# Year 2020 (with COVID): Observation of Scientific Literature on Clinical Natural Language Processing

Natalia Grabar[1, 2] Cyril Grouin[1]
(1) Université Paris Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France
(2) STL, CNRS, Université de Lille, Domaine du Pont-de-bois, 59653 Villeneuve-d'Ascq cedex, France

natalia.grabar@univ-lille.fr, cyril.grouin@limsi.fr

**Summary**

Objectives. To analyze the content of publications within the medical NLP domain in 2020. Methods. Automatic and manual preselection of publications to be reviewed, and selection of the best NLP papers of the year. Analysis of the important issues. Results. Three best papers have been selected in 2020. We also propose an analysis of the content of the NLP publications in 2020, all topics included. Conclusion. The two main issues addressed in 2020 are related to the investigation of COVID-related questions and to the further adaptation and use of transformer models. Besides, the trends from the past years continue, such as diversification of languages processed and use of information from social networks.

**Keywords**. Natural Language Processing, Semi-automatic Selection of Publication, Topics, Issues, 2020.

## 1 Introduction

Natural Language Processing (NLP) aims at providing methods, tools and resources designed in order to mine textual and narrative documents, and to make it possible to access the information they convey [1]. While human languages are complex (as an example, learning a human language requires many years in order to be fluent), the importance of using NLP approaches to mine documents produced by humans has been pointed out since a long time [2]. In this synopsis, we first present the selection process applied this year and then we analyze the content of some publications. More particularly, we will focus on several important issues such as robustness of the methods, reproducibility of the results, as well as the originality of the research questions addressed in 2020.

## 2 The Selection Process

In order to identify all papers published during the year 2020 in the field of NLP, we queried two databases: MEDLINE[1], specifically dedicated to the biomedical domain, and the Association for Computational Linguistics (ACL) anthology[2], a database that brings together the major NLP conferences (ACL, International Conference for Computational Linguistics (COLING), Empirical Methods in Natural Language Processing (EMNLP), International Conference on Language Resources and Evaluation (LREC), Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), …etc.) and journals, since some NLP studies concerning the biomedical domain are published in conferences and journals which are not indexed by PubMed.

(English[LA] AND journal article[PT] AND 2020[DP]
AND (medical OR clinical OR natural) AND "language processing")
Figure 1: Query used for collecting candidate publications to be reviewed.

We applied the basic query we defined last year on MEDLINE (Figure 1): all journal papers published in English in 2020, having abstract, and composed of sequences "clinical language processing" or "medical language processing" or "natural language processing". As of 2020, January 9th, we collected 767 entries. We applied a similar query on the ACL anthology database and collected 10 entries. In order to process those 777 papers, we automatically scored the papers. Indeed, all the candidate papers are not specifically related to the NLP domain despite the use of one of the three sequences from the query. For instance, they can be related to other sections from the IMIA Yearbook (Public Health and Epidemiology Informatics, Decision Support, Knowledge Representation and Management, …etc.) without providing major issues for

1 https://pubmed.ncbi.nlm.nih.gov/

2 https://www.aclweb.org/anthology/

the NLP section. Hence, we applied three sets of rules we defined in 2018 while identifying best papers in a previous edition, in order to compute global scores for each publication.

The first set of rules is based upon the name of the journal (both full name and concepts found in the name):

- the positive score is assigned to the main journals in which the biomedical NLP research is usually published by the NLP community (Biomedical informatics insights, International Journal of Medical Informatics, Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, BMC Bioinformatics);
- the negative score is assigned to journals not specifically related to NLP, but to other domains such as Cognitive studies and Communication disorders (Neuroscience, Human brain mapping, Operative neurosurgery, Speech therapy, …etc.). We also dismiss survey papers and papers published in the IMIA Yearbook. We manually defined this set of journals in order to rule out those false positives.

The second set of rules relies on concepts found in both the title and abstract of papers:

- the positive score is assigned to concepts typically involved in papers related to NLP. Those concepts may be related to objectives, resources, and tools (such as *natural language processing, NLP, named entity recognition, NER, part of speech, POS, tagged words, semantic, syntax, biomedical entity, meanings, electronic health record, EHR, reports, notes, clinical text, text corpus, free text, unstructured text, tweets, PubMed, Social Media, MedDRA, UMLS, annotated data, Metamap*);
- the negative score is assigned to concepts that are usually involved in studies related to disorders involving anatomical parts or language abilities (such as *word processing, language production, language comprehension, voice quality, left posterior superior temporal gyrus, pSTG, posterior superior temporal sulcus, pSTS, inferior fronto-occipital fasciculus, IFOF, dorsolateral prefrontal cortex, cortex, language lateralization, chemical fragment, fragment chemistry, brain structures, verbal intelligence, cerebral, positive mismatch responses, pMMRs, prelingual, postlingual, cochlear, aphasia, SAPS, cortical, language function, infants*).

