Romanian micro-blogging named entity recognition including health-related entities

Vasile Păiș, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Carol Luca Gasan, Roxana Micu
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
vasile,vergi,elena,maria@racai.ro

Abstract

This paper introduces a manually annotated dataset for named entity recognition (NER) in micro-blogging text for the Romanian language. It contains gold annotations for 9 entity classes and expressions: persons, locations, organizations, time expressions, legal references, disorders, chemicals, medical devices and anatomical parts. Furthermore, word embedding models computed on a larger micro-blogging corpus are made available. Finally, several NER models are trained and their performance is evaluated against the newly introduced corpus.

1 Introduction

Social media networks can present an alternative to formal data sources, and prove to be a reliable resource for different natural language processing (NLP) tasks. Since users often submit domain-specific information in a wide variety of social networking resources, such as specific communities, blogs, microblogs, public news websites, or forums, there has been a significant interest in extracting information from these resources. Micro-blogging platforms, such as Twitter, Reddit or Gab, as a source of such comments, in general, contain relevant information that can be used to develop NLP resources for both general and specific domains, such as health (dos Santos et al., 2020) and legal aspects (Altoaimy, 2018). Furthermore, mining and studying health-related information can ultimately be used to improve public health (Magge et al., 2021).

Our contribution is threefold. First, we introduce the MicroBloggingNERo corpus (Păiș et al., 2022a), comprised of micro-blogging specific messages, manually annotated with 9 named entity (NE) classes and expressions, including 4 health-related classes. This is the largest available Romanian micro-blogging NE corpus to date and the only one containing health-related entities. Second, we release word embeddings computed on a larger micro-blogging corpus. Finally, we train several NER models using the newly released corpus and we make these available for the research community.

This paper is organized as follows: Section 2 presents related work, Section 3 describes the manually annotated dataset, and Section 4 introduces the experiments performed, including the generated word embedding representations and NER models. Finally, Section 5 gives conclusions.

2 Related work

Biomedical corpora have been collected for languages with different extent of technological support: Hmong (White, 2022), Romanian (Mitrofan et al., 2019), Spanish (Carrino et al., 2021) and others. There are even parallel medical corpora available[1]. Many of them are annotated with UMLS medical entities (Mohan and Li, 2019; Ohta et al., 2002).[2] Most of the biomedical corpora were collected from abstracts (Ohta et al., 2002), but also from full articles (Alex et al., 2008; Bada et al., 2012); others combine various sources (Mitrofan et al., 2019).

The interest in gathering corpora made up of micro-blogging texts characterizes the research dedicated to various languages: Arabic (Zaatari et al., 2016), Chinese (Wang et al., 2012), English (Sharma et al., 2020), French (Mazoyer et al., 2020), German and Danish (Bick, 2020), Italian (Sanguinetti et al., 2018), Romanian (Manolescu and Çöltekin, 2021; Ciobotaru et al., 2022), Turkish (Çöltekin, 2020), etc.

Micro-blogging corpora are used for various tasks: sentiment analysis application development (Sharma et al., 2020; Cieliebak et al., 2017), emotion annotation (Roberts et al., 2012), credibility

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus

[2] Such a collection is available at https://github.com/BaderLab/Biomedical-Corpora.

Proceedings of the 29th International Conference on Computational Linguistic, pages 190–196
October 12–17, 2022.

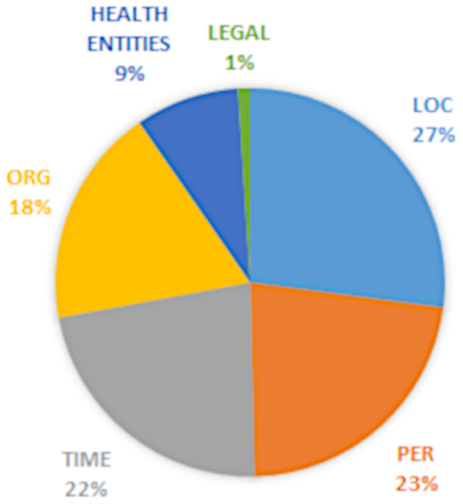

Figure 1: Label distribution of entity classes.

analysis (Zaatari et al., 2016), event detection (Mazoyer et al., 2020), hate speech detection (Bick, 2020; Çöltekin, 2020; Manolescu and Çöltekin, 2021; Sanguinetti et al., 2018) and others. Another vivid line of research is identifying certain types of entity names in these corpora, such as occupations (Miranda-Escalada et al., 2021; Santamaría Carrasco and Cuervo Rosillo, 2021) and symptoms (Kumar et al., 2021).

