | Keywords vs. Zero-Shot | Language of Labels and Prompts |
|---|---|---|
| Accuracy | $1.577262 \times 10^{-11}$ | 0.000386 |
| Macro-F1 | $6.632477 \times 10^{-12}$ | 0.000359 |
| Total Labels | $1.397020 \times 10^{-28}$ | $2.697203 \times 10^{-30}$ |

**Table 13.** Results of adjusted *p*-value for Tukey's HSD post hoc test on the effect of changing the type of labeling function.

| Experiment Group 1 | Experiment Group 2 | P-adj Accuracy | P-adj Macro-F1 | P-adj Labels |
|---|---|---|---|---|
| KHLF | KHZSLF | 0.0 | 0.0 | 0.0 |
| KHLF | ZSLF | 0.0 | 0.0 | 0.0 |
| KHZSLF | ZSLF | 0.0 | 0.0 | 0.5 |

**Table 14.** Results of adjusted *p*-value for Tukey's HSD post hoc test on the effect of changing the the language used in labels and prompts of zero-shot models.

| Experiment Group 1 | Experiment Group 2 | P-adj Accuracy | P-adj Macro-F1 | P-adj Labels |
|---|---|---|---|---|
| AA | AAAEEE | 0.9999 | 1.0 | 1.0 |
| AA | AE | 0.8400 | 0.6911 | 1.0 |
| AA | EE | 1.0 | 0.9986 | **0.0004** |
| AA | NN | **0.0001** | **0.0001** | 0 |
| AAAEEE | AE | 0.9583 | 0.9688 | 0.9999 |
| AAAEEE | EE | 1.0 | 0.9994 | 0.0667 |
| AAAEEE | NN | **0.0005** | **0.0009** | 0 |
| AE | EE | 0.8634 | 0.6021 | **0.0005** |
| AE | NN | **0.0003** | **0.0006** | 0 |
| EE | NN | **0.0001** | **0.0001** | 0 |

Bold values represents statically significant experiments.

As presented in Table 12, the ANOVA test results show that using keyword detection vs. zero-shot models as labeling functions and changing the language of labels and templates used in zero-shot models is statistically significant at a significance level of 0.05 in regard to the accuracy, macro-F1, and total number of labels predicted by the model.

Furthermore, the Tukey's HSD post hoc test results in Table 13 show that when using zero-shot models and keyword detection as labeling functions (KHZSLF), the experiments had significantly better performance than when using only the keyword detection labeling functions (KHLF) or using only the zero-shot model labeling functions (ZSLF) for all three evaluation metrics (accuracy, macro-averaged F1 score, and total number of labels). In addition, the results shows that there is no significant statistical difference between the

total generated labels when using keyword and zero-shot models (KHZSLF) compared to using only zero-shot models as labeling functions (ZSLF).

Table 14 illustrates the results when changing the language used in labels and prompts in zero-shot models; the results show that the total number of generated labels is affected when using Arabic in both labels and prompts (AA) or mixed Arabic and English labels and prompts (AE). The effect is statistically significant; more labels are generated than when using English language in both labels and prompts (EE).

Furthermore, the results indicate that there is a statistically significant difference between the means of the three evaluation metrics (accuracy, macro-averaged F1 score, and the total number of labels) when using zero-shot model labeling functions with any language (AA, AE, or EE) compared to not using zero-shot models (NN), indicating that experiments using zero-shot model labeling functions outperform experiments using only keyword labeling functions.

Therefore, we can conclude that when using mixed zero-shot models with mixed language labels and prompts (AAAEEE), the differences between the experiments are not statistically significant compared to using only zero-shot models, indicating that this experimental setup does not significantly improve the evaluation metrics.

## 5. Conclusions

In this study, we have attempted to fill a gap in the field of NLP by creating Kuwaiti dialect language resources, as currently the Kuwaiti dialect is underrepresented in the available Arabic language models. These language resources are critical for developing high-performance approaches and systems for different NLP problems. To overcome data annotation challenges, we have proposed an automated system to programmatically label a tweet dataset to detect the stance towards vaccines in the Kuwaiti dialect (Q8VaxStance). The proposed system is based on an approach combining the benefits of weak supervised learning and zero-shot learning.

This research is an essential part of a more comprehensive attempt to understand the elements that cause vaccine hesitancy in Kuwait and to create practical approaches for addressing it. This labeled dataset is the first Kuwaiti dialect dataset for vaccine stance detection. In this research, we conducted 52 experiments to identify the best experimental setup and the main factors that affect the annotation system's performance metrics by comparing the accuracy value, Macro-F1 score, Cohen's kappa score, and total number of generated labels. In addition, we studied the statistical significance of the experiments by applying ANOVA and pairwise Tukey's HSD post hoc statistical tests.

Based on our results, we achieved the best accuracy, Macro-F1 score, and Cohen's kappa score values in the experiments when using both zero-shot models and keyword detection as labeling functions; experiments KHZSLF-EE4 and KHZSLF-EA1 had nearly the same accuracy, and had Macro-F1 scores of 0.83 and 0.83, respectively. The Cohen's kappa scores achieved in these experiments were 0.66 and 0.67, respectively, which are considered good annotator agreement scores. As part of our future work, we plan to conduct additional experimentation and refinement in order to achieve perfect agreement and improved performance metrics.

