The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora

Luis Gascó†, Darryl Estrada-Zavala, Eulàlia Farré-Maduell,
Salvador Lima-López, Antonio Miranda-Escalada, Martin Krallinger
Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Corresponding author: luis.gasco@bsc.es

Abstract

There is a pressing need to exploit health-related content from social media, a global source of data where key health information is posted directly by citizens, patients and other healthcare stakeholders. Use cases of disease-related social media mining include disease outbreak/surveillance, mental health and pharmacovigilance. Current efforts address the exploitation of social media beyond English. The SocialDisNER task, organized as part of the SMM4H 2022 initiative (Weissenbacher et al., 2022), has applied the LINKAGE methodology to select and annotate a Gold Standard corpus of 9,500 tweets in Spanish enriched with disease mentions generated by patients and medical professionals. As a complementary resource for teams participating in the SocialDisNER track, we have also created a large-scale corpus of 85,000 tweets, where in addition to disease mentions, other medical entities of relevance (e.g., medications, symptoms and procedures, among others) have been automatically labelled. Using these large-scale datasets, co-mention networks or knowledge graphs were released for each entity pair type. Out of the 47 teams registered for the task, 17 teams uploaded a total of 32 runs. The top-performing team achieved a very competitive 0.891 f-score, with a system trained following a continue pretraining strategy. We anticipate that the corpus and systems resulting from the SocialDisNER track might further foster health-related text mining of social media content in Spanish and inspire disease detection strategies in other languages. Corpus: https://doi.org/10.5281/zenodo.6359365

1 Introduction

With more than 4.2 billion users worldwide, social media have become the most widely used digital platform for interacting with peers as well as accessing information relevant to specific groups (Kemp, 2021). Specifically Twitter, an online micro-blogging social network (OSN), has been widely used to extract information about people: from opinions and effects of environmental pollution (Gasco et al., 2019; Otero et al., 2021) to biomedical aspects such as adverse drug reactions (OConnor et al., 2014; MacKinlay et al., 2017), public health (Collier et al., 2008), and the psychological effects of a pandemic on the population (Aiello et al., 2021).

Most analysis and information extraction studies are carried out in English, the main language used by the platform's users. However, other major languages such as Spanish have generated a large amount of potentially usable data for analysis (Alshaabi et al., 2021), which increases the impact of NLP systems able to extract information in this language.

To date, the resources for working with Twitter data in Spanish have focused on corpora to extract information related to emotions (Plaza-del Arco et al., 2020; Martinez-Camara et al., 2015) and professions (Miranda-Escalada et al., 2021). Recently, language models trained with Spanish tweets have been developed to improve the performance of NLP tasks applied to data obtained from this OSN (Huertas-Tato et al., 2022). However, there is a clear lack of corpora to train systems capable of extracting biomedical information from Spanish content that could be used for real-time mining of health information posted by patients, which is of growing interest for different scenarios such as screening for rare diseases (Miller et al., 2021).

One of the main entities to effectively recognise health-related content is diseases. Shared-task such as DisTEMIST (Miranda-Escalada et al., 2022; Nentidis et al., 2022), which focused on the detection of disease mentions in clinical cases, fostered the creation of many tools for this purpose. These systems focus on extracting information from technical and scientific texts, but their performance is limited to the more lay language used in social networks. In the SocialDisNER shared task, we have applied the knowledge and experience from clinical cases acquired in DisTEMIST to detect disease mentions in Spanish tweets written by patients and

Proceedings of the 29th International Conference on Computational Linguistic, pages 182–189
October 12–17, 2022.

