Identifying professions & occupations in Health-related Social Media using Natural Language Processing

J. Alberto Mesa Murgado and **Ana Belén Parras Portillo** and **Pilar López-Úbeda** and **M. Teresa Martín-Valdivia** and **L. Alfonso Ureña López**
SINAI Research Group - CEATIC - Universidad de Jaén
Campus Las Lagunillas s/n. E-23071, Jaén, Spain
{jmurgado,abparras,plubeda,maite,laurena}@ujaen.es

Abstract

This paper describes the entry of the research group SINAI at SMM4H's ProfNER task on identifying professions and occupations in social media data related to health. Specifically, we participated in Task 7a: Tweet Binary Classification to determine whether a tweet contains mentions of occupations or not and also in Task 7b: NER Offset Detection and Classification aimed at predicting occupations mentions and classify them as either professions or working statuses.

1 Introduction

Natural Language Processing (NLP) and Machine Learning (ML) techniques are becoming essential in critical fields such as the one of healthcare, considering that they perform tasks faster than a human agent and at a very high level of reliability. Some of these tasks include the automatic assignment of International Classification of Diseases (ICD) codes to health related texts (Perea-Ortega et al., 2020) or the detection of negative and positive emotions in medical documents (Plaza-del Arco et al., 2019).

Automatic text classification and Named Entity Recognition (NER) are two tasks in which NLP has proved to have a relevant impact. In both cases we are given a certain set of documents and while for the first task we aim to classify them distinguishing by a certain criteria, the second seeks to detect and tag specific entities.

The Social Media Mining for Health (SMM4H) 2021 ProfNER Shared Task (Miranda-Escalada et al., 2021) emphasizes the importance of identifying professions and occupations within social media content related to health, this knowledge could later be applied to determine which of them are at risk due to direct exposure to the COVID-19 pandemic and/or state what professional sectors are more prone to mental health issues due to the uncertainty of the current situation.

This issue has been further subdivided into two tracks: determine whether the social media textual content contains mentions of professions or not (a binary classification task) and to identify professions and working statuses within the text in order to extract its text span and tag it accordingly (NER). Our research group has used NLP and ML approaches for both tasks in combination with two dictionaries which we have developed.

Considering this information, this paper is structured as follows: section 2 introduces work related to this challenge and research field. Section 3 briefly describes the dataset provided and its characteristics. Section 4 states the systems we have developed for each task. Section 5 exhibits the results from our systems using the test dataset and finally, in Section 6 we present our conclusions and future work.

2 Related work

Social media plays an important role where people can share information related to health. This information can be used for public health monitoring tasks through the use of NLP techniques.

On one hand, in terms of document classification in the medical field, many researchers have used social networks as a source of information to develop and evaluate systems. For example, to predict mental illnesses such as depression or anorexia (Aldarwish and Ahmad, 2017; López-Úbeda et al., 2021) and to detect nonmedical prescription medication (Al-Garadi et al., 2021). More recently, new studies analyzed health, psychosocial, and social issues emanating from the COVID-19 pandemic from social network comments using NLP (López-Úbeda et al., 2020a; Müller et al., 2020; Oyebode et al., 2020).

On the other hand, several NER systems have been developed using NLP-based systems such as MedLEE (Friedman, 1997), MetaMap (Aronson and Lang, 2010) and cTAKES (Savova et al.,

1
Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 141–145
June 10, 2021. ©2021 Association for Computational Linguistics

2010)). Most of these are rule-based systems that use extensive medical vocabularies. Current state-of-the-art approaches to the NER task propose the use of RNNs to learn useful representations automatically because they facilitate the modeling of long-distance dependencies between words in a sentence (López-Úbeda et al., 2019; López-Úbeda et al., 2020b).

Since there is currently a great growth in demand for classification and extraction of information from medical texts, the NLP community has organized a series of open challenges with a focus on biomedical entity extraction and document classification tasks such as DDIExtraction (Segura Bedmar et al., 2013), the N2C2 - National NLP Clinical Challenges shared task (Henry et al., 2020) and the CHEMDNER challenge (Krallinger et al., 2015). Finally, SMM4H (Weissenbacher et al., 2019) provided tasks for the extraction of adverse effects using Twitter as a source of information. In this workshop, participants were first required to identify whether a tweet contained an Adverse Drug Reaction (ADR). Subsequently, the challenge provided the NER task to locate the specific ADR.