The results of the ANOVA and pairwise Tukey's HSD post hoc statistical tests showed that the experiments using both zero-shot models and keyword detection as labeling functions (KHZSLF) significantly outperformed those using only the keyword detection labeling functions (KHLF) or only the zero-shot models labeling functions (ZSLF) for all evaluation metrics. When changing the language of the labels and prompts used in zero-shot models, our results showed that the mean total number of generated labels when using Arabic in both labels and prompts (AA) or mixed Arabic English labels (AE) and prompts was statistically significant compared to using English in both labels and prompts (EE), indicating that our proposed annotation system generates more labels when the Arabic language is used in both prompts and labels or in at least one of them.

In our future research, we first intend to experiment more with the proposed annotation system by applying zero-shot and few-shot learning on large language models supporting the Arabic language. Second, we plan to use this generated dataset to fine-tune and compare available Arabic BERT-based language models and large multilingual models to create a trained model for Kuwaiti dialect stance detection. Finally, we plan to use graph neural network algorithms to predict vaccine stances and compare the findings with the results of this research.

**Author Contributions:** Conceptualization, H.A. and H.D; methodology, H.A.; software, H.A.; validation, H.A; formal analysis, H.A; investigation, H.A.; data curation, S.D.; writing—original draft preparation, H.A.; writing—review and editing: H.A., S.D. and H.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Datasets and the list of keywords and hashtags used in this research are publicly available at the following link: https://github.com/hanaalostad/Q8Stance, accessed on 4 September 2023.

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

## References

1. Alibrahim, J.; Awad, A. COVID-19 vaccine hesitancy among the public in Kuwait: A cross-sectional survey. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8836. [PubMed]
2. Sallam, M.; Dababseh, D.; Eid, H.; Al-Mahzoum, K.; Al-Haidar, A.; Taim, D.; Yaseen, A.; Ababneh, N.A.; Bakri, F.G.; Mahafzah, A. High Rates of COVID-19 Vaccine Hesitancy and Its Association with Conspiracy Beliefs: A Study in Jordan and Kuwait among Other Arab Countries. *Vaccines* **2021**, *9*, 42. [CrossRef] [PubMed]
3. Al-Ayyadhi, N.; Ramadan, M.M.; Al-Tayar, E.; Al-Mathkouri, R.; Al-Awadhi, S. Determinants of hesitancy towards COVID-19 vaccines in State of Kuwait: An exploratory internet-based survey. *Risk Manag. Healthc. Policy* **2021**, *14*, 4967–4981. [CrossRef]
4. Cascini, F.; Pantovic, A.; Al-Ajlouni, Y.A.; Failla, G.; Puleo, V.; Melnyk, A.; Lontano, A.; Ricciardi, W. Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *eClinicalMedicine* **2022**, *48*, 101454. [CrossRef]
5. Greyling, T.; Rossouw, S. Positive attitudes towards COVID-19 vaccines: A cross-country analysis. *PLoS ONE* **2022**, *17*, e0264994. [CrossRef] [PubMed]
6. AlAwadhi, E.; Zein, D.; Mallallah, F.; Bin Haider, N.; Hossain, A. Monitoring COVID-19 vaccine acceptance in Kuwait during the pandemic: Results from a national serial study. *Risk Manag. Healthc. Policy* **2021**, *14*, 1413–1429. [CrossRef]
7. Putra, C.B.P.; Purwitasari, D.; Raharjo, A.B. Stance Detection on Tweets with Multi-task Aspect-based Sentiment: A Case Study of COVID-19 Vaccination. *Int. J. Intell. Eng. Syst.* **2022**, *15*, 515–526.
8. Muric, G.; Wu, Y.; Ferrara, E. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health Surveill* **2021**, *7*, e30642. [CrossRef]
9. Hayawi, K.; Shahriar, S.; Serhani, M.; Taleb, I.; Mathew, S. ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* **2022**, *203*, 23–30. [CrossRef]
10. Jun, J.; Zain, A.; Chen, Y.; Kim, S.H. Adverse Mentions, Negative Sentiment, and Emotions in COVID-19 Vaccine Tweets and Their Association with Vaccination Uptake: Global Comparison of 192 Countries. *Vaccines* **2022**, *10*, 735. [CrossRef]
11. Moubtahij, H.E.; Abdelali, H.; Tazi, E.B. AraBERT transformer model for Arabic comments and reviews analysis. *IAES Int. J. Artif. Intell. (IJ-AI)* **2022**, *11*, 379–387. [CrossRef]
12. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. 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), Bangkok, Thailand, 1 August 2021; pp. 7088–7105. [CrossRef]
13. Obeid, O.; Zalmout, N.; Khalifa, S.; Taji, D.; Oudah, M.; Alhafni, B.; Inoue, G.; Eryani, F.; Erdmann, A.; Habash, N. CAMeL tools: An open source python toolkit for Arabic natural language processing. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 7022–7032.
14. Salamah, J.B.; Elkhlifi, A. Microblogging opinion mining approach for kuwaiti dialect. In Proceedings of the The International Conference on Computing Technology and Information Management (ICCTIM), Dubai, United Arab Emirates, 9 April 2014; p. 388.
15. Almatar, M.G.; Alazmi, H.S.; Li, L.; Fox, E.A. Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. *ISPRS Int. J.-Geo-Inf.* **2020**, *9*, 702. [CrossRef]
16. Husain, F.; Al-Ostad, H.; Omar, H. A Weak Supervised Transfer Learning Approach for Sentiment Analysis to the Kuwaiti Dialect. In Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8 December 2022; pp. 161–173.

