Project Proposal Template

First Author
G01377914
Lokeshwar Sai Sreeramdas
lsreeram@gmu.edu

1 Introduction

The primary objective of my research is to examine NLP methodologies in health education for Native Americans in Latin America, especially Covid-19. To achieve goal, I asked survey participants to respond to questions about health-related subjects. I used the responses of participants to assess their level of health education. In this paper, the summary of NLP application's findings regarding the participants' responses. In the first experiment, they used embedding-based techniques to quantify the semantic similarity between participants' responses and "expert" or "reference" answers. In the second experiment, they used algorithms based on synonyms to group answers under subjects.

1.1 Context and Goal

The COVID-19 epidemic has disproportionately affected groups, most importantly Latin American indigenous peoples (United Nations and Affairs, 2020). Observing the problems that they face such as limited access to healthcare facilities and socioeconomic marginalization—my research aims to support COVID-19 health education for indigenous populations.

1.2 Communication Challenges

Due to cultural differences and less information in indigenous languages, miscommunication can be a big problem (Garca et al., 2020; Afifi et al., 2020). This misunderstanding may make it challenging to comprehend and implement advised health precautions.

1.3 Importance of Reliable Information

It is very important to offer valid COVID-19-related information that is appropriate for the linguistic and cultural backgrounds of indigenous peoples (UN, 13 April 2020). Initiatives from the World Health Organization (WHO) and the European Union (EU) highlight how critical it is to address communication problems when transferring important health-related terms and concepts.

1.4 Likely challenges and mitigations

The most important aspect of preventing the spread of infectious diseases is health literacy.The recent COVID-19 epidemic highlighted the importance of misinformation in determining the severity of the issue (UN, 13 April 2020). Understanding people's attitudes, such as antibiotic misuse, is critical for reducing health-illiteracy (Calderón-Parra J, 2021).

1.5 Research Objective

The need for Natural Language Processing (NLP) tools is urgent given the urgency of accurate health-related communication. The process of measuring health education can be accelerated and streamlined with the help of these tools, particularly in relation to the COVID-19 pandemic.

I ran a survey questionnaire to members of vulnerable groups—specifically, indigenous people in Latin America—in order to meet my research objectives. In order to gauge the participants' level of health education, a series of health-related questions were included in the questionnaire.

My study's main objective is to evaluate NLP approaches for assessing COVID-19-related health education. I pay particular attention to how similar the participants' answers are semantically and how they are categorized under different topics. To evaluate the effectiveness of NLP techniques, I compare the outcomes with human annotations.

2 Approach

To achieve our research goals, we conducted a survey questionnaire among participants from vulnerable groups, specifically indigenous people from Latin America. The questionnaire sought answers on various health-related topics, serving as the foundation for measuring participants' health education status.

NLP-Based Semantic Similarity Measurement, In the first experiment, we employed embeddings-based tools to measure semantic similarity between participants' answers and ”expert” or ”reference” answers. This involved leveraging advanced NLP techniques to capture nuanced semantic relationships within the responses. Topic Classification Using Synonym-Based Methods,The second experiment focused on topic classification, utilizing synonym-based methods. We aimed to categorize participants' answers under specific health-related topics. This approach involved identifying synonymous terms and mapping responses to predefined topics.

To validate the effectiveness of our NLP methodologies, we compared the results obtained from both experiments with human annotations. Human annotators, with their pragmatic inferencing skills, were considered the benchmark for semantic similarity and topic classification accuracy.

2.1 Methodologies

Survey Study Design: To assess the health education status of indigenous groups speaking Quechua or Kichwa from Peru and Ecuador, we conducted a survey study in collaboration with our partners from Latin America: Marleen Haboud, Claudia Crespo, and Fernando Ortega Pérez.

Participant Demographics: Approximately 150 participants from each country participated in the survey. The chosen indigenous communities provided responses to questions about COVID-19, including 10 yes-no questions and 10 open-ended questions. The study aimed to capture a diverse range of perspectives within these communities.