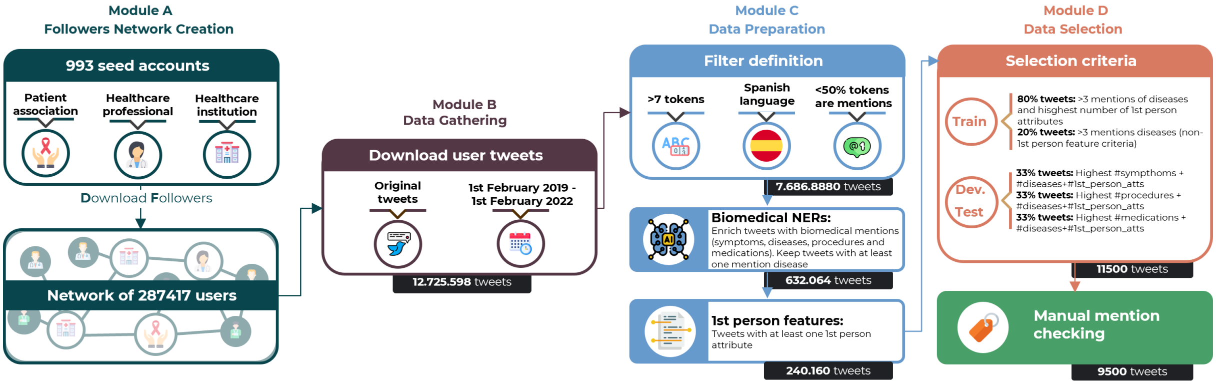
Module A
Followers Network Creation
993 seed accounts
Patient association
Healthcare professional
Healthcare institution
Download Followers
Network of 287417 users
Module B
Data Gathering
Download user tweets
Original tweets
1st February 2019 - 1st February 2022
12.725.598 tweets
Module C
Data Preparation
Filter definition
>7 tokens
Spanish language
<50% tokens are mentions
7.686.8880 tweets
Biomedical NERs:
Enrich tweets with biomedical mentions (symptoms, diseases, procedures and medications). Keep tweets with at least one mention disease
632.064 tweets
1st person features:
Tweets with at least one 1st person attribute
240.160 tweets
Module D
Data Selection
Selection criteria
Train
80% tweets: >3 mentions of diseases and hisghest number of 1st person attributes
20% tweets: >3 mentions diseases (non-1st person feature criteria)
Dev.
Test
33% tweets: Highest #sympthoms + #diseases+#1st_person_atts
33% tweets: Highest #procedures + #diseases+#1st_person_atts
33% tweets: Highest #medications + #diseases+#1st_person_atts
11500 tweets
Manual mention checking
9500 tweets

Figure 1: LINKAGE (reLevant socIal NetworK dAta GathEring) methodology for twitter data selection.

medical professionals.

2 Task Description

Shared Task Goal. SocialDisNER focuses on the recognition of disease mentions in Twitter posts. Tweets originate from the following: patients' accounts, with firsthand health reports; friends, support network and relatives, who share the difficulties faced by patients; and medical professionals, who disseminate reliable information about diseases. Tweets in this task include information on rheumatic diseases such as lupus erythematosus, highly prevalent diseases such as cancer, diabetes, obesity and mental disorders, fibromyalgia and autism spectrum conditions.

Shared Task Setting. The track comprised a single challenge in which participants were required to train NLP systems capable of detecting disease mentions automatically. We conducted the task using CodaLab in a three-stage scenario. In the *practice phase*, the training and evaluation set were released so that participants could build their systems and evaluate their models on the validation set. For the *evaluation phase*, the test and background set were released without annotations for the participant teams to compute and submit their predictions. Teams were only evaluated on the test set, using the background data to avoid possible manual corrections and to evaluate the scalability of the systems. In this phase, the participants were only allowed to upload 2 runs. In order to further develop disease mention recognition systems in tweets, the competition has been kept open in the *post-evaluation phase*, so that researchers continue to measure and compare the performance of their systems.

Evaluation metrics. Teams were evaluated and ranked using micro-average F1 score. In addition to this metric, micro-averaged precision and recall were computed. The evaluation script is available on Github[1]. We also compared the systems versus a baseline model following a lexical search approach (TeMU-BSC, 2022).

3 Corpus and resources

3.1 SocialDisNER Gold Standard

Gold Standard selection methodology. SocialDisNER has compiled a set of 9,500 tweets written in Spanish containing patient and family members experiences about diseases and relevant content written by clinical professionals. The LINKAGE methodology was created to obtain a larger number of tweets posted by patients themselves and relevant professionals. This methodology has been designed to avoid noisy health twitter content and biases associated with keyword-based selection (Kowald and Lex, 2018). The procedure, shown in the Figure 1, consists of 4 stages:

1. **Followers Network Creation**: First, a user network with a common interest in the health field is downloaded. For SocialDisNER, a total of 993 seed Twitter accounts of patient associations, health professionals and healthcare institutions were selected and manually curated by task organizers. We obtained the followers from this pool of accounts, resulting in a community of 287,417 users.