The third set of rules is also applied on titles and abstracts, and covers the concepts describing the methodology used in papers:

- the positive score is assigned to papers using classical NLP methods or evaluation metrics (such as *annotation tool, text-mining, rule-based, regular expression, lexicon, CRF, recall, precision, F1-score, F-measure, accuracy, Inter-annotator agreement, Kappa, classify/classifier, detect, extract, extraction, predict, predicting, text simplification, lexical simplification*);
- the negative score is assigned to papers claiming to use the NLP methods, such as pointed out by sequences like *using natural language processing, using NLP, perform a Natural Language Processing analysis.* Such papers are downgraded because the NLP claims are usually limited to the use of existing and ready-to-use NLP tools while the main contribution of papers is related to the analysis of tool results rather than to the improvements made to NLP methods by researchers who take advantage of existing tools.

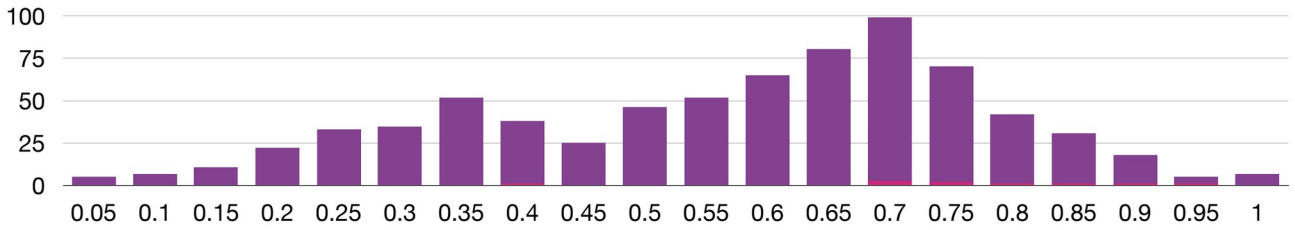


Figure 2: Distribution of papers according to the filter scores (violet bars indicate the total number of papers, pink bars indicate papers kept in the top-15 best papers).

For each of the 777 candidate papers, the final score ranked from 0.05 to 1 (Figure 2). On this figure, the violet bars indicate the total number of papers for each computed grade, while the pink bars indicate the papers we kept in the top-15 best papers' list. This score has been used as a meta element during the

manual selection of the top-15 papers. Indeed, the section editors did not fully rely on the scores but only used them as additional information. Hence, both section editors independently browsed the abstracts, keywords and automatic scores, and assigned a *Yes / Maybe / No* score to each paper. All papers having at least one *Yes* or *Maybe* score have been kept for the next step of the selection. At this stage, 143 candidate papers remain (i.e., a subset of 18.4% of the whole dataset). We then performed an adjudication process, in order to choose the final 15 candidates to be proofread by external reviewers. We payed attention to the topics addressed by the researchers and to their geographic origin so as to provide enough diversity. As a result, out of the fifteen papers, four come from the USA, three from China, and one from each of the following countries: Finland, Germany, Italy, Peru, Spain, Turkey, and United-Kingdom. This is the first time we select a paper from South America in our top-15 best paper candidates.

In the next sections, we present the main issues and approaches addressed in the preselected publications.

## 3 Current Trends in Biomedical NLP

Since a few years, we observed the increasing use of transformer models based on word embeddings. Those models are useful to capture the context of words. When they achieve to cover adequately the domain (e.g., clinical domain) or the properties of a given corpus (e.g., clinical texts), they allow to obtain high performances w.r.t. approaches that do not use such models. Several transformer models have been released during the last years: currently, the most famous one is BERT, a multilingual generic model provided by Google [3].

Based on this model, several other models were produced: (i) either specific to a domain such as BioBERT [4] for the biomedical domain; or (ii) specific to a language, such as BERTje [5] for Dutch, FlauBERT [6] and CamemBERT [7] for French, GottBERT [8] for German, AlBERTo [9] and UmBERTo[3] for Italian, or BETO [10] for Spanish; or (iii) specific to a given task such as MRCBert [11] for summarization or SQuADBERT[4] for question-answering, trained on the Stanford Question-Answering Dataset [12]. Conversely, almost all previous methods still disappeared, due to their relative high lifespan: Word2vec [13], GloVe [14], and ELMo [15]. New approaches are coming, with an increasing number of parameters in those new models, especially the Generative Pretrained Transformer (GPT) series provided by OpenAI: GPT-2[5] [16] for text generation, and GPT-3 [17] for all tasks of NLP (using up to 175 billion parameters).

