Use of NLP in the Context of Belief states of Ethnic Minorities in Latin America

Olga Kellert and **Md Mahmud Uz Zaman**
University of Göttingen, Germany
olga.kellert@phil.uni-goettingen.de and
mail.mahmuduzzaman@gmail.com

Abstract

The major goal of our study is to test methods in NLP in the domain of health care education related to Covid-19 of vulnerable groups such as indigenous people from Latin America. In order to achieve this goal, we asked participants in a survey questionnaire to provide answers about health related topics. We used these answers to measure the health education status of our participants. In this paper, we summarize the results from our NLP-application on the participants' answers. In the first experiment, we use embeddings-based tools to measure the semantic similarity between participants' answers and "expert" or "reference" answers. In the second experiment, we use synonym-based methods to classify answers under topics. We compare the results from both experiments with human annotations. Our results show that the tested NLP-methods reach a significantly lower accuracy score than human annotations in both experiments. We explain this difference by the assumption that human annotators are much better in pragmatic inferencing necessary to classify the semantic similarity and topic classification of answers.

1 Introduction

Indigenous people belong to the particularly vulnerable groups in the COVID-19 era and are disproportionally affected by epidemics and other crises, as acknowledged by the United Nations (United Nations and Affairs, 2020). Beyond the general problems related to the socio-economic marginalization and the concomitant inaccessibility of health-care services (in particular in rural regions and remote communities), a major threat for indigenous people arises through miscommunication, either due to the sparsity of information material in indigenous languages or due to cultural differences hindering the interpretation/application of the recommended health measures(García et al., 2020) (Afifi et al., 2020). Dissemination of reliable COVID-19- related information, adapted to cultural and linguistic background of indigenous peoples, is a major priority in epidemic crisis; (García et al., 2020) (Afifi et al., 2020) (UN, 13 April 2020). Several initiatives of the European Union (EU) and World Health Organization (WHO) address the problems in communication of health related information (Baccolini, 2021). These initiatives target communication of key health-related terms and concepts underlying them such as understanding of medical instructions. In the recent covid pandemic, it was documented that misconceptions about preventive measures against the spread of covid had a strong impact on the severity of the pandemic (UN, 13 April 2020). In order to reduce health-illiteracy and avoid unnecessary spread of infectious diseases, it is necessary to observe people's understandings of infectious diseases and their treatments. For instance, some individuals have the perception that antibiotics are a "cure-all" drug and might take antibiotics to cure diseases caused by viruses, which is an improper use of antibiotics and can lead to severe damaging effects(Calderón-Parra J, 2021).

Given the urgency of measuring the accuracy of health-related concepts and uses, it is necessary to develop NLP tools that can ease and speed up the process related to health education measurement. The key outcome of our research project is testing NLP methodology targeting measurement of health education related to the COVID-19 pandemics.

2 State-of-the-art

Accuracy measurement of medical terms uses like *antibiotics* is **currently missing** due to two main reasons: a) missing **data sources and methodologies** that enable researchers to identify, characterize and measure **actual** uses of health related topics and concepts and b) **missing statistical** (in)accuracy measures of actual information status related to infectious diseases. It is thus not surprising that the initiative *the Social Media Mining 4*

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 1–5

Health (#SMM4H) is addressing these problems in its agenda(Klein, 2021) (Magge et al., 2021). This initiative uses social media data as a data source for solving health-related tasks and problems such as finding disease mentions and symptoms(Klein, 2021) (Magge et al., 2021) (Weissenbacher et al., 2019). However, this rich data source does not have demographic information necessary for the statistics on social variation in the health literacy study. In addition, social media does not represent all social groups including indigenous population that often has low internet access or uses other tools for communication. As a consequence, data from indigenous communities related to Covid pandemics is very rare (Ojha et al., 2021). In order to address these problems, we used a traditional methodology in social sciences in order to access the information about the health education status, namely the survey methodology. We asked health-related questions such as questions about virus propagation and treatment to our participants. In order to be able to measure the accuracy of health-related concepts and uses of our participants', it is necessary to compare their information status with "expert" knowledge or uses.
In recent years, big progress has been made in semantic comparison of linguistic units such as words and sentences due to recent developments in **neural language models** such as BERT(Devlin et al., 2019) (Giulianelli et al., 2020). BERT is a language model trained on a large amount of natural language data to predict words that have been masked out as shown in Table 1 for the word *coach* (Devlin et al., 2019).