2. **Data Gathering**: Second, the content published by this community of users was downloaded using the Twitter API. In the task, only the original tweets written by users were

[1] https://github.com/TeMU-BSC/socialdisner_evaluation_script

downloaded, ignoring retweets and replies to other tweets. Tweets posted between February 2019 and February 2022 were downloaded to prevent having data only about the COVID-19 pandemic, as we have tweets prior to the beginning of it. After this downloading process, more than 12.7 million tweets were obtained.

3. **Data Preparation**: Third, rules were applied to filter out irrelevant content as follows: a) tweets written in a language other than Spanish according the Twitter API; b) tweets with less than 7 tokens, as they might not contain substantial information about diseases; and c) documents in which more than 50% of tokens were mentions. A data enrichment process was applied to the resulting set to produce the final candidates. On the one hand, biomedical mentions such as symptoms, procedures, diseases and drugs were extracted using NER systems developed in previous works (Gonzalez-Agirre et al., 2019). Only tweets with at least one disease mention were selected. First-person characteristics, such as the presence of first-person pronouns, were also calculated, considering only tweets that at least had a first-person attribute. After these constraints, a set of more than 240k tweets were obtained.

4. **Data Selection**: Finally, criteria were determined for selecting candidate tweets to be annotated by experts. Selection strategies were applied so that there was content with several mentions of diseases written in the first person and also had mentions of other biomedical diseases such as symptoms, procedures and medications. The selection criteria for the development and test sets were the same. At the end of this phase, 11,500 tweets were selected for annotation.

Gold Standard statistics. Tweets were annotated by a medical expert during 3 months using an adaptation of the DisteMIST annotation guidelines, whose agreement was tested among medical and linguistic experts with a IAA score of 0.823. A total of 9,500 tweets were finally selected for the task. This set of tweets was divided into a training, a development and a test set. Table 1 shows the distribution and statistics of the corpus,

	Corpus name	Documents	Annotations	Tokens
Gold Stand.	**Training**	5,000	15,173	211,555
	Development	2,500	4,252	84,478
	Test	2,000	3,859	70,244
	Total	9,500	23,284	366,277
Silver Stand.	**Socialdisner-Diseases**	85,077	116,260	3,236,411
	Socialdisner-Symptoms	12,624	12,896	521,503
	Socialdisner-Procedures	11,464	10,080	467,059
	Socialdisner-Pharma	1,759	1,029	68,269
	Socialdisner-Morphology_neoplasms	8,518	8,943	332,539
	Socialdisner-Professions	15,831	18,590	660,071
	Socialdisnerv-Person	41,033	58,007	1,689,479
	Socialdisner-Species	12,118	14,014	486,249

Table 1: SocialDisNER corpora summary.

3.2 SocialDisNER Large Scale corpus

A set of 85,000 tweets was selected to generate large-scale corpora of Spanish tweets with several biomedical mentions. These datasets were published as Silver Standard and were generated using NER systems trained on data previously published by our team. Since those NERs were not trained with tweets, programmatic cleanups were performed by eliminating mentions containing URLs and more than one twitter mention. Additionally, we conducted a manual review of the most recurrent mentions to eliminate false positives. The statistics for each large-scale corpus are shown in Table 1. Due to the selection process of the SocialDisNER data, which focused on diseases, there is a higher presence of content with this entity.

3.3 SocialDisNER co-mention networks

Inspired by (Hope et al., 2020), several co-occurrence matrices of mentions present in the large-scale corpus have been made available to the community. On the one hand, we have published a disease co-occurrence matrix, which could be used to analyze the comorbidity of diseases among users. Co-mention matrices between diseases and other entities have also been published, providing interesting associations between diseases and symptoms, professions and medicines, among others. To the best of our knowledge, this is the first graph of social network biomedical mentions in Spanish.