As a consequence, there is a harmful race for being the first to produce such resources, which implies to rapidly publish a paper in the first available conference or workshop, or on the arXiv deposit, in order to be cited. Nevertheless, this also implies that such papers would certainly never be identified as best paper candidates for the NLP chapter of the IMIA Yearbook, unless they are indexed in PubMed, or published in an NLP conference that we, co-editors of the NLP section, used to visit.

Nevertheless, an opposite way has also been observed, with authors searching for a green research that does not use models which are not environmentally friendly (expensive in terms of hardware, running time, and $CO_2$ footprint). This way has been investigated by Poerner *et al.* [18] which proposed a GreenBioBERT model that has been produced using Word2vec to train a model on a new target domain (namely, on the COVID-19 issue) along with an alignment of vectors from the existing BioBERT model and the model trained with Word2vec.

In terms of architecture, the most commonly used for NLP tasks currently is a bidirectional LSTM, which permits to capture both left and right contexts, plus a final Conditional Random Field (CRF) layer to refine the outputs (BiLSTM-CRF). Those methods achieve excellent results but need high computation resources to train new models, currently only available in huge data centers, which implies an important $CO_2$ footprint. The main issue is how to use such models on new domains, new languages, or for a new task? Instead of training new models, alternative solutions exist, such as fine-tuning of existing models, domain adaptation, transfer learning, as done by Jin *et al.* [19] to predict clinical trial results. Nevertheless, such models remain expensive.

### 3.1 Languages Addressed

3 https://github.com/musixmatchresearch/umberto

4 https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

5 https://openai.com/blog/better-language-models/

In relation with the languages processed, work with texts written in English represents the largest part of publications. Indeed, there is a significant number of existing corpora, datasets and resources available in English. Yet, we observe an increasing number of publications dedicated to other languages and a greater variety of languages: Arabic [20], Chinese [21-26], Croatian [27], Finnish [28,29], French [30,31], German [32-34], Hebrew [35], Italian [36-38], Japanese [39,40], Korean [41,42], Norwegian [43], Portuguese [44], Spanish [45-48], Swedish [49], and Turkish [28]. Overall, we believe that the trend observed in previous years is continuing. We expect it will develop further in the future.

### 3.2 COVID-19

A lot of work has been done on the COVID-19 issue. In order to help such work, several papers present COVID-19 resources, such as the COVID-19 Open Research Dataset (CORD-19) produced by Wang *et al.* [50] and available through the Kaggle[6] platform. Besides, several works are anchored in hospitals and propose to predict conditions and events related to COVID, such as prediction of admission of patients with COVID in an intensive care unit [51], detection of COVID cases from radiological text reports [46], prognosis prediction for patients with chronic obstructive pulmonary disease [45] or for patients with hypertension [31].

In addition to clinical and hospital information, researchers investigate scientific literature looking for instance for drug repurposing recommendations [52], and for temporal evolution of research work on COVID-19 [53]. Besides, the first systematic reviews related to COVID are proposed [54], including the focus on research needs [55].

The researches also investigate social networks, focusing on analyzing public opinion and emotions on the COVID pandemics in Twitter posts [56,57], monitoring illicit sales of COVID medication on Twitter and Instagram [58], and observing COVID symptoms and disease histories collected from a large population in Reddit, which may provide more reliable insights [59]. Another important issue is that, in the current situation, the surveillance of emerging epidemiological events becomes again very important, as around 60% of all outbreaks are detected using informal sources, which motivated online epidemiological surveillance [32].

### 3.3 Neurological and Psychiatric Disorders

We observed an interest for neurological and psychiatric disorders, which are mainly issued from clinical context: detection of duration of untreated psychosis [38], analysis of language in patients with aphasia [60], Alzheimer's disease [61], and autism spectrum disorder [62], generation of artificial mental health records and their evaluation [63], detection and prediction of suicide in mental illness [64-66], automatic detection of agitation and related symptoms among hospitalized patients [67], analysis of COVID impact on people with epilepsy [36], prediction of care cost in mental health setting [68], and, more generally, the use of artificial intelligence in mental health and its biases [69].