Task Objectives: Our primary task was to measure the accuracy of key health-related concepts. Specifically, we sought to understand how well the information status of indigenous groups aligns with information and suggestions from reliable sources such as the World Health Organization (WHO). The WHO guidelines, forming our Reference Corpus, served as a benchmark for evaluating the participants' responses.

Example Task: As an illustration, participants were asked about the distribution of the COVID-19 virus, and their responses were compared against WHO recommendations. For instance, the WHO suggests that the virus is transmitted through contact, thereby recommending social distancing. By posing questions related to this aspect, we aimed to gauge the participants' understanding and alignment with authoritative health information.

Data Collection Method: The survey data were collected in rural areas through free interviews conducted by a local person familiar with indigenous communities. This approach was crucial for inclusivity, ensuring the involvement of individuals less accustomed to highly controlled tasks, such as older and/or illiterate participants.

Interview Summarization: Due to constraints in time and resources, we opted not to transcribe the interviews. Instead, a local interviewer summarized the answers in a digital form in Spanish. While this method provides valuable insights, it is essential to note that the answers may not directly reflect the information state of indigenous minorities.

The accuracy measurement of medical terms, including crucial concepts like antibiotics, faces significant challenges. Two primary obstacles contribute to this gap: a) Missing Data Sources and Methodologies:

Current methodologies lack the necessary data sources and techniques to identify, characterize, and measure the actual uses of health-related topics and concepts. b) Missing Statistical (In)Accuracy Measures: There is a lack of statistical measures assessing the accuracy of information related to infectious diseases.

To bridge these gaps, initiatives like Social Media Mining 4 Health (SMM4H) have emerged (Klein, 2021; Magge et al., 2021). This initiative leverages social media data to address health-related tasks, such as identifying disease mentions and symptoms. However, social media data presents limitations:

Demographic Information Gap: It lacks demographic information crucial for studying social variation in health literacy. Representation Gap: Not all social groups, including indigenous populations, are adequately represented, given factors like low internet access or alternative communication tools. Utilizing Traditional Methodology: To overcome these challenges, we adopted a traditional survey methodology commonly used in social sciences. This approach allows us to access information about the health education status of participants. Survey questions, covering topics like virus propagation and treatment, were posed to participants.

Comparison with ”Expert” Knowledge: To measure the accuracy of health-related concepts and the uses of participants, a comparison was made with ”expert” knowledge or uses. This is essential for understanding the disparities between participant information status and established expertise.

Incorporating Neural Language Models: Recent advancements in semantic comparison, particularly through neural language models like BERT (Devlin et al., 2019; Giulianelli et al., 2020), have enabled significant progress. BERT, trained on extensive natural language data, has proven effective in predicting masked words and discerning lexical meaning variations. However, its application in the low-resource scenario, particularly in survey questionnaires within specific domains like ethnic minorities, remains underexplored.

Vector-Based Approaches for Answer Similarity Measurement: This paper contributes to the field by testing vector-based approaches in the measurement of answer similarity within the low-resource domain. The focus is on understanding how well these models perform in scenarios with specific topic domains and the unique format of answers provided in survey questionnaires.

2.2 Advanced Natural Language Processing model

Incorporating Advanced NLP Model: all-mpnet-base-v2 In our pursuit of understanding health education status among vulnerable groups, we employed an advanced Natural Language Processing (NLP) model known as all-mpnet-base-v2. This model is part of the Megatron series, designed for handling large-scale language tasks.

Overcoming Data Source Limitations: Recognizing the limitations of traditional data sources, especially in the context of low-resource domains and specific linguistic characteristics, the all-mpnet-base-v2 model serves as a valuable tool. Its capacity to capture complex linguistic relationships and nuances makes it particularly well-suited for our study.

