Spark NLP: natural language understanding at scale

Kocaman, V.; Talby, D.

Software Impacts 8 (2021) 100058

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

Spark NLP: Natural Language Understanding at Scale

Veysel Kocaman *, David Talby

John Snow Labs Inc., 16192 Coastal Highway, Lewes, DE, 19958, USA

ARTICLE INFO

Keywords:
Spark
Natural language processing
Deep learning
Tensorflow
Cluster

ABSTRACT

Spark NLP is a Natural Language Processing (NLP) library built on top of Apache Spark ML. It provides simple, performant & accurate NLP annotations for machine learning pipelines that can scale easily in a distributed environment. Spark NLP comes with 1100+ pretrained pipelines and models in more than 192+ languages. It supports nearly all the NLP tasks and modules that can be used seamlessly in a cluster. Downloaded more than 2.7 million times and experiencing 9x growth since January 2020, Spark NLP is used by 54% of healthcare organizations as the world's most widely used NLP library in the enterprise.

Code metadata

Current code version	v2.7.1
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2021-5
Permanent link to Reproducible Capsule	https://codeocean.com/capsule/1573505/tree/v1
Legal Code License	Apache-2.0 License
Code versioning system used	git, maven
Software code languages, tools, and services used	scala, python, java, R
Compilation requirements, operating environments & dependencies	jdk 8, spark
If available Link to developer documentation/manual	https://nlp.johnsnowlabs.com/api/
Support email for questions	info@johnsnowlabs.com

Software metadata

Current software version	2.7.1
Permanent link to executables of this version	https://github.com/JohnSnowLabs/spark-nl
Permanent link to Reproducible Capsule	https://codeocean.com/capsule/1573505/tree/v1
Legal Software License	Apache-2.0 License
Computing platforms/Operating Systems	Linux, Ubuntu, OSX, Microsoft Windows, Unix-like
Installation requirements & dependencies	jdk 8, spark
If available, link to user manual - if formally published include a reference to the publication in the reference list	https://nlp.johnsnowlabs.com/api/
Support email for questions	info@johnsnowlabs.com

1. Spark NLP library

Natural language processing (NLP) is a key component in many data science systems that must understand or reason about a text. Common use cases include question answering, paraphrasing or summarizing, sentiment analysis, natural language BI, language modeling, and disambiguation. Nevertheless, NLP is always just a part of a bigger data processing pipeline and due to the nontrivial steps involved in this process, there is a growing need for all-in-one solution to ease the burden of text preprocessing at large scale and connecting the dots between various steps of solving a data science problem with NLP. A good NLP library should be able to correctly transform the free text into structured features and let the users train their own NLP models that are easily fed into the downstream machine learning (ML) or deep learning (DL) pipelines with no hassle.

Spark NLP is developed to be a single unified solution for all the NLP tasks and is the only library that can scale up for training and

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: veysel@johnsnowlabs.com (V. Kocaman), david@johnsnowlabs.com (D. Talby).

https://doi.org/10.1016/j.simpa.2021.100058
Received 21 January 2021; Received in revised form 22 January 2021; Accepted 24 January 2021

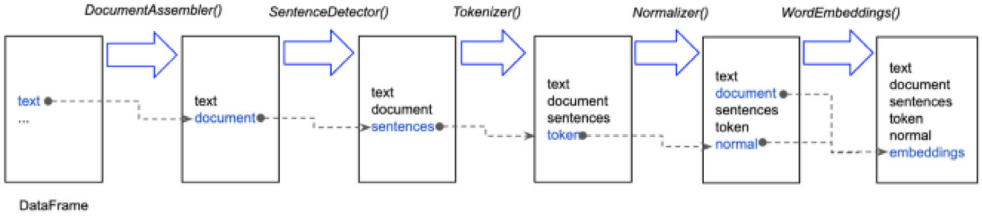

Fig. 1. The flow diagram of a Spark NLP pipeline. When we fit() on the pipeline with a Spark data frame, its text column is fed into the DocumentAssembler() transformer and a new column *document* is created as an initial entry point to Spark NLP for any Spark data frame. Then, its document column is fed into the SentenceDetector() module to split the text into an array of sentences and a new column "sentences" is created. Then, the "sentences" column is fed into Tokenizer(), each sentence is tokenized, and a new column "token" is created. Then, Tokens are normalized (basic text cleaning) and word embeddings are generated for each. Now data is ready to be fed into NER models and then to the assertion model.