3.4 SocialDisNER guidelines

We have also published the SocialDisNER guidelines. This document shows the annotation criteria used by medical experts when creating the corpus and ensure its quality and replicability. The guidelines were created by adapting the DisTEMIST annotation rules to the special features of social me-

Team	Country	A/I	Tool	P	R	F1	Ref
CASIA	China	A	-	**0,906**	0,876	**0,891**	(Fu et al., 2022)
READ-BioMed	Australia	A	(READ-BioMed, 2022)	0,868	0,875	0,871	(Yepes and Verspoor, 2022)
Clac	Canada	A	-	0,851	**0,888**	0,869	(Verma et al., 2022)
PLN CMM	Chile	A	(PLN-CMM, 2022)	0,882	0,843	0,862	(Rojas et al., 2022)
NLP-CIC-WFU	México	A	(Tamayo, 2022)	0,842	0,860	0,851	(Tamayo et al., 2022)
dezzai	Spain	I	-	0,828	0,845	0,836	(Ortega-Martín et al., 2022)
RACAI	Romania	A	(RACAI, 2022)	0,868	0,779	0,821	(Avram et al., 2022)
KU_EDI	Korea	A	-	0,809	0,798	0,803	(Lain et al., 2022)
SINAI	Spain	A	(SINAI, 2022)	0,756	0,795	0,775	(Chizhikova et al., 2022)
ITAINNOVA	Spain	I	(ITAINNOVA, 2022)	0,779	0,769	0,774	(Montañés-Salas et al., 2022)
FRE	Spain	I	-	0,680	0,805	0,738	(Cetina and García-Santa, 2022)
baseline			(TeMU-BSC, 2022)	0,776	0,701	0,737	
HIBA	Argentina	I	-	0,759	0,644	0,697	(Castano et al., 2022)[a]
TEAM IAI	France	A	-	0,640	0,655	0,647	(Sinha et al., 2022)
CAISA[b]	Germany	A	-	0,836	0,494	0,621	(Karimi and Flek, 2022)
ResearchX	India	A	-	0,505	0,625	0,559	-
AILAB Udine	Italy	A	-	0,504	0,461	0,481	(Portelli et al., 2022)
JSL[c]	USA	I	-	0,004	0,004	0,004	(Kocaman et al., 2022)

[a] Same system that the one used in DisTEMIST.

[b] The best system of this team was a post-workshop evaluation.

[c] This team had problems in the evaluation phase due to the format used in the submission.

Table 2: SocialDisNER ranking with the best submission per team. Best result bolded, second best underlined. A/I stands for Academy/Industry.

dia. The final version contains 56 annotation and restriction rules in relation to disease concepts. The guidelines are freely available at Zenodo (Farré-Maduell et al., 2022).

4 Results

Participation. SocialDisNER achieved a considerable impact in the scientific-technical community. A total of 47 teams were registered, out of which 17 submitted a total of 32 runs for evaluation. Although most of the teams came from academic environments, a significant number (4) came from industry. Interestingly, 10 teams came from non-Spanish-speaking countries, which indicates the community interest in developing systems in languages other than English.

Results. Table 2 shows the results of SocialDisNER. Eleven teams achieved better performance than the baseline of the task. The best performing system was developed by the CASIA team, with a micro-average F1-score of 0.891. The Clac team obtained the best performance in terms of Recall, with a value of 0.888. The top-performing team developed a Unified Named Entity Recognition system by following a continual pre-training strategy based on transformers architecture. This system was able to correctly predict ≈93% (2871/3083) of disease mentions from the test set that also appeared in the training. Out of 776 mentions that the model had not previously seen, it was able to properly predict 511 (≈66%).

Error analysis. In common with similar biomedical entity recognition tasks, the longer the mentions, the more difficult is for models to predict them correctly (Augenstein et al., 2017). In SocialDisNER we have found a correlation of -0.24 between the prediction errors of the systems and the length of the mentions, with a tendency to error when the length of the mention increases. Regarding specific issues of detection, we have detected 4 common detection problems in several participating systems:

1. *Difficulty recognizing capitalized mentions*: Systems are able to detect a lowercase mention, but not its uppercase version if they have not seen it before in the training set.