### 3.4 Place of Patients

The place of patients in the healthcare context is increasing and several publications place patients at the center of their investigations. Hence, in addition to the patient-centered healthcare process, we can also mention work focusing on patient outcome [70], studying public opinion on use of free-text data in electronic medical records for research [71], studying patient feedback on the quality of care [72], analyzing the ways the patients describe their pain [73], measuring the quality of patient-doctor communication [74], analyzing the developmental crisis episodes that occur during early adulthood in social media [75], and analyzing patient experience in order to define some guidelines [76].

### 3.5 Social Networks

As noticed above, the social networks continue to provide important information on several issues: public opinion and emotions on the COVID pandemics in Twitter posts [56,57], surveillance of illicit sales of COVID medication on Twitter and Instagram [58], observation of COVID symptoms and disease histories collected from a large population in Reddit [59], surveillance of emerging epidemiological events [32], analysis of HIV-related tweets and of their relation to the HIV incidence [77], analysis of drug use on Twitter [78], analysis of developmental crisis episodes during early adulthood in social media [75].

---

6 https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

We assume that the investigation of social networks will go further in the future, as these networks provide independent and massive information on various events.

## 4 Conclusion

The NLP publications in 2020 have been heavily marked by the sanitary situation, which motivated an increasing number of works related to the COVID-19 pandemic. The authors addressed all types of content (clinical texts, clinical trials, scientific papers, social media, etc.) so as to mine information related to this major issue (adverse drug reactions, usefulness of existing treatments, psychological impact of the pandemic, …etc.).

From a scientific perspective, we observe an increasing use of transformer models based on word embeddings. In continuation of the trend already observed in previous years, we notice that the variety of languages processed is also increasing. This observation is also related to the use of multilingual transformer models, among them BERT is the most used since it allows to process more than one hundred languages. In addition, several authors adapted those multilingual models to their data (specific domain for a given language), which also increases the number of publications related to new languages. An opposite way also appeared with papers focusing on green research, especially to propose methods for processing new domains or new data, using existing transformer models without any new training steps for these complex models.

In the coming years, we hope that environmentally friendly solutions will be preferred to the production of new transformer models which still need more and more computing resources. In addition, we also urge the NLP community to go back to a qualitative analysis of their outputs, rather than a basic and harmful race to present numerical gains over other similar studies. The positive impact of NLP research on clinical issues should be highlighted rather than the search for better results computed by evaluation metrics.

## References


1. Nadkarni P.M., Ohno-Machado L., Chapman W.W. (2011). Natural Language Processing: an introduction. *J Am Med Inform Assoc*, **18**, 544–51.
2. Friedman C. and Hripcsak G. (1999). Natural language processing and its future in medicine. *Academic Medicine*, **74**(8), 890–5.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810. 04805v2.
4. Lee J., Yoon W., Kim S., Kim D., Kim S., So C. H., Kang J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, p. 1–7.
5. De Vries W., Van Cranenburgh A., Bisazza A., Caselli T., Van Noord G., Nissim M. (2019). BERTje: A Dutch BERT Model. https://arxiv.org/abs/1912.09582v1.
6. Le H., Vial L., Frej J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proc. of LREC*, p. 2479–2490, Marseille, France: European Language Resources Association.
7. Martin L., MULLER B., Ortiz SUÁREZ P. J., DUPONT Y., Romary L., De La CLERGERIE É., SEDDAH D. & SagotT B. (2020). CamemBERT: a tasty French language model. In *Proce. of ACL*, p. 7203–7219: Association for Computational Linguistics.
8. SCHEIBLE R., THOMCZYK F., TIPPMANN P., JARAVINE V. & BOEKER M. (2020). GottBERT: a pure German Language Model. *CoRR*, abs/2012.02110.
9. POLIGNANO M., Basile P., DE Gemmis M., SEMERARO G. & Basile V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proc. of CLiC-it*, volume 2481: CEUR.
10. CAÑETE J., CHAPERON G., FUENTES R., Ho J.-H., Kang H. & PÉREZ J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR*.

11. Jain S., Tang G. & Chi L. S. (2021). MRCBert: A Machine Reading Comprehension Approach for Unsupervised Summarization. https://arxiv.org/abs/2105.00239v1.

12. RAJPURKAR P., Zhang J., LOPYREV K. & Liang P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392, Austin, Texas: Association for Computational Linguistics.

13. MIKOLOV T., Chen K., Corrado G. & Dean J. (2013). Efficient Estimation of Word Represen- tations in Vector Space. https://arxiv.org/abs/1301.3781v3.