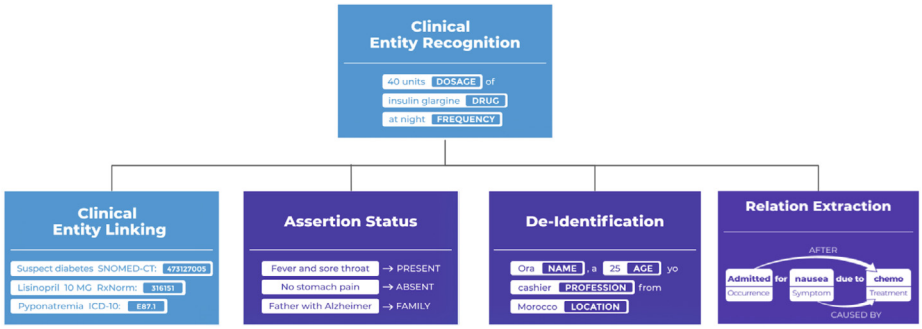

Fig. 2. Named Entity Recognition is a fundamental building block of medical text mining pipelines, and feeds downstream tasks such as assertion status, entity linking, de-identification, and relation extraction.

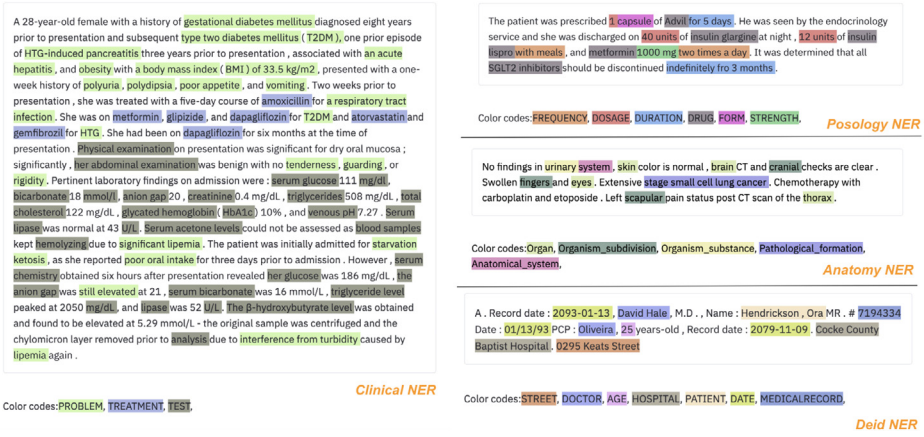

Fig. 3. Sample clinical entities predicted by a clinical NER model trained on various datasets. There are more than 40 pretrained NER models in Spark NLP Enterprise edition.

inference in any Spark cluster, take advantage of transfer learning and implementing the latest and greatest algorithms and models in NLP research, and deliver a mission-critical, enterprise-grade solutions at the same time. It is an open-source natural language processing library, built on top of Apache Spark and Spark ML. It provides an easy API to integrate with ML pipelines and it is commercially supported by John Snow Labs Inc, an award-winning healthcare AI and NLP company based in USA.

Spark NLP's annotators utilize rule-based algorithms, machine learning and deep learning models which are implemented using TensorFlow that has been heavily optimized for accuracy, speed, scalability, and memory utilization. This setup has been tightly integrated with Apache Spark to let the driver node run the entire training using all the available cores on the driver node. There is a CuDA version of each TensorFlow component to enable training models on GPU when available. The Spark NLP is written in Scala and provides open-source API's in Python, Java, Scala, and R - so that users do not need to be aware of the underlying implementation details (TensorFlow, Spark, etc.) in order to use it. Since it has an active release cycle (released 26 new versions in 2019 and another 26 in 2020), the latest trends and research in NLP field are embraced and implemented rapidly in a way that could scale well in a cluster setting to allow common NLP pipelines run orders of magnitude faster than what the inherent design limitations of legacy libraries allowed.