2. *Mentions containing punctuation marks and/or special Twitter characters*: Mentions with internal punctuation marks are difficult to be correctly extracted by systems. They are usually detected as several independent mentions or detecting only one of the segments.

3. *Composite mentions*: Mentions that refer to more than one disease using conjunctions or prepositions are noted as a single mention, but participating systems tend to split such mentions when extracted.

4. *Detection of mention boundaries*:The systems developed in the task predict longer mentions than expected when there are symbols such as hashtag "#" or emojis before and/or after the mention.

Some examples of these errors can be found in Table 3 in appendix A. The systems also fail to properly detect mentions semantically similar to those seen in the training phase, but expressed with another verbal construction. For example, the models are able to detect "suicide" but not "thinking about suicide"

5 Discussion

SocialDisNER is the first task focused on extracting diseases from social media content written in Spanish. The first Gold Standard corpus of tweets with diseases annotated by medical experts has been built specially for the task. The corpus documents were selected to contain first-person patient experiences and relevant biomedical information written by experts.

We also published additional resources such a large-scale Silver Standard corpus annotated with additional biomedical entities that might be used to train systems to detect entities in Spanish tweets that previously could not be detected due to lack of resources. This large-scale corpus made it possible to generate co-mention networks that can be used toward Knowledge Graph mining to replicate studies on adverse drug effects in Spanish-speaking people (Nikfarjam et al., 2015). A knowledge graph similar to the one in Figure 2 might allow descriptive analysis of disease co-morbidities, to detect new symptoms in rare diseases, to discover diseases associated with specific professions, and even to discover adverse effects of biomaterials and prosthetics.

SocialDisNER has been very well received by the community. There have been participants from the academy and the industry, probably enticed by the methodology followed to collect content, which ensured that a significant percentage of the documents were of high relevance. Nevertheless, the modularity with which the methodology has been defined enables its improvement for future shared-tasks and projects. For example, more sophisticated first-person detection systems such as those developed in Al-Garadi et al. (2020) could be used, based on manual annotation of previous data. NER systems could also be trained with the large-scale corpus to improve the data enrichment process. This data selection methodology to retrieve relevant information could also be easily transferred to other languages as it relies on the presence of patient associations on Twitter, which is relatively common.

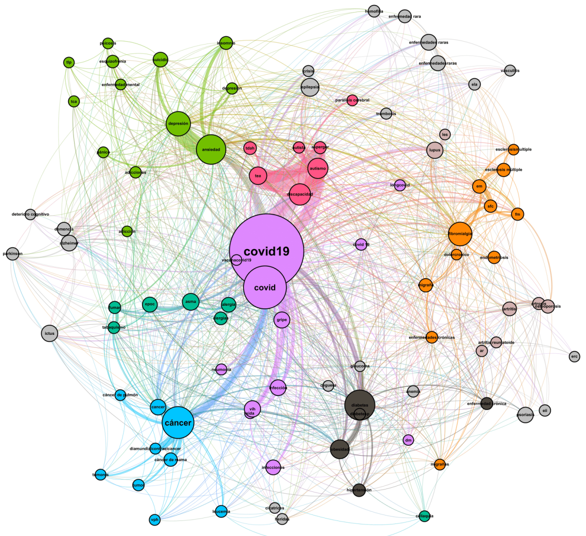

Figure 2: Simplified SocialDisNER co-morbility network,

When organizing the task, we contacted several patient associations with Twitter accounts. These groups showed interest in SocialDisNER, its results and how the output of the task may benefit the patients. The interest shown by associations, which in many cases helped to disseminate the event, shows the importance of involving all relevant stakeholders during the development and organization phases in order to increase the impact and use cases of our work.

Acknowledgements

We acknowledge the Encargo of PlanTL (SEDIA) to BSC for supporting SocialDisNER and the scientific committee for their valuable guidance and help. Due to the relevance of detecting disease content and symptoms from social media to analyze adverse effects of prosthetics and biomaterials, this work has been supported by BIOMATDB, a European Unions Horizon Europe Coordination Support Action under Grant Agreement No 101058779. Due to the relevance of the detection of diseases related to work environments, this project has also been supported by the AI4ProfHealth project (PID2020-119266RA-I00). Icons from Figure 1 created by Flaticon.