In this work, we developed MicroBloggingNERo the first Romanian micro-blogging corpus enriched with health information.

3 Dataset

The MicroBloggingNERo dataset contains 7,800 messages manually annotated with the following named entity classes and expressions, totalling 11,099 annotations (see figure 1):

- Organization (ORG) entities representing companies, agencies, and political parties. Examples: *Facebook*, *Guvernul* (“the government”), *PSD* (Partidul Social Democrat “the Socialist Democratic Party”), *#ConsConRo*.
- Person (PER) entities are regularly limited to humans, but may include fictional characters and references to religious figures. Examples: *Adela*, *Moș Crăciun* (“Santa Claus”), *Niculina Stoican*.
- Location (LOC) entities represent addresses, geographical areas and landmasses, bodies of water, and geological formations, denoted by a proper name. Examples: *România* (“Romania”), *Parcul Tineretului* (“the Youth Park”), *Lacul Sfânta Ana* (“Lake Saint Ana”).
- Time (TIME) expressions identify when something happened, how long something lasted, or how often something occurs. Sometimes the precise date cannot be determined, allowing for expressions indicating periods of time. Examples: *astăzi* (“today”), *15 septembrie* (“September 15”), *Crăciun* (“Christmas”).
- Legal references (LEGAL) are expressions indicating a legal document. Examples: *legea 13/2021* (“law 13/2021”), *constituția* (“the Constitution”).
- Anatomical parts (ANAT) class marks parts of the human body, organs, components of organs, tissues, cells, cellular components. Examples: *cap* (“head”), *mâini* (“hands”), *ficat* (“liver”).
- Chemical and drugs (CHEM) class contains mentions of amino acids, peptides, proteins, antibiotics, active substances, drugs, enzymes, hormones, receptors. Examples: *sodiu* (“sodium”), *vaccin* (“vaccine”).
- Disorders (DISO) class is intended primarily for diseases but includes also things such as anatomical abnormalities, congenital anomalies, syndromes, lesions, symptoms. Examples: *diabet* (“diabetes”), *COVID*.
- Medical devices (MED_DEVICE) class contains mentions of any device intended to be used for medical purposes. Example: *stetoscop* (“stethoscope”).

The dataset was gathered by using a crawling process based on specific queries, aiming to gather messages containing the entities of interest. It was then cleaned by removing duplicates and very short messages (too short to be actual sentences) or messages containing only hashtags. Furthermore, all URLs were replaced with a special “<url>” tag.

The annotation process involved 7 annotators working under the guidance and supervision of 3 senior researchers. Parts of the corpus were common between several of the annotators, allowing us to compute inter-annotator agreement. Each pair of annotators had to annotate 300 common files. The Cohen Kappa coefficient, indicating the inter-annotator agreement was computed at the token level and led to an average Kappa of 0.80. According to (Cohen, 1960), a value between 0.61–0.80 indicates substantial agreement between the annotators. Moreover, Landis and Koch (1977) stated that a value of Cohen Kappa coefficient greater than 0.81 represents an almost perfect agreement

and Fleiss et al. (2013) also characterizes kappas over 0.75 as excellent.

The remaining disagreements account for mistakes made by individual annotators when specific domain information was needed in order to identify and classify mainly entities from both medical and legal domains. For example, medical entities such as "virus" was annotated with labels "DISO" or "CHEM". In most situations the correct label is "DISO", however when the mechanism of the action of the virus is explained the correct label is "CHEM", thus creating confusion for the annotators. A similar situation can be found in the case of "vaccine" mentions when this entity was annotated with both "CHEM" and "MED_DEVICE" labels. In this case the correct label is "CHEM", but some annotators considered it "MEDICAL_DEVICE" because the term was used in a context of diagnosing Covid-19 together with tests, masks, gowns, gloves, sterilizers, and ventilators. Entity classes such as ORG, PER, LOC and TIME have an inter-annotator agreement greater than 0.85.