Spark NLP library has two versions: Open source and enterprise. Open source version has all the features and components that could be expected from any NLP library, using the latest DL frameworks and research trends. Enterprise library is licensed (free for academic purposes) and designed towards solving real world problems in healthcare domain and extends the open source version. The licensed version has the following modules to help researchers and data practitioners in various means: Named entity recognition (NER), assertion status (negativity scope) detection, relation extraction, entity resolution (SNOMED, RxNorm, ICD10 etc.), clinical spell checking, contextual parser, text2SQL, deidentification and obfuscation. High level overview of the components from each version can be seen at Fig. 4.

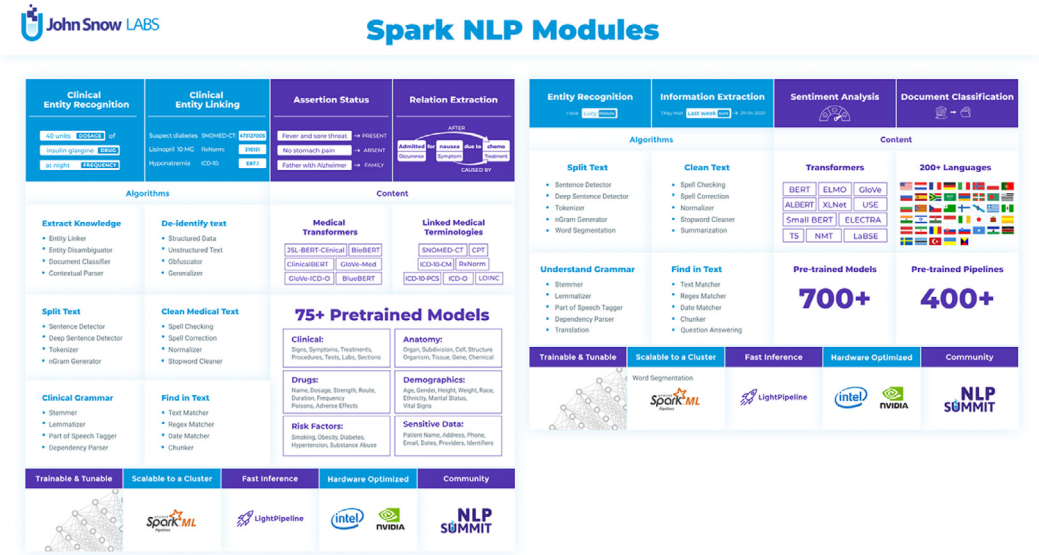

Fig. 4. Spark NLP library has two versions (open source and enterprise) and each comes with a set of pretrained models and pipelines that could be used out of the box with no further training or dataset.

2. The impact to research fields

The COVID-19 pandemic brought a surge of academic research about the virus — resulting in 23,634 new publications between January and June of 2020 [1] and accelerating to 8,800 additions per week from June to November on the COVID-19 Open Research Dataset [2]. Such a high volume of publications makes it impossible for researchers to read each publication, resulting in increased interest in applying natural language processing (NLP) and text mining techniques to enable semi-automated literature review [3].

In parallel, there is a growing need for automated text mining of Electronic health records (EHRs) in order to find clinical indications that new research points to. EHRs are the primary source of information for clinicians tracking the care of their patients. Information fed into these systems may be found in structured fields for which values are inputted electronically (e.g. laboratory test orders or results) [4] but most of the time information in these records is unstructured making it largely inaccessible for statistical analysis [5]. These records include information such as the reason for administering drugs, previous disorders of the patient or the outcome of past treatments, and they are the largest source of empirical data in biomedical research, allowing for major scientific findings in highly relevant disorders such as cancer and Alzheimer's disease [6]. Despite the growing interest and ground breaking advances in NLP research and NER systems, easy to use production ready models and tools are scarce in biomedical and clinical domain and it is one of the major obstacles for clinical NLP researchers to implement the latest algorithms into their workflow and start using immediately. On the other hand, NLP tool kits specialized for processing biomedical and clinical text, such as MetaMap [7] and cTAKES [8] typically do not make use of new research innovations such as word representations or neural networks discussed above, hence producing less accurate results [9,10]. We introduce Spark NLP as the one-stop solution to address all these issues.