BERT has been used to find out which word vectors are responsible for lexical meaning variation such as *coach* used as 'trainer' and 'vehicle'. A word vector is essentially a mathematical representation of the meaning of a word based on learning or memorizing the frequency at which a word appears in a particular linguistic context. The differences or similarities of word vectors have been used to predict semantic (dis)similarity of words (Giulianelli et al., 2020) and sentences (Reimers and Gurevych, 2020). However, previous approaches mainly focus on meaning differences in Big Data sources such as social media and very few of them address meaning differences in survey questionnaires of ethnic minorities. It is thus not known yet how well these models work in the low resource scenario given the specific topic domain and the specific format of answers. This paper presents results from testing vector-based approaches in the measurement of answer similarity in the low resource domain.

3 Methodology

We carried out a survey study with our cooperation partners from Latin America (Marleen Haboud, Claudia Crespo, Fernando Ortega Pérez), in which indigenous groups speaking Quechua or Kichwa from Peru and Ecuador (around 150 people from each country) answered questions about Covid-19 (10 yes-no questions and 10 open-ended questions). Our task was to measure the accuracy of key concepts related to health. We tested how well the information status of indigenous groups matches the information and suggestions from reliable sources such as the World Health Organization (WHO), henceforth our Reference Corpus. For instance, according to the WHO, the virus COVID-19 is distributed through contact, hence the suggestion to keep social distancing. We asked our participants about how the virus COVID-19 is distributed in order to see how well their answer matches the information from WHO. The answers were collected in rural areas via free interviews by a local person knowing indigenous communities. The method of free interviews was particularly important in order to include individuals who are less accustomed to performing highly controlled tasks such as older and/or illiterate participants. Due to lack of time and resources we did not transcribe the interviews. Instead, the local interviewer summarized the answers to the questions in a digital form in Spanish. Consequently, the answers in this survey study do not **directly** reflect the information state of indigenous minorities.

4 Experiments and Results

We ran two experiments. The data and the code for both experiments can be found on GitHub[1]. In our first experiment, we tested the SBERT Model for measuring the semantic similarity between the participants' answers and the "expected" answers from the reference corpus via cosine similarity (see Sentence Transformers based on Reimers and Gurevych, 2020). The following examples demonstrate some results of cosine similarity from the chosen method:

[1] https://github.com/mahmuduzzamanDE/ACLAmericaNLP

Word	Before mask	After mask
coach 'vehicle'	I have driven my coach into the garage.	I have driven my <mask> into the garage.
coach 'trainer'	I have a female coach.	I have a female <mask>.

Table 1: MASK TASK

Question : 8. When should a mask be used?
Reference text : Especially in closed public places, but it is also useful in outdoor public places."
Answers by participants:
"['Whenever we are in contact with another person.'] # participant 1
"Similarity: tensor([[0.1775]])", # similarity between reference text and participant 1

"['All the time when leaving home.'] # participant 2
"Similarity: tensor([[0.0477]])", # similarity between reference text and participant 2

"['Especially in closed public places, but it is also useful in outdoor public places']", "Similarity: tensor([[0.9961]])", match between reference text and reference text

"['When we are in public places where social distancing cannot be maintained.']", participant 3
"Similarity: tensor([[0.2265]])", # similarity between reference text and participant 3

In order to evaluate the validity of the similarity measure by SBERT, we asked human annotators to annotate participants' answers from 0-5 as not similar (0) or similar (5). The annotators were four students of linguistics and one expert in medical anthropology. We divided the human ratings into three categories: similar (4-5), dissimilar (0-2), ambiguous (3) and selected the answers with high inter-speaker agreement. We translated the human ratings into correspondent cosine similarity scores: similar (>0.6), dissimilar (<0.4), ambiguous (> 0.4 and < 0.6). Our results show that the semantic similarity measured by cosine similarity using SBERT is significantly lower **(mean 0.2)** than the semantic similarity acquired by human annotation **(mean 0.7)**.