14. PENNINGTON J., SOCHER R. & MANNING C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

15. PETERS M. E., NEUMANN M., IYYER M., GARDNER M., Clark C., Lee K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

16. RADFORD A., Wu J., Child R., Luan D., Amodei D. & SUTSKEVER I. (2019). Language Mod- els are Unsupervised Multitask Learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

17. BROWN T. B., Mann B., Ryder N., Subbiah M., Kaplan J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-Voss A., KRUEGER G., HENIGHAN T., Child R., RAMESH A., ZIEGLER D. M., Wu J., Winter C., Hesse C., Chen M., SIGLER E., LITWIN M., Gray S., Chess B., Clark J., Berner C., McCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165v4.

18. POERNER N., WALTINGER U. & SCHÜTZE H. (2020). Inexpensive Domain Adaptation of Pretrained Language Models: Case Studies on Biomedical NER and Covid-19 QA. In *Proc. of EMNLP*, p. 1482–90.

19. Jin Q., Tan C., Chen M., Liu X. & Huang S. (2020). Predicting Clinical Trial Results by Implicit Evidence Integration. In *Proc. of EMNLP*, p. 1461–77.

20. Faris H., Habib M., Faris M., Alomari M. & Alomari A. (2020). Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines. *J Biomed Inform*, **109**, 103525–5.

21. Chen C.-H., Hsieh J.-G., Cheng S.-L., Lin Y.-L., Lin P.-H. & Jeng J.-H. (2020). Early short-term prediction of emergency department length of stay using natural language processing for low-acuity outpatients. *Am J Emerg Med*, **38**(11), 2368–2373.

22. Li X., Lin X., Ren H. & Guo J. (2020). Ontological organization and bioinformatic analysis of adverse drug reactions from package inserts: Development and usability study. *J Med Internet Res*, **22**(7), 20443–3.

23. Wang Z., Huang H., Cui L., Chen J., An J., Duan H., Ge H. & Deng N. (2020). Using natural language processing techniques to provide personalized educational materials for chronic disease patients in China: Development and assessment of a knowledge-based health recommender system. *JMIR Med Inform*, **8**(4), 17642–2.

24. Wu C., Luo G., Guo C., Ren Y., Zheng A. & Yang C. (2020). An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions. *J Biomed Inform*, **108**, 103511–1.

25. Xia H., An W., LI J. & Zhang Z. J. (2020). Outlier knowledge management for extreme public health events: Understanding public opinions about COVID-19 based on microblog data. *Socioecon Plann Sci*, **10**(09), 41–1.

26. Zhang Z., Zhu L. & Yu P. (2020). Multi-level representation learning for Chinese medical entity recognition: Model development and validation. *JMIR Med Inform*, **8**(5), 17637–7.

27. KRSNIK I., GLAVAŠ G., KRSNIK M., MILETIC´ D. & ŠTAJDUHAR I. (2020). Automatic annotation of narrative radiology reports. *Diagnostics (Basel)*, **10**(4), 196–6.

28. GÜNGÖR O., GÜNGÖR T. & USKUDARLI S. (2020). Exseqreg: Explaining sequence-based nlp tasks with regions with a case study using morphological features for named entity recognition. *PLoS One*, **15**(12), 244179–9.

29. Moen H., Hakala K., PELTONEN L.-M., MATINOLLI H.-M., SUHONEN H., Terho K., DANIELSSON-OJALA R., Valta M., Ginter F., SALAKOSKI T. & SALANTERÄ S. (2020). Assisting nurses in care documentation: from automated sentence classification to coherent document structures with subject headings. *J Biomed Semantics*, **11**(1), 10–30.

30. Grabar N., Dalloux C. & Claveau V. (2020). CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, **11**(1), 1–7.

31. NEURAZ A., LERNER I., Digan W., Paris N., TSOPRA R., ROGIER A., BAUDOIN D., Cohen K. B., Burgun A., GARCELON N. & RANCE B. (2020). Natural language processing for rapid response to emergent diseases: Case study of calcium channel blockers and hypertension in the COVID-19 pandemic. *J Med Internet Res*, **22**(8), 20773–3.

32. Abbood A., ULLRICH A., BUSCHE R. & GHOZZI S. (2020). EventEpi-A natural language processing framework for event-based surveillance. *PLoS Comput Biol*, **16**(11), 1008277–7.

33. FERRARIO A., DEMIRAY B., YORDANOVA K., Luo M. & Martin M. (2020). Social rem- iniscence in older adults' everyday conversations: Automated detection using natural language processing and machine learning. *J Med Internet Res*, **22**(9), 19133–3.

34. WULFF A., Mast M., HASSLER M., MONTAG S., MARSCHOLLEK M. & Jack T. (2020). Designing an openEHR-based pipeline for extracting and standardizing unstructured clinical data using natural language processing. *Methods Inf Med*, **59**(S02), 64–78.