The use of Spanish as the main language of a challenge has emerged in recent years providing important workshops such as the DIANN (Fabregat et al., 2018) (Disability Annotation Task) task, PharmaCoNER (Agirre et al., 2019) (Pharmacological Substances, Compounds and proteins and NER), Cantemist (Miranda-Escalada et al., 2020) and eHeatlh-KD (Piad-Morffis et al., 2020) (eHealth knowledge discovery).

3 Dataset

Organizers provided us with a dataset consisting of 8,000 tweets from Twitter subdivided into two subsets: 6,000 tweets with which to train our systems and 2,000 tweets to validate them. Namely, train set and dev set, accordingly.

Besides the tweet's identifier and text span, a binary value was used to determine whether it contained a profession or not as well as its corresponding annotated entities tagged using the Inside-outside-beginning (IBO) format.

To compare the performance of all the systems presented at this shared task, we were provided with another dataset, namely test set, consisting of 2,000 processed tweets and 25,000 raw tweets for the background set.

4 Methodology

For our participation we employ ML models enriched with custom made dictionaries consisting of 776 professions such as *"Farmacéutico"* (Pharmacist), *"Dentista"* (Dentist), *"Cajera"* (Cashier) and *"Veterinario"* (Vet) recovered from the *"Listado de profesiones reguladas en el ámbito sanitario"*[1] provided by the *Ministerio de de Sanidad y Política Social* of the Spanish Government and from the occupations listed by the European Commission in their International Standard Classification of Occupations (ISCO)[2].

The second dictionary contains 26 working statuses including *"Autónomo"*, *"Funcionario"* (Public employee) and *"Erte"* (record of temporary Employment regulation) among others based on the Workshop's annotation guidelines[3].

4.1 Task 7a: Binary Classification

To classify tweets whether if they contained mentions of professions or not, we used two approaches: a Support Vector Machine (SVM) and bidirectional Long Short Term Memory (BiLSTM) Recurrent Neural Network (RNN), both combined with our professions dictionary.

The SVM approach applies the scikit-learn library (Pedregosa et al., 2011) using its default parameters, as stated in the documentation, the words for each tweet, also stated as document, were transformed into vectors considering the frequency of the terms within each document (TFIDF). This structure was later enriched using our professions dictionary through a vector for each document consisting of as much binary values as the number of terms in the dictionary such as each binary value represented if the term within the document was in the dictionary or not.

The BiLSTM model is implemented using the Tensorflow library (Abadi et al., 2015) which we explored through different batch sizes, ranging from 2^7 (2^6 and 2^8), and different number of epochs. Although the best accuracy obtained for training (0.8695) determined a batch size of 128 and 5 epochs. This model makes use of the GloVe (Pennington et al., 2014) word embeddings 200d vector,

[1] https://www.mscbs.gob.es/eu/profesionales/formacion/docs/Anexo_X_del_Real_Decreto_1837.pdf

[2] https://ec.europa.eu/esco/portal/occupation

[3] https://zenodo.org/record/4306017#.YE3vpJ1KhhE

pre-trained using Twitter's tweets, this approach is also uses our professions dictionary.

4.2 Task 7b: NER offset detection and classification

To detect and classify entities within those same tweets and retrieve them discerning the text span of the entities as well as their initial and end position within the text, we opted for a Conditional Random Fields (CRF) approach.

This approach was implemented using scikit-learn crfsuite library (Wijffels and Okazaki, 2007-2018), transforming words to features and using the words that came before and after each term, then enriched the system using both our dictionaries (added features as binary values). For this implementation we considered the L-BFGS method for the gradient descent, a 100 iterations and values 0.1 for both c1 and c2.

Our system assigned an IBO tag to each term of every tweet and we later searched for the entity text span within the same tweet to extract its initial and end position.

5 Evaluation results

We have been provided with the median of the participants' score using three metrics: precision (P), recall (R) and F1-scoring (F1) (Magge et al., 2021). Using them we built 2 tables consisting of two sections: the upper section shows the score obtained by our systems on the test set, while the latter fits the same purpose for the dev set.

5.1 Task 7a: Binary Classification

We submitted 2 runs for the evaluation phase: a combination of SVM (SVM+Dic) and an BiLSTM model (BiLSTM+Dic), both combined with our professions dictionary, the scoring associated to these systems applied to the provided datasets is displayed in Table 1.