The annotation scheme, as well as the guidelines, were inspired by existing Romanian NE corpora (built without social media text), such as MoNERo (Mitrofan et al., 2019; Mitrofan, 2017), SiMoNERo (Mitrofan, 2019) and LegalNERo (Păiș et al., 2021; Păiş et al., 2021). The annotation guidelines[3] for the MicroBloggingNERo corpus had to be adjusted to take into account the particularities (Păiș et al., 2022b) of micro-blogging text, such as hashtags (*#Cluj, #LaculNoua*), text written with emojis instead of numbers, unusual abbreviation, elongated words (commonly used for emphasis in microblogging), code-mixed text, etc.

Since the MicroBloggingNERo corpus was created with multiple types of annotations, the health-related entities represent only 9% of the total entity classes, as depicted in Figure 1. This accounts for 958 annotations, half of them being in the DISO class (466 annotations).

After the annotation process was completed, the corpus was anonymized by replacing all person names, specific locations (such as addresses or street numbers), and organizations with randomly generated ones. Any user mentions were removed and replaced with a special tag "<user>", regardless of how it was included in the original message. Furthermore, any other identifiers (such as message IDs) were removed from the corpus.

The dataset was released with several usage scenarios in mind, reflected in the presence of multiple formats, including span-based annotations with all the entities, span-based annotations with general entities (PER, LOC, ORG, TIME), span-based annotations for the bio-medical domain, and tokenized text with token-based NE annotations attached (using tools available inside the RELATE platform (Păiș et al., 2019, 2020; Păiș, 2020)).

4 Experiments

We used the presented NER annotated micro-blogging dataset to train NER models, using different architectures and configurations. For this purpose, the corpus was split into train (70%), valid (15%) and test (15%). The splits were selected randomly while trying to preserve the general distribution of entities in each split. They are made available within the corpus release. Then, we experimented with classical word embedding representations. For this purpose, we used a recurrent neural network architecture, implemented with Long Short-Term Memory (LSTM) cells, as provided by the NeuroNER package (Dernoncourt et al., 2017). We used pre-trained word embeddings (Păiș and Tufis, 2018) trained on the Representative Corpus of the Contemporary Romanian Language (CoRoLa) (Tufiș et al., 2019). Then, we trained new word representations based on a larger corpus of raw micro-blogging text (comprising 853k messages), using the FastText tool (Bojanowski et al., 2017). Finally, based on observations such as those of (Păiș and Mitrofan, 2021), indicating that multiple representations can help the system achieve better results, we combine the two representations into a single, larger representation. We make the generated embedding models freely available[4].

Following the experiments with static word embeddings, we performed two additional experiments using contextualized embeddings provided by the XLM-RoBERTa model (Conneau et al., 2020). This model provides contextualized representations for multiple languages, including Romanian, and has proven to be a good choice for different NER experiments (Malmasi et al., 2022; Shaffer, 2021; Adelani et al., 2021; Suppa and Jarabka, 2021). Our experiments made use of a basic NER system, employing a linear layer followed by

[3] https://relate.racai.ro/resources/microblogging/Annotation_Guide_MicroBloggingNERo.pdf

[4] https://relate.racai.ro/resources/microblogging

System	ANAT	CHEM	DISO	LEG	LOC	MED DEV	ORG	PER	TIME	Total	Ep.
Neuroner CoRoLa	42.86	60.47	75.47	45.71	77.69	72.73	66.21	84.18	63.96	72.03	23
Neuroner Microblogging	22.54	58.82	71.43	47.37	81.27	61.54	65.95	80.95	63.64	71.26	28
Neuroner CoRoLa+ MicroBlogging	21.43	52.87	73.47	36.36	81.21	66.67	62.00	83.51	61.50	70.75	32
XLM-RoBERTa	**49.46**	**70.97**	**82.00**	68.29	86.88	62.50	**77.37**	88.56	68.32	**78.96**	35
XLM-RoBERTa with LI	45.24	67.42	81.00	**68.42**	**87.05**	**76.92**	75.39	**88.70**	**68.50**	78.62	61