35. Barash Y., GURALNIK G., Tau N., SOFFER S., Levy T., SHIMON O., ZIMLICHMAN E., Kohen E. & Klang E. (2020). Comparison of deep learning models for natural language processing-based classification of non-english head ct reports. *Neuroradiology*, **62**(10), 1247–1256.

36. LANZONE J., Cenci C., TOMBINI M., Ricci L., Tufo T., PICCIOLI M., MARRELLI A., MECARELLI O. & ASSENZA G. (2020). Glimpsing the impact of COVID19 lock-down on people with epilepsy: A text mining approach. *Front Neurol*, **11**, 870–0.

37. Mensa E., Colla D., DALMASSO M., GIUSTINI M., Mamo C., PITIDIS A. & RADICIONI D. P. (2020). Violence detection explanation via semantic roles embeddings. *BMC Med Inform Decis Mak*, **20**(1), 263–3.

38. Viani N., Kam J., Yin L., Bittar A., Dutta R., Patel R., STEWART R. & VELUPILLAI S. (2020). Temporal information extraction from mental health records to identify duration of untreated psychosis. *J Biomed Semantics*, **11**(1), 2–22.

39. NAKATANI H., NAKAO M., UCHIYAMA H., TOYOSHIBA H. & OCHIAI C. (2020). Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: Case-control study. *JMIR Med Inform*, **8**(4), 16970.

40. UJIIE S., Yada S., WAKAMIYA S. & ARAMAKI E. (2020). Identification of adverse drug event-related Japanese articles: Natural language processing analysis. *JMIR Med Inform*, **8**(11), 22661–1.

41. Cho I., Lee M. & Kim Y. (2020). What are the main patient safety concerns of healthcare stakeholders: a mixed-method study of web-based text. *Int J Med Inform*, **140**, 104162–2.

42. Lee K. H., Kim H. J., Kim Y. J., Kim J. H. & Song E. Y. (2020). Extracting structured genotype information from free-text hla reports using a rule-based approach. *J Korean Med Sci*, **35**(12), 78–8.

43. ESKILDSEN N. K., ERIKSSON R., CHRISTENSEN S. B., AGHASSIPOUR T. S., BYGSØ M. J., BRUNAK S. & HANSEN S. L. (2020). Implementation and comparison of two text mining methods

with a standard pharmacovigilance method for signal detection of medication errors. *BMC Med Inform Decis Mak*, **20**(1), 4–4.

44. Lopes F., TEIXEIRA C. & OLIVEIRA H. G. (2020). Comparing different methods for named entity recognition in Portuguese neurology text. *J Med Syst*, **44**(4), 77–90.

45. GRAZIANI D., SORIANO J. B., Rio-BERMUDEZ C. D., MORENA D., DÍAZ T., CASTILLO M., ALONSO M., ANCOCHEA J., LUMBRERAS S. & IZQUIERDO J. L. (2020). Characteristics and prognosis of COVID-19 in patients with COPD. *J Clin Med*, **9**(10), 3259–9.

46. LÓPEZ-ÚBEDA P., Diaz-GALIANO M. C., MARTÍN-NOGUEROL T., Luna A., UREÑA-LÓPEZ L. A. & MARTÍN-VALDIVIA M. T. (2020). COVID-19 detection in radiological text reports integrating entity recognition. *Comput Biol Med*, **127**, 104066–6.

47. NAJAFABADIPOUR M., Zanin M., RODRÍGUEZ-GONZÁLEZ A., TORRENTE M., GARCÍA B. N., BERMUDEZ J. L. C., PROVENCIO M. & MENASALVAS E. (2020). Reconstructing the patient's natural history from electronic health records. *Artif Intell Med*, **105**, 101860–0.

48. SANTISO S., PÉREZ A., CASILLAS A. & ORONOZ M. (2020). Neural negated entity recognition in Spanish electronic health records. *J Biomed Inform*, **105**, 1–15.

49. Caccamisi A., JØRGENSEN L., DALIANIS H. & ROSENLUND M. (2020). Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. *Ups J Med Sci*, **125**(4), 316–324.

50. Wang L. L., Lo K., Chaddrasekhar Y., Reas R., Yang J., Burdick D., Eide D., Funk K., Katsis Y., Kinney R., Li Y., Liu Z., Merrill W., Mooney P., Murdick D., Rishi D., Sheehan J., Shen Z., Stilson B., Wade A., Wang K., Wang N. X. R., Wilhelm C., Xie B., Raymond D., Weld D. S., Etzioni O. & Kohlmeier S. (2020). CORD-19: The COVID-19 Open Research Dataset. https://arxiv.org/abs/2004.10706v4.