Vector-Based Semantic Comparison: The all-mpnet-base-v2 model utilizes vector-based approaches for semantic comparison, similar to BERT. However, it extends beyond by incorporating multi-modal learning capabilities. This allows it to not only understand variations in word meanings but also consider contextual information across different modalities, enhancing its performance in diverse scenarios.

Testing in Low-Resource Domain: To assess the effectiveness of the all-mpnet-base-v2 model, we conducted experiments in the low-resource domain of health education among indigenous populations. By leveraging the model’s strengths in semantic understanding and its ability to handle diverse linguistic expressions, we aimed to measure answer similarity more effectively compared to traditional methodologies.

Comparative Analysis: Our study involves a comparative analysis between the results obtained using all-mpnet-base-v2 and other vector-based approaches. This comparative approach helps us evaluate the model’s performance in the specific context of our survey questionnaires and its ability to bridge the gap in accuracy measurement for health-related concepts.

By incorporating the all-mpnet-base-v2 model into our approach, we leverage advanced NLP capabilities to enhance the accuracy and effectiveness of our measurements in the unique context of health education among vulnerable populations.

Multi-Engine Implementation: To further enrich the accuracy measurement process, we implemented a multi-engine approach. This involved leveraging diverse Natural Language Processing (NLP) models, including the advanced all-mpnet-base-v2 model and other relevant engines capable of semantic analysis.

Diversity in NLP Models: Recognizing the complexity and nuances of indigenous languages, we employed a range of NLP models with the capacity to comprehend and analyze linguistic diversity. The inclusion of multiple engines allowed us to capture a broader spectrum of semantic meanings and variations.

Task Distribution among Engines: Different NLP models were assigned specific tasks based on their strengths. For example, the all-mpnet-base-v2 model, with its multi-modal learning capabilities, was particularly well-suited for capturing contextual nuances in responses, while other engines focused on semantic similarity and topic classification.

Comparative Analysis: The multi-engine implementation facilitated a comparative analysis of the results obtained from each engine. This comparison aimed to discern variations in performance, strengths, and weaknesses among the different NLP models.

Robustness Testing: To ensure the robustness of our measurements, we conducted extensive testing on the multi-engine implementation. This involved evaluating the models across various linguistic contexts, survey question formats, and demographic characteristics to gauge their adaptability and reliability.

Integration of Results: The results from individual engines were integrated, considering the strengths of each model. This approach allowed us to harness the collective power of multiple engines, enhancing the overall accuracy and reliability of our health education status measurements.

Ethical Considerations: Throughout the implementation of multi-engine models, we adhered to strict ethical guidelines. Ensuring the privacy and confidentiality of participants was paramount. Additionally, efforts were made to minimize biases and promote inclusivity in the analysis process.

3 Experiments

Our study comprised two distinct experiments, each contributing to the comprehensive measurement of health education status among indigenous groups.

Experiment 1: SBERT Model for Semantic Similarity Measurement In the first experiment, we employed the Sentence-BERT (SBERT) model to measure the semantic similarity between participants' answers and the "expected" answers from the reference corpus.

3.1 Datasets

Please list which datasets you are using, whether or not you have access to them, and whether or not they are publicly available with the same preprocessing and train / dev / tests as the work you will be reproducing.

The dataset originates from a survey study aimed at assessing the health education status of indigenous groups, specifically those speaking Quechua or Kichwa from Peru and Ecuador in the context of COVID-19. With approximately 150 participants from each country, the study focuses on linguistic and cultural diversity within indigenous communities. The survey comprises 10 yes-no questions and 10 open-ended questions, covering various aspects of COVID-19, including transmission, preventive measures, and treatment.

Data collection involved free interviews in rural areas, conducted by a local person familiar with indigenous communities. Responses, provided in Quechua or Kichwa, were summarized in Spanish to address resource constraints. Challenges such as the absence of accurate health-related measurements and limitations in social media data were acknowledged. In response, a traditional survey methodology in social sciences was employed, emphasizing inclusivity for participants less accustomed to controlled tasks.