References

Luca Maria Aiello, Daniele Quercia, Ke Zhou, Marios Constantinides, Sanja Šćepanović, and Sagar Joglekar. 2021. How epidemic psychology works on twitter: Evolution of responses to the covid-19 pandemic in the us. *Humanities and Social Sciences Communications*, 8(1):1–15.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2020. A text classification approach for the automatic detection of twitter posts containing self-reported covid-19 symptoms.

Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. 2021. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ data science*, 10(1):15.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Andrei-Marius Avram, Vasile Pais, and Maria Mitrofan. 2022. Racai@smm4h'22: Tweets disease mention detection using a neural lateral inhibitory mechanism. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 1–3.

Jose Castano, Maria Laura Gambarte, Carlos Otero, and Daniel Luna. 2022. A simple terminology-based approach to clinical entity recognition.

Kendrick Cetina and Nuria García-Santa. 2022. Fre at socialdisner: Joint learning of language models for named entity recognition. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 68–70.

Mariia Chizhikova, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2022. Sinai@smm4h'22: Transformers for biomedical social media text mining in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 27–30.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Eulàlia Farré-Maduell, Salvador Lima, Luis Gascó, Antonio Miranda-Escalada, and Martin Krallinger. 2022. SocialDisNER Guidelines: detection of disease mentions in spanish social media content. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Jia Fu, Sirui Li, Hui Ming Yuan, Zhucong Li, Zhen Gan, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2022. Casia@smm4h'22: A uniform health information mining system for multilingual social media texts. In *Proceedings of the Seventh Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 143–147.

Luis Gasco, Chloé Clavel, Cesar Asensio, and Guillermo de Arcas. 2019. Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise. *Science of The Total Environment*, 658:69–79.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. 2020. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*.

Javier Huertas-Tato, Alejandro Martin, and David Camacho. 2022. Bertuit: Understanding spanish language in twitter through a native transformer. *arXiv preprint arXiv:2204.03465*.

ITAINNOVA. 2022. Socialdisner code. https://github.com/ITAINNOVA/SocialDisNER.

Akbar Karimi and Lucie Flek. 2022. Caisa@smm4h'22: Robust cross-lingual detection of disease mentions on social media with adversarial methods. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 168–170.

Simon Kemp. 2021. [link].

Veysel Kocaman, Cabir Celik, Damla Gurbaz, Gursev Pirge, Bunyamin Polat, Halil Saglamlar, Meryem Vildan Sarikaya, Gokhan Turer, and David Talby. 2022. John_snow_labs@smm4h'22: Social media mining for health (#smm4h) with spark nlp. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 44–47.

Dominik Kowald and Elisabeth Lex. 2018. Studying confirmation bias in hashtag usage on twitter. *arXiv preprint arXiv:1809.03203*.

Antoine Lain, Wonjin Yoon, Hyunjae Kim, Jaewoo Kang, and Ian Simpson. 2022. Ku_ed at socialdisner: Extracting disease mentions in tweets written in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

Andrew MacKinlay, Hafsah Aamer, and Antonio Jimeno Yepes. 2017. Detection of adverse drug reactions using medical named entities on twitter. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1215. American Medical Informatics Association.

Eugenio Martinez-Camara, M Teresa Martín-Valdivia, L Alfonso Urena-Lopez, and Ruslan Mitkov. 2015. Polarity classification for spanish tweets using the cost corpus. *Journal of Information Science*, 41(3):263–272.

Emily G Miller, Amanda L Woodward, Grace Flinchum, Jennifer L Young, Holly K Tabor, and Meghan C Halley. 2021. Opportunities and pitfalls of social media research in rare genetic diseases: a systematic review. *Genetics in Medicine*, 23(12):2250–2259.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 13–20.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

Rosa M. Montañés-Salas, Irene López-Bosque, Luis García-Garcés, and Rafael del Hoyo-Alonso. 2022. Itainnova at socialdisner: A transformers cocktail for disease identification in social media in spanish. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 71–74.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2022. Overview of bioasq 2022: The tenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 337–361, Cham. Springer International Publishing.