Table 1: Results (% F1 scores) from different experiments using the MicroBloggingNERo corpus

a classification head, and the same NER system enhanced with a biologically inspired lateral inhibition layer, as introduced by (Păiș, 2022). This system was previously used for NER in the Romania bio-medical domain (Mitrofan and Păiș, 2022). Intuitively, similar to the biological process, we considered that the lateral inhibition layer would allow the system to better focus on difficult, and less represented, NE classes. Results are given in Table 1, in terms of F1 percent scores, while also indicating the training epoch associated with each model.

The results indicate that Romanian NER in a micro-blogging context is still far from achieving high-scores. Currently, even systems making use of contextualized embeddings do not achieve over 90% F1 score for any of the classes. On regular texts, previous works, such as (Păiş et al., 2021), indicate over 90% F1 scores for persons and legal references. Furthermore, Mitrofan and Păiș (2022) show an F1 score of 85% for health-related entities in regular Romanian texts.

An analysis of the errors associated with the best performing model, with regard to health-related entities, indicates that a large number of entities are not recognized (the predicted label is "O"). This suggests a need for a larger corpus with more diverse entities. Other types of errors account for predicting ORG as MED_DEVICE or CHEM, which is an error found also with the human annotators, where a company name is used also for a vaccine name.

5 Conclusion

Micro-blogging texts pose unique challenges in terms of natural language processing. This is particularly true with regard to bio-medical and other health-related entities. These are more difficult to detect when compared to more common NEs, such as person and organization names, due to their inherent complexity. Erhardt et al. (2006) consider this to be a general issue with regard to information extraction in the biomedical domain, given the complex entities and changing nomenclatures. This was also the case in the process of manual annotation, annotators having difficulties to identify and classify both health-related and legal NEs. Furthermore, as noted by Klein et al. (2020), imbalanced data remains a challenge for training deep neural network models. This is particularly true with the MicroBloggingNERo dataset, considering the distribution of classes depicted in Figure 1. Furthermore, this suggests that future work should focus on expanding the dataset, aiming for a more balanced class distribution.

When looking at micro-blogging word embeddings, these seem to be beneficial for recognizing certain entity classes, possibly accounting for spelling particularities in micro-blogging texts. General representations, trained on the large CoRoLa corpus, provide better results when detecting anatomical parts, chemicals and medical devices. This is due to the presence of bio-medical texts in the CoRoLa corpus, as well as to the reduced number of mentions in the micro-blogging corpus. Future research will explore more word representation models, particularly those trained on larger micro-blogging corpora. Currently, there is no contextualized word model available for the Romanian language built on micro-blogging texts or specifically on health-related texts, such as BioBERT (Lee et al., 2019).

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. The iti txm corpora: Tissue expressions and protein-protein interactions.

Lama Altoaimy. 2018. Driving change on twitter: A corpus-assisted discourse analysis of the twitter debates on the saudi ban on women driving. *Social Sciences*, 7(5):81.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the craft corpus. *BMC Bioinformatics*, 13.

Eckhard Bick. 2020. An annotated social media corpus for German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6127–6135, Marseille, France. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Alexandra Ciobotaru, Mihai Vlad Constantinescu, Liviu P. Dinu, and Stefan Dumitrescu. 2022. Red v2: Enhancing red dataset for multi-label emotion detection. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1392–1399, Marseille, France. European Language Resources Association.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Wesley Ramos dos Santos, Amanda MM Funabashi, and Ivandré Paraboni. 2020. Searching brazilian twitter for signs of mental health issues. In *LREC*, pages 6111–6117.

Ramón A-A. Erhardt, Reinhard Schneider, and Christian Blaschke. 2006. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 11(7):315–325.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36, Barcelona, Spain (Online). Association for Computational Linguistics.

Deepak Kumar, Nalin Kumar, and Subhankar Mishra. 2021. NLP@NISER: Classification of COVID19 tweets containing symptoms. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 102–104, Mexico City, Mexico. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Mihai Manolescu and Çağrı Çöltekin. 2021. ROFF - a Romanian Twitter dataset for offensive language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 895–900, Held Online. INCOMA Ltd.