The dataset seeks to measure the accuracy of health-related concepts by comparing participant information status with "expert" knowledge. A comparative analysis is crucial for evaluating the reliability of participant responses. The dataset, along with experiment code, is openly accessible on GitHub, fostering transparency and collaboration. In essence, this dataset represents a concerted effort to bridge gaps in health education measurement among indigenous populations using culturally sensitive survey methodologies.

3.2 Implementation

The implementation of this study incorporates a thoughtful and comprehensive approach to address challenges in measuring the accuracy of health-related concepts among indigenous populations. Two experiments were conducted, and the data and code are openly available on GitHub.

In the first experiment, the Sentence-BERT (SBERT) model was employed to gauge semantic similarity between participants' answers and those from a reference corpus. This model, known for its ability to capture semantic relationships, utilized cosine similarity as a metric. Examples of results from this experiment showcase the effectiveness of the SBERT model in measuring the semantic alignment of responses.

A multi-engine implementation strategy was adopted, leveraging diverse Natural Language Processing (NLP) models, including the advanced all-mpnet-base-v2 model. This choice reflects an understanding of the linguistic complexity within indigenous languages. Different engines were assigned specific tasks based on their strengths, facilitating a comparative analysis of results. Rigorous testing ensured the robustness of the multi-engine approach across various linguistic contexts, survey question formats, and demographic characteristics.

The integration of results from individual engines, considering their unique strengths, aimed to enhance the overall accuracy and reliability of health education status measurements. Ethical considerations were paramount throughout the implementation, ensuring participant privacy and minimizing biases.

In summary, the implementation combines advanced NLP models, a multi-engine strategy, and ethical considerations to comprehensively measure health education status, bridging gaps in existing methodologies and offering a nuanced understanding of indigenous perspectives on health-related concepts in the context of COVID-19.

3.3 Comparing the models

The new approach, incorporating the all-mpnet-base-v2 model, marks a significant evolution from the previous SBERT-centric methodology. While the previous approach relied solely on SBERT for semantic analysis, the new strategy employs the advanced all-mpnet-base-v2 model, known for its multi-modal learning capabilities and improved understanding of diverse linguistic contexts. Notably, the new approach adopts a multi-engine implementation, allocating specific tasks to different engines, including all-mpnet-base-v2, for a comprehensive analysis. Rigorous testing ensures the robustness of the multi-engine strategy across linguistic nuances and demographic variations. Ethical considerations, including privacy and bias minimization, remain central in both approaches. Overall, the new methodology enhances semantic analysis, embraces a multi-engine approach, and rigorously tests for robustness, providing a nuanced understanding of health education status among indigenous populations.

3.4 Multilingualism Approach

The study employed a multilingual approach by translating the data into Spanish and reapplying the same methods to enhance the model's performance. This strategy suggests an effort to improve the model's understanding across linguistic variations, acknowledging the importance of linguistic diversity in the dataset. By leveraging translation and reapplying existing methodologies, the study aims to refine the model's effectiveness in capturing semantic nuances and cultural contexts within the Spanish-translated dataset. This approach reflects a commitment to enhancing the model's versatility and performance across multiple languages, contributing to a more robust and inclusive analysis of health education status among indigenous populations.

3.5 Results

Provide a table comparing your results to the published results.

3.6 Discussion

Discuss any issues you faced. Do your results differ from the published ones? If yes, why do you think that is? Did you do a sensitivity analysis (e.g. multiple runs with different random seeds)?

3.7 Resources

Discuss the cost of your reproduction in terms of resources: computation, time, people, development effort, and communication with the authors (if applicable).

3.8 Error Analysis

Perform an error analysis on the model. Include at least 2-3 instances where the model fails. Discuss the error analysis in the paper – what other analyses could the authors have run? If you were able to perform additional error analyses, report it here.

4 Conclusion

Is the paper reproducible?

References