51. Izquierdo J. L., Ancochea J., Savana COVID-19 RESEARCH GROUP & Soriano J. B. (2020). Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: Retrospective study using machine learning and natural language processing. *J Med Internet Res*, **22**(10), 21801–1.

52. Gates L. E. & Hamed A. A. (2020). The anatomy of the SARS-CoV-2 biomedical literature: Introducing the CovidX network algorithm for drug repurposing recommendation. *J Med Internet Res*, **22**(8), 21169–9.

53. Ebadi A., Xi P., Tremblay S., Spencer B., Pall R. & Wong A. (2020). Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. *Scientometrics*, **11**, 1–15.

54. Wang L. L. & Lo K. (2020). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief Bioinform*, **22**(2), 781–799.

55. Doanvo A., Qian X., Ramjee D., Piontkivska H., Desai A. & Majumder M. (2020). Machine learning maps research needs in COVID-19 literature. *Patterns (N Y)*, **1**(9), 100123–3.

56. Boon-Itt S. & Skunkan Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill*, **6**(4), 21978–8.

57. Dyer J. & Kolic B. (2020). Public risk perception and emotion on Twitter during the Covid-19 pandemic. *Appl Netw Sci*, **5**(1), 99–9.

58. Mackey T. K., Li J., Purushothaman V., Nali M., Shah N., Bardier C., Cai M. & Luang B. (2020). Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram. *JMIR Public Health Surveill*, **6**(3), 20794–4.

59. Picone M., Inoue S., Defelice C., Naujokas M. F., Sinrod J., Cruz V. A., Stapleton J., Sinrod E., Diebel S. E. & Wassman E. R. (2020). Social listening as a rapid approach to collecting and analyzing COVID-19 symptoms and disease natural histories reported by large numbers of individuals. *Popul Health Manag*, **23**(5), 350–360.

60. Themistocleous C., Webster K., Afthinos A. & Tsapkini K. (2020). Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *Am J Speech Lang Pathol*, **30**(15), 466–480.

61. Reeves S., Williams V., Costela F. M., Palumbo R., Umoren O., Christopher M. M., Blacker D. & Woods R. L. (2020). Narrative video scene description task discriminates between levels of cognitive impairment in Alzheimer's disease. *Neuropsychology*, **34**(4), 437–446.

62. Chojnicka I. & Wawer A. (2020). Social language in autism spectrum disorder: A computational analysis of sentiment and linguistic abstraction. *PLoS One*, **15**(3), 229985–5.

63. Ive J., Viani N., Kam J., Yin L., Verma S., Puntis S., Cardinal R. N., Roberts A., Srewart R. & Velupillai S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digit Med*, **3**, 69–9.

64. Senior M., Burghart M., Yu R., Kormilitzin A., Liu Q., Vaci N., Nevado-Holgado A., Pandit S., Zlodre J. & Fazel S. (2020). Identifying predictors of suicide in severe mental illness: A feasibility study of a clinical prediction rule (oxford mental illness and suicide tool or OxMIS). *Front Psychiatry*, **11**, 268–8.

65. Levis M., Westgate C. L., Gui J., Watts B. V. & Shiner B. (2020). Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychol Med*, **2**, 1–10.

66. Jayasinghe L., Bittar A., Dutta R. & Stewart R. (2020). Clinician-recalled quoted speech in electronic health records and risk of suicide attempt: a case-crossover study. *BMJ Open*, **10**(4), 36186–6.

67. Hart K. L., Pellegrini A. M., Forester B. P., Berretta S., Murphy S. N., Perlis R. H. & McCoy T. H. (2020). Distribution of agitation and related symptoms among hospitalized patients using a scalable natural language processing method. *Gen Hosp Psychiatry*, **68**, 46–51.

68. Colling C., Khondoker M., Patel R., Fok M., Harland R., Broadbent M., McCrone P. & Stewart R. (2020). Predicting high-cost care in a mental health setting. *BJPsych Open*, **6**(1), 10–0.

69. Straw I. & Callison-Burch C. (2020). Artificial intelligence in mental health and the biases of language based models. *PLoS One*, **15**(2), 240376–6.

70. Hernandez-Boussard T., Blayney D. W. & Brooks J. D. (2020). Leveraging digital data to inform and improve quality cancer care. *Cancer Epidemiol Biomarkers Prev*, **29**(4), 816–822.

71. Ford E., Oswald M., Hassan L., Ad Goran Nenadic K. B. & Cassell J. (2020). Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. *J Med Ethics*, **46**(6), 367–377.