Azadeh Nikfarjam, Abeed Sarker, Karen Oconnor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Miguel Ortega-Martín, Alfonso Ardoiz, Jorge Álvarez, Oscar García-Sierra, and Adrián Alonso. 2022. dezzai@smm4h'22: Tasks 5 10 - hybrid models everywhere. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

P Otero, J Gago, and P Quintas. 2021. Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution. *Marine Pollution Bulletin*, 170:112620.

Karen OConnor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

PLN-CMM. 2022. Socialdisner. https://github.com/plncmm/socialdisner.

Beatrice Portelli, Simone Scaboro, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. Ailab-udine@smm4h'22: Limits of transformers and bert ensembles. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 130–134.

RACAI. 2022. Rner. https://github.com/racai-ai/RNER.

READ-BioMed. 2022. socialdisner-2022. https://github.com/READ-BioMed/socialdisner-2022.

Matias Rojas, Jose Barros, Kinan R. Martin, Mauricio Araneda-Hernandez, and Jocelyn Dunstan. 2022. Pln cmm at socialdisner: Improving detection of disease mentions in tweets by using document-level features. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 52–54.

SINAI. 2022. Spanish_disease_finder. https://huggingface.co/chizhikchi/Spanish_disease_finder.

Aman Sinha, Cristina Garcia Holgado, Marianne Clausel, and Matthieu Constant. 2022. Iai @ socialdisner : Catch me if you can! capturing complex disease mentions in tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 85–89.

Antonio Tamayo. 2022. Nlp-cic-wfu at socialdisner: Disease mention extraction in spanish tweets using transfer learning and search by propagation. https://github.com/ajtamayoh/NLP-CIC-WFU-Contribution-to-SocialDisNER-shared-task-2022.

Antonio Tamayo, Diego Burgos, and Alexander Gelbukh. 2022. Nlp-cic-wfu at socialdisner: Disease mention extraction in spanish tweets using transfer learning and search by propagation. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 19–22.

TeMU-BSC. 2022. Socialdisner-st baseline 1 - lookup. https://github.com/TeMU-BSC/socialdisner_baseline_lookup.

Harsh Verma, Parsa Bagherzadeh, and Sabine Bergler. 2022. Claclab at socialdisner: Using medical gazetteers for named-entity recognition of disease mentions in spanish tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 55–57.

Davy Weissenbacher, Juan M. Banda, Vera Davydova, Darryl Estrada-Zavala, Luis Gascó, Yao Ge, Yuting Guo, Ari Z. Klein, Martin Krallinger, Mathias Leddin, Argun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Ana Lucía Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*.

Antonio Jimeno Yepes and Karin Verspoor. 2022. Readbiomed@socialdisner: Adaptation of an annotation system to spanish tweets. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 48–51.

A Appendix prediction errors

Name	Tweet with mention	Example of participants extractions
Capitalized mentions	LAS CICATRICES No hay cicatriz, (...)	Systems predict the mention "cicatrices", but not "CICATRICES", only present in the test set.
	(...) NO SOMOS DEPRESIVAS, TENEMOS DEPRESIÓN! Sabías (...)	Systems predict "depresión", but not "DEPRESIÓN"
Mentions with punctuation marks and/or special characters	visibilidad para esta enfermedad crónica asociada al #dolor	*enfermedad crónica* *enfermedad crónica* and *dolor*]
	(...)teniendo una enfermedad crónica (Crohn) y en tratamiento(...)	*enfermedad crónica* *enferemedad crónica (Crohn* *enfermedad cronica* and *Crohn*
	(...)Antraciclinas en Her2+ en ca de mama temprano(...)	*ca de mama* *ca de mama temprano* *her2* and *ca de mama*
Composite mentions	(...)debido a una malformación o disfunción de los órganos que(...)	*malformación* *disfunción de los órganos* *malformación* and *disfunción de los órganos* *malformación o disfunción de los órganos*
	(...)cuidar a su marido con cáncer y metástasis. No(...)	*cancer* and *metástasis*
Detection of mention boundaries	(...)y el consiguiente daño NEURONAL🧠...)	*daño NEURONAL*🧠
	(...)lo agradecemos 👏👏👏😂 ✌#ConferenciaCovid19 (...)	*ConferenciaCovid19*

Table 3: Example of prediction errors of SocialDisNER systems