Béatrice Mazoyer, Julia Cagé, Nicolas Hervé, and Céline Hudelot. 2020. A French corpus for event detection on Twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6220–6227, Marseille, France. European Language Resources Association.

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, Mexico City, Mexico. Association for Computational Linguistics.

Maria Mitrofan. 2017. Bootstrapping a Romanian corpus for medical named entity recognition. In *RANLP*, pages 501–509.

Maria Mitrofan. 2019. *Extragere de cunoștințe din texte în limba română și date structurate cu aplicații în domeniul medical*. Ph.D. thesis, Romanian Academy.

Maria Mitrofan, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2019. MoNERo: a biomedical gold standard corpus for the Romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79.

Maria Mitrofan and Vasile Păiș. 2022. Improving Romanian BioNER using a biologically inspired system. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 316–322, Dublin, Ireland. Association for Computational Linguistics.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vasile Păiş, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vasile Păiș. 2020. Multiple annotation pipelines inside the relate platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.

Vasile Păiș. 2022. Racai at semeval-2022 task 11: Complex named entity recognition using a lateral inhibition mechanism. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1562–1569, Seattle, United States. Association for Computational Linguistics.

Vasile Păiș, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France. European Language Resources Association.

Vasile Păiș and Maria Mitrofan. 2021. Assessing multiple word embeddings for named entity recognition of professions and occupations in health-related social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 128–130, Mexico City, Mexico. Association for Computational Linguistics.

Vasile Păiș, Maria Mitrofan, Verginica Barbu-Mititelu, Elena Irimia, Carol Luca Gasan, Roxana Micu, Laura Marin, Maria Dicusar, Bianca Florea, and Ana Badila. 2022a. Romanian micro-blogging named entity recognition (MicroBloggingNERo).

Vasile Păiș, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia, Roxana Micu, and Carol Luca Gasan. 2022b. Challenges in creating a representative corpus of Romanian micro-blogging text. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, pages 1–7, Marseille, France. European Language Resources Association.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2021. Romanian Named Entity Recognition in the Legal domain (LegalNERo).

Vasile Păiș and Dan Tufis. 2018. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, 19(2):185–191.

Vasile Păiș, Dan Tufiș, and Radu Ion. 2019. Integration of Romanian NLP tools into the RELATE platform. In *International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 181–192.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3806–3813, Istanbul, Turkey. European Language Resources Association (ELRA).

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sergio Santamaría Carrasco and Roberto Cuervo Rosillo. 2021. Word embeddings, cosine similarity and deep learning for identification of professions & occupations in health-related social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 74–76, Mexico City, Mexico. Association for Computational Linguistics.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 40–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rolly Sharma, Archana Verma, Reena Grover, Digvijay Pandey, Binay Pandey, and Lavanya A. 2020. Microbloging as a corpus for sentiment analysis structure and feeling mining. *Journal of Xi'an Shiyou University, Natural Science Edition*, pages 229–234.

Marek Suppa and Ondrej Jariabka. 2021. Benchmarking pre-trained language models for multilingual NER: TraSpaS at the BSNLP2021 shared task. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 105–114, Kiyv, Ukraine. Association for Computational Linguistics.

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Onofrei Mihaela. 2019. Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary Romanian. *Revue Roumaine de Linguistique*, 64(3):227–240.

Longyue Wang, Derek F. Wong, Lidia S. Chao, and Junwen Xing. 2012. CRFs-based Chinese word segmentation for micro-blog with small-scale data. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 51–57, Tianjin, China. Association for Computational Linguistics.

N.M. White. 2022. The hmong medical corpus: a biomedical corpus for a minority language. *Language Resources and Evaluation*.

Ayman Al Zaatari, Rim El Ballouli, Shady ELbassouni, Wassim El-Hajj, Hazem Hajj, Khaled Shaban, Nizar Habash, and Emad Yahya. 2016. Arabic corpora for credibility analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4396–4401, Portorož, Slovenia. European Language Resources Association (ELRA).