17. Aldihan, H.; Gaizauskas, R.; Fitzmaurice, S. A Pilot Study on the Collection and Computational Analysis of Linguistic Differences Amongst Men and Women in a Kuwaiti Arabic WhatsApp Dataset. In Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), Abu Dhabi, United Arab Emirates, 8 December 2022; pp. 372–380.
18. Shimizu, A.; Wakabayash, K. Effect of Label Redundancy in Crowdsourcing for Training Machine Learning Models. *J. Data Intell.* **2022**, *3*, 301–315. [CrossRef]
19. Zhang, Z.; Strubell, E.; Hovy, E. A Survey of Active Learning for Natural Language Processing. *arXiv* **2022**, arXiv:2210.10109. https://doi.org/10.48550/arxiv.2210.10109.
20. Simmler, N.; Sager, P.; Andermatt, P.; Chavarriaga, R.; Schilling, F.P.; Rosenthal, M.; Stadelmann, T. A Survey of Un-, Weakly-, and Semi-Supervised Learning Methods for Noisy, Missing and Partial Labels in Industrial Vision Applications. In Proceedings of the 2021 8th Swiss Conference on Data Science (SDS), Lucerne, Switzerland, 9 June 2021. [CrossRef]
21. Hang, D.; Victor, S.P.; Huayu, Z.; Minhong, W.; Arlene, C.; Emma, D.; Jiaoyan, C.; Beatrice, A.; William, W.; Honghan, W. Ontology-Driven and Weakly Supervised Rare Disease Identification From Clinical Notes. *BMC Med Inform. Decis. Mak.* **2023**, *23*, 86.
22. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid training data creation with weak supervision. *Proc. Vldb Endow.* **2017**, *11*, 269–282. [CrossRef]
23. Naeini, E.K.; Subramanian, A.; Calderon, M.D.; Zheng, K.; Dutt, N.; Liljeberg, P.; Salanterä, S.; Nelson, A.M.; Rahmani, A.M. Pain Recognition With Electrocardiographic Features in Postoperative Patients: Method Validation Study. *J. Med. Internet Res.* **2021**, *23*, e25079. [CrossRef]
24. Datta, S.; Roberts, K. Weakly Supervised Spatial Relation Extraction From Radiology Reports. *JAMIA Open* **2023**, *6*, ooad027. [CrossRef]
25. Yu, F.; Xiu, X.; Li, Y. A survey on deep transfer learning and beyond. *Mathematics* **2022**, *10*, 3619. [CrossRef]
26. Tunstall, L.; von Werra, L.; Wolf, T. *Natural Language Processing with Transformers*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.
27. Yildirim, S.; Asgari-Chenaghlu, M. *Mastering Transformers: Build State-of-the-Art Models from Scratch with Advanced Natural Language Processing Techniques*; Packt Publishing: Birmingham, UK, 2021.
28. Ranasinghe, T.; Zampieri, M. An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India. *Information* **2021**, *12*, 306. [CrossRef]
29. Kuo, C.C.; Chen, K.Y. Toward Zero-Shot and Zero-Resource Multilingual Question Answering. *IEEE Access* **2022**, *10*, 99754–99761. [CrossRef]
30. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv* **2021**, arXiv:2111.09543. https://doi.org/10.48550/arxiv.2111.09543.
31. Gruzd, A.; Mai, P. Communalytic: A Research Tool For Studying Online Communities and Online Discourse. 2022. Available online: https://communalytic.org/ (accessed on 14 September 2023).
32. 2022. Available online: https://nlp.johnsnowlabs.com/docs/en/alab/quickstart (accessed on 14 September 2023).
33. Ratner, A.; De Sa, C.; Wu, H.; Davison, D.; Wu, X.; Liu, Y. Language Models in the Loop: Incorporating Prompting into Weak Supervision. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1776–1787.
34. Davison, J. XLM-Roberta-Large-XNLI. Available online: https://huggingface.co/joeddav/xlm-roberta-large-xnli (accessed on 4 September 2023).
35. Laurer, M.; van Atteveldt, W.; Casas, A.; Welbers, K. Less Annotating, More Classifying–Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Anal.* **2022**, 1–33. [CrossRef]
36. Gallego, V. XLM-RoBERTa-Large-XNLI-ANLI. Available online: https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli. (accessed on 4 September 2023).