Our second experiment had the goal to find a computational method to classify a topic of an answer to an open-ended question. Here is an example. Survey question: Why do you not want to be vaccinated? Topics: a) afraid of side effects, b) my own decision, c).... An automatic classification of answers under the correspondent topics can ease the process of survey data analysis and provide a uniform way of measuring answers to open-ended questions. We asked human annotators to create topics for the interview questions and then to annotate answers according to these topics, e.g. "I can get thrombosis" was classified by human annotators as a) afraid of side effects.

We tested automatic methods to classify answers under suggested topics. The underlying idea was to look for key words in the answers that semantically correspond to suggested topics. For this aim, we performed a synonym-based similarity task without stemming (Task 1) and with stemming (Task 2). In the first task, if the topic was a synonym of one of the tokens in the given answer, the classification was TRUE. In the second task, if the topic stem was a synonym of the token stem in the given answer, the classification was TRUE. The latter case ignores morphological variation of words and focuses only on the lexical stem. We preprocessed the given answers by tokenization, removing stop words and case lowering. The synonyms were taken from the NLTK wordnet.

```
  print(set(synonyms))
{'impinging','contact','reach',
'get_through',                    'inter-
group_communication'    ,'contact_lens',
...}
```

We used a Stemmer from NLTK, to stem the synonym words:

```
   print(Stem)
{contact|saliv|aglomer|tos|segur|
mascarill|distanci|comun|familiar|
friccion|intim|relacion|roc|
tocamient|....}
```

Table 2 demonstrates which answers the synonym-based approach by stemming correctly identified and which answers the system did not correctly identify.

Our results in Table 3 show that stemming gives slightly better results than the absence of stemming, namely a correct classification of additional 10 answers. However, despite this light improvement, the accuracy is still very low, or more precisely, the system could not make a link between a given

Used Sentence (Spanish)	Translated (English)	w/o Stem	Stem
por no seguir medidas de bioseguridad	*for not maintaining social distancing*	✓	✓
mediante contacto de persona a persona	*through contact from one person to another*	✓	✓
por saliba secreciones nasales, tos, falta de aseo	*through salive, secretion, cough, no cleanliness*	✗	✓
cuando estamos juntos	*when we are together*	✗	✓
transmisión aérea de persona a persona, vias respiratorias principalmente.	*through air transmission from person to person, mostly through respiration*	✗	✗
no acercándose mucho a otras personas	*we should not come too close to other people*	✗	✗

Table 2: Example Sentences

answer and a topic in around 50 % of the cases.

5 Discussion

The computational approaches we tested have shown much lower accuracy compared to human annotations. The biggest problem we have identified is the lack of pragmatic inferencing humans are good at, but automatic models we tested are not. For instance, people answered to the question about how the virus distributes by saying “through crowd”. Due to a pragmatic inference human annotators can evaluate this answer as similar to the answer given by the reference corpus. “A crowd” implies pragmatically that social distancing cannot be obtained adequately and this can promote virus infection. However, none of our automatic models was able to predict a high similarity between the reference answer "through contact" and the participant’s answer “through crowd”. Another example illustrating problems with pragmatic inferences is the annotation of vaccination side effects. While human annotators had no difficulties to classify “thrombosis” as a possible vaccination side-effect, our automatic methods were not able to do it. To sum up, one of the biggest challenges in our tasks was the lack of Natural Language Understanding and Inferencing (NLI and NLU) by the computational models we tested. Using NLI and NLU in the context of low resource is reserved for future research. In the near future, we will test models trained on health-related topics, fragmented answers that represent the majority of our answers and models trained on NLI-and NLU-datasets (Kochkina et al., 2023).