Model	P	R	F1
Median	**0.9185**	**0.8553**	**0.85**
BiLSTM+Dic	0.7612	0.4752	0.59
SVM+Dic	0.8995	0.4255	0.58
BiLSTM+Dic	0.77	**0.72**	**0.74**
SVM+Dic	**0.86**	0.70	**0.74**
SVM	**0.86**	0.64	0.66

Table 1: Scores obtained by our systems on the SMM4H ProfNER Shared Task - Task 7a (binary classification) applied over the test and dev set, accordingly.

While the performance of our systems on the training dataset was, at average, close to 0.74 (F1), on the test set it was decreased in a 21%. Therefore, resulting in a value close to 0,59 (again, F1), 31% below the median.

5.2 Task 7b: NER Offset Classification

We were closer to the median and in line with what our systems obtained on the training dataset (1% decrease in performance compared to the test set). These results are displayed in Table 2 in the same way in which Table 1 was: the upper section refers to the scoring for our systems on the test set while the latter, exhibits the scoring for the dev set.

Model	P	R	F1
Median	**0.842**	**0.7265**	**0.7605**
CRF+Dic	0.824	0.652	0.728
CRF+Dic	**0.861**	**0.647**	**0.739**
CRF	0.852	0.597	0.702

Table 2: Score obtained by our systems on the SMM4H ProfNER Shared Task - Task 7b (NER) applied over the test and dev set, accordingly.

6 Conclusion

For our participation in the SMM4H Task 7 ProfNER Shared Task on identifying professions and occupations we implemented three systems. First two are aimed at the binary classification task (Task A) using an SVM and a BiLSTM approach, both combined with our professions dictionary. The latter system follows a CRF approach combined with our professions and working statuses dictionaries and it is applied to the NER task (Task B).

Our predictions for the training set were consistent with those obtained on the test set for the second task (NER) whereas our approaches for the first task (binary classification) fell short of our expectations by 21% below our training results.

For future work, we will use the gold test in order to perform a deeper analysis to assess why did this event happened and therefore improve the performance of our systems.

Acknowledgements

This work has been partially supported by the LIVING-LANG project [RTI2018-094653-B-C21] of the Spanish Government and the Fondo Europeo de Desarrollo Regional (FEDER).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. 2021. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC medical informatics and decision making*, 21(1):1–13.

M. M. Aldarwish and H. F. Ahmad. 2017. Predicting depression levels using social media posts. In *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, pages 277–280.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Hermenegildo Fabregat, Juan Martinez-Romo, and Lourdes Araujo. 2018. Overview of the DIANN Task: Disability Annotation Task. In *IberEval@ SEPLN*, pages 1–14.

Carol Friedman. 1997. Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, 7(1):S1.

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2020a. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066.

Pilar López-Úbeda, Manuel Carlos Díaz Galiano, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. 2019. Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 102–106.

Pilar López-Úbeda, José M Perea-Ortega, Manuel C Díaz-Galiano, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2020b. Sinai at ehealth-kd challenge 2020: Combining word embeddings for named entity recognition in spanish medical records.

Pilar López-Úbeda, Flor Miriam Plaza-del Arco, Manuel Carlos Díaz-Galiano, and Maria-Teresa Martín-Valdivia. 2021. How successful is transfer learning for detecting anorexia on social media? *Applied Sciences*, 11(4):1838.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, 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 Applications Workshop & Shared Task*.

A Miranda-Escalada, E Farré, and M Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

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

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Oladapo Oyebode, Chinenye Ndulue, Ashfaq Adib, Dinesh Mulchandani, Banuchitra Suruliraj, Fidelia Anulika Orji, Christine Chambers, Sandra Meier, and Rita Orji. 2020. Health, psychosocial, and social issues emanating from covid-19 pandemic based on social media comments using natural language processing. *arXiv preprint arXiv:2007.12144*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

José M Perea-Ortega, Pilar López-Úbeda, Manuel C Díaz-Galiano, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2020. Sinai at clef ehealth 2020: testing different pre-trained word embeddings for clinical coding in spanish.

Alejandro Piad-Morffis, Yoan Gutiérrez, Hian Cañizares-Diaz, Suilan Estevez-Velarde, Rafael Muñoz, Andres Montoyo, Yudivian Almeida-Cruz, et al. 2020. Overview of the ehealth knowledge discovery challenge at iberlef 2020. CEUR.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2019. Sinai at semeval-2019 task 3: Using affective features for emotion classification in textual conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 307–311.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30.

Jan Wijffels and Naoaki Okazaki. 2007-2018. crfsuite: Conditional random fields for labelling sequential data in natural language processing based on crfsuite: a fast implementation of conditional random fields (crfs). R package version 0.1.