Alexis Quesada-Arencibia
José Carlos Rodríguez
Gabriele Salvatore de Blasio
Carmelo Rubén García
Roberto Moreno-Díaz (Eds.)

EUROCAST 2024

Computer Aided Systems Theory

EXTENDED ABSTRACTS

19th International Conference on Computer Aided Systems Theory
Las Palmas de Gran Canaria, Spain, February 2024

Building Spanish Trustworthy Question-Answer Datasets for Suicide Information*

Pablo Ascorbe[1], María S. Campos[2], César Domínguez[1], Jónathan Heras[1], Magdalena Pérez[3], and Ana Rosa Terroba-Reinares[1,4]

[1] Departamento de Matemáticas y Computación, Universidad de La Rioja, Spain
{pablo.ascorbe, cesar.dominguez, jonathan.heras}@unirioja.es
[2] Unidad de Salud Mental Espartero, Logroño, La Rioja
mscampos@riojasalud.es
[3] Teléfono de la Esperanza
magdalenaperez@telefonodelaesperanza.org
[4] Fundación Rioja Salud
arterroba@riojasalud.es

Suicide is a public health problem since worldwide more than 800,000 suicides are estimated to occur every year, one every 40 seconds, figures of epidemic proportions [4]. Such is the scale of the problem that specialists tasked with answering questions related to it are overwhelmed and overburdened. Moreover, much of the information that can be found about suicide on the Internet could be more harmful than helpful. These problems could be tackled by means of automatic Question-Answering (Q&A) systems [3]; however, it is necessary to have trustworthy corpora that help to validate, train, and guide the construction of such systems [2]. Since, up to the best of the authors knowledge, there is not a Spanish Q&A corpus for suicide information, the aim of this work is to create such a corpora for suicide information. In particular, we have considered three levels of quality [1]: *bronze-standard*, when the entire data has been generated automatically and has little processing; *silver-standard*, when starting from a bronze corpora, a processing stage is applied to refine the data followed by an annotation and validation stage conducted by experts; and, *gold-standard*, when the corpus has been manually generated and validated by experts.

The starting point to obtain the corpora is a set of trustworthy Spanish documents provided by suicide experts (in our case, it is composed of a total of 151 documents). Then, a bronze-standard Q&A corpora was built using three large language models, two in Spanish (bertin-gpt-j-6B-alpaca and Llama-2-7b-ft-instruct-es) and one in English (t5-base-squad-qag), all of them freely available at Hugging Face platform. We split the documents into chunks, and for each chunk, we asked the language models to generate a Q&A pair. Using this procedure, a total of 22,920 Q&A pairs were obtained, but many generated questions were incomplete, repeated or contained essentially the same information among them. Therefore, some filters were applied to increase the quality of

* This work was partially supported by Grant PID2020-115225RB-I00 funded by MCIN/AEI/ 10.13039/501100011033, and by funds for the 2023 strategies of the Spanish Ministry of Health, which were approved in the CISNS on June 23, 2023, to support the implementation of the Mental Health Action Plan.

the generated corpus. A first filter was basic data processing operations, such as the elimination of empty, duplicated or incomplete pairs. Then, we trained a deep learning classification model (named bertin-roberta-base-spanish-spanish-suicide-intent, again freely available at Hugging Face) using a suicidal behaviour dataset to determine whether a question-answer pair contains information about suicide. This model allowed us to filter out not suicide related pairs. Finally, we removed the pairs that were semantically similar by using an embedding and the cosine distance. After all these steps, we obtained the final version of our bronze-standard corpora, leaving us with 4,901 Q&A pairs.

From that bronze corpora, a manual filtering was conducted by non-experts to eliminate Q&A uninteresting pairs; i.e., they talked about suicide but seemed not to be useful or were too ambiguous. This left 484 pairs to be evaluated by experts. In order to perform such an evaluation, a web application was developed for the validation of the corpus by a group of psychologists and psychiatrists, allowing them to update or remove the pairs that did not pass their filter; thus obtaining the silver-standard corpus with 380 Q&A pairs.

In order to evaluate the models performance to generate interesting Q&A pairs, we describe the number of pairs obtained by each model in each step of the process. As a starting point, 7,806 were generated from Bertin, 4,558 from Llama-2, and 10,557 from t5-base-squad. After automatic filters, the number of pairs were reduced to 760 from Bertin, 1,166 from Llama-2, and 2,977 from t5-base-squad. After manual filtering by a non-expert, we had 337 from Bertin, 55 from Llama-2, and 92 from t5-base-squad. Finally, the silver corpus had 272 from Bertin, 45 from Llama-2, and 63 from t5-base-squad. We can observe that the Bertin model had the best performance.

In addition, a gold-standard corpus directly extracted from the FAQs sections contained in some of the documents provided by experts (i.e. they are not automatically generated) has been also provided. This dataset consists of 118 Q&A items.

The built Q&A corpus and models described previously are freely available at `https://huggingface.co/PrevenIA`. These corpus are the first step towards the validation, training, and construction of automatic Q&A systems that provide information about suicide.

References

1. Casola, S., Lavelli, A., Saggion, H.: Creating a silver standard for patent simplification. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1045–1055 (2023)
2. Das, B., Nirmala, S.J.: Improving healthcare question answering system by identifying suitable answers. In: 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon). pp. 1–6. IEEE (2022)
3. Rogers, A., Gardner, M., Augenstein, I.: QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. ACM Computing Surveys **55**(10), 1–45 (2023)
4. WHO: Suicide worldwide in 2019: global health estimates (2021)