Future Work

There are several issues of our methodology that need to be addressed in future research. The absence of good resources for indigenous languages has forced us to work with local translators who digitized the answers the way they perceived them. In future we will use transcribed oral data for our experiments.

Another issue is the use of few human annotations that have provided us the human similarity score necessary to evaluate computational models. Even though the inter-speaker agreement was comparatively high in our study due to very explicit training and discussion of annotation guidelines, we suspect that the inter-speaker agreement will show a much higher variation in the perception of semantic similarity if the annotation guidelines are missing as is often the case in crowd-sourced human annotations. The trade-off between expensive human annotators with long training for annotation and cheap crowd-sourced human annotations without any training is an issue that needs to be addressed in the future research.

Ethics Statement

Scientific work carried out in our project complies with the ACL Ethics Policy and with the ethic guidelines from the German Research Foundation (DFG). We have informed our participants about the goals of our project and they signed an agreement with us. In addition, the data acquisition by interviewing indigenous people was approved by Ethic committees at the universities of our cooperation partners.

Task	Cosine sim.	Synonym-token-sim.	Synonym-token-sim.+ stemming
Value	0.2	0.4	0.5
Human Annot.	0.7	1	1
Accuracy loss	0.5	0.6	0.5

Table 3: Accuracy values and accuracy loss per task

Acknowledgements

We acknowledge the funding support from the German Research Foundation (DFG) (Grant number: 468416293).

References

Rima A. Afifi, Nicole Novak, Paul A. Gilbert, Bernadette Pauly, Sawsan Abdulrahim, Sabina Faiz Rashid, Fernando Ortega, and Rashida A. Ferrand. 2020. 'most at risk' for covid19? the imperative to expand the definition from biological to social factors for equity. *Preventive Medicine*, 139:106229.

Rosso A. Di Paolo C. et al. Baccolini, V. 2021. What is the prevalence of low health literacy in european union member states? a systematic review and meta-analysis. doi: 10.1007/s11606-020-06407-8. epub 2021 jan 5. pmid: 33403622; pmcid: Pmc7947142. *Journal of General Internal Medicine*, 36:753–761.

Bendala-Estrada AD Ramos-Martínez A Muñez-Rubio E Fernández Carracedo E et al. Calderón-Parra J, Muiño-Miguez A. 2021. Inappropriate antibiotic use in the covid-19 era: Factors associated with inappropriate prescribing and secondary complications. analysis of the registry semi-covid. *PLoS ONE*, 16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gerardo M. García, Marleen Haboud, Rosaleen Howard, Antonia Manresa, and Julieta Zurita. 2020. Miscommunication in the covid-19 era. *Bulletin of Latin American Research*, 39(S1):39–46.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

G. Gonzalez Hernandez Klein, A.Z.; A. Magge; K. O'Connor; J.I. Flores Amaro; D. Weissenbacher. 2021. Toward using twitter for tracking covid-19: A natural language processing pipeline and exploratory data set. *Journal of medical Internet research*, 23.

Elena Kochkina, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing Management*, 60(1):103116.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.

Atul Kr. Ojha, Chao-Hong Liu, Katharina Kann, John Ortega, Sheetal Shatam, and Theodorus Fransen. 2021. Findings of the LoResMT 2021 shared task on COVID and sign language for low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 114–123, Virtual. Association for Machine Translation in the Americas.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

News UN. 13 April 2020. During this coronavirus pandemic, 'fake news' is putting lives at risk: Unesco.

Department of Economic United Nations and Social Affairs. 2020. Indigenous peoples and the covid-19 pandemic: considerations.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez. 2019. Deep neural networks ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*, 26(12):1618–1626.