72. Nawab K., Ramsey G. & Schreiber R. (2020). Natural language processing to extract meaningful information from patient experience feedback. *Appl Clin Inform*, **11**(2), 242–252.

73. Tighe P. J., Sannapaneni B., Fillingim R. B., Doyle C., Kent M., Shickel B. & Rashidi P. (2020). Forty-two million ways to describe pain: Topic modeling of 200,000 PubMed pain-related abstracts using natural language processing and deep learning-based text generation. *Pain Med*, **21**(11), 3133–3160.

74. Cuffy C., Hagiwara N., Vrana S. & McInnes B. T. (2020). Measuring the quality of patient-physician communication. *J Biomed Inform*, **112**, 103589–9.

75. Agarwal S., Guntuku S. C., Robinson O. C., Dunn A. & Ungar L. H. (2020). Examining the phenomenon of quarter-life crisis through artificial intelligence and the language of twitter. *Front Psychol*, **11**, 241–1.

76. Cammel S. A., De Vos M. S., Van Soest D., Hettne K. M., Boer F., Steyerberg E. W. & Boosman H. (2020). How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Med Inform Decis Mak*, **20**(1), 97–7.

77. Stevens R., Bonett S., Bannon J., Chittamuru D., Slaff B., Browne S. K., Huang S. & Bauermeister J. A. (2020). Association between HIV-related tweets and HIV incidence in the United States: Infodemiology study. *J Med Internet Res*, **22**(6), 17196–6.

78. Tassone J., Yan P., Simpson M., Mendhe C., Mago V. & Choudhury S. (2020). Utilizing deep learning and graph mining to identify drug use on Twitter data. *BMC Med Inform Decis Mak*, **20**(11), 304–4.

## Content Summaries of Best Papers for the Natural Language Processing Section of the 2021 IMIA Yearbook

**Qiao Jin, Chuanqi Tan, Mosha Chen, Xiaozhong Liu, Songfang Huang (2020). Predicting Clinical Trial Results by Implicit Evidence Integration. In: Proc of Empirical Methods in NLP. doi: 10.18653/v1/2020.emnlp-main.114**

The clinical trial result prediction (CTRP) task is based on medical literature containing PICO (how the Intervention group compares with the Comparison group in terms of the measured Outcomes in the studied Population). The authors proposed an EBM-Net model which is a transformer model that uses unstructured sentences as implicit evidences and a fine-tuning approach. They compared their fine-tuned model w.r.t. the BioBERT model and other approaches (MeSH ontology, bag-of-words, …etc.) and achieved better results on the COVID-19 clinical trials' dataset (22 clinical trials from the CORD-19 dataset).

**Nina Poerner, Ulli Waltinger, Hinrich Schütze (2020). Inexpensive Domain Adaptation of Pre- trained Language Models: Case Studies on Biomedical NER and Covid-19 QA. In: Proc of Empirical Methods in NLP. doi: 10.18653/v1/2020.findings-emnlp.134**

The authors highlight the expensive cost of domain adaptation while training a model on target-domain text. This cost is expressed in terms of hardware requirement, high running time and negative impact on the $CO_2$ footprint. The authors investigated a solution they called GreenBioBERT that relies first on a Word2vec training stage on target-domain texts (PubMed, PMC, CORD-19), and second on the alignment of word vectors with the vectors from BioBERT. They applied their model on two issues: 8 biomedical NER tasks in English, and question-answering (QA) on COVID-19 issue. The authors achieved competitive results using BioBERT and a better precision on a few tasks; on the COVID-19 QA task, their model achieved better results than the SQuADBERT model (designed for QA). In this paper, the authors proposed a useful method to use existing pretrained language models in order to adapt them to new datasets, new tasks, new languages, …etc.

**Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf Car- dinal, Angus Roberts, Robert Stewart, Sumithra Velupillai (2020). Generation and evaluation of artificial mental health records for Natural Language Processing. In: NPJ Digital Medicine. doi: 10.1038/s41746-020-0267-x**

The main problem for biomedical NLP is the difficult access to clinical documents and the inherent complexity to completely de-identify documents. The solution proposed by the authors consists in generating artificial discharge summaries. In this paper, the authors produced artificial summaries in mental health, based on the MIMIC-III data. Then, they used their artificial texts in order to train models using the Keras toolkit for classification tasks. The authors observed that models trained on their synthetic data perform as well as models trained on real data. Since the synthetic discharge summaries have been produced taking as input the MIMIC-III data, the authors cannot share their resources. Nevertheless, the method is reproducible.