A primary building block in such text mining systems is named entity recognition (NER) — which is regarded as a critical precursor for question answering, topic modeling, information retrieval, etc [11]. In the medical domain, NER recognizes the first meaningful chunks out of a clinical note, which are then fed down the processing pipeline as an input to subsequent downstream tasks such as clinical assertion status detection [12], clinical entity resolution [13] and de-identification of sensitive data [14]. However, segmentation of clinical and drug entities is considered to be a difficult task in biomedical NER systems because of complex orthographic structures of named entities [15]. Sample NER predictions from a clinical text can be found at Fig. 3.

The next step following an NER model in the clinical NLP pipeline is to assign an assertion status to each named entity given its context. The status of an assertion explains how a named entity (e.g. clinical finding, procedure, lab result) pertains to the patient by assigning a label such as present ("patient is diabetic"), absent ("patient denies nausea"), conditional ("dyspnea while climbing stairs"), or associated with someone else ("family history of depression"). In the context of COVID-19, applying an accurate assertion status detection is crucial, since most patients will be tested for and asked about the same set of symptoms and comorbidities — so limiting a text mining pipeline to recognizing medical terms without context is not useful in practice. The flow diagram of such a pipeline can be seen in Fig. 1.

In our previous study [16], we showed through extensive experiments that NER module in Spark NLP library exceeds the biomedical NER benchmarks reported by Stanza in 7 out of 8 benchmark datasets and in every dataset reported by SciSpacy without using heavy contextual embeddings like BERT. Using the modified version of the well known BiLSTM-CNN-Char NER architecture [17] into Spark environment, we also presented that even with a general purpose GloVe embeddings (GloVe6B) and with no lexical features, we were able to achieve state-of-the-art results in biomedical domain and produces better results than Stanza in 4 out of 8 benchmark datasets (Table Table 1).

In another study [19], we introduced a set of pre-trained NER models that are all trained on biomedical and clinical datasets using the same deep learning architecture. We then illustrated how to extract knowledge and relevant information from unstructured electronic health records (EHR) and COVID-19 Open Research Dataset (CORD-19) by combining these models in a unified & scalable pipeline and shared the results to illustrate extracting valuable information from scientific papers. The results suggest that papers present in the CORD-19 include a wide variety of the many entity types that this new NLP pipeline can recognize, and that assertion status detection is a useful filter on these entities (Fig. 2). The most frequent phrases from the selected entity types can be found at Table 2. This bodes well for the richness of downstream analysis that can be done using this now structured and normalized data — such as clustering, dimensionality reduction, semantic similarity, visualization, or graph-based analysis to identity correlated concepts. Moreover, in order to evaluate how fast the pipeline works and how effectively it scales to make use of a compute cluster, we ran the same Spark NLP prediction pipelines in

Table 1
NER performance across different datasets in the biomedical domain. All scores reported are micro-averaged test F1 excluding O's. Stanza results are from the paper reported in [9], SciSpaCy results are from the scispacy-medium models reported in [10]. The official training and validation sets are merged and used for training and then the models are evaluated on the original test sets. For reproducibility purposes, we use the preprocessed versions of these datasets provided by [18] and also used by Stanza. Spark-x prefix in the table indicates our implementation. Bold scores represent the best scores in the respective row.

Dataset	Entities	Spark - Biomedical	Spark - GloVe 6B	Stanza	SciSpacy
NCBI-Disease	Disease	**89.13**	87.19	87.49	81.65
BC5CDR	Chemical, Disease	**89.73**	88.32	88.08	83.92
BC4CHEMD	Chemical	**93.72**	92.32	89.65	84.55
Linnaeus	Species	86.26	85.51	**88.27**	81.74
Species800	Species	**80.91**	79.22	76.35	74.06
JNLPBA	5 types in cellular	**81.29**	79.78	76.09	73.21
AnatEM	Anatomy	**89.13**	87.74	88.18	84.14
BioNLP13-CG	16 types in Cancer Genetics	**85.58**	84.30	84.34	77.60

Table 2
The most frequent 10 terms from the selected entity types predicted through parsing 100 articles from CORD-19 dataset [2] with an NER model named *jsl_ner_wip* in Spark NLP. Getting predictions from the model, we can get some valuable information regarding the most frequent disorders or symptoms mentioned in the papers or the most common vital and EKG findings without reading the paper. According to this table, the most common symptom is *cough* and *inflammation* while the most common drug ingredients mentioned is *oseltamivir* and *antibiotics*. We can also say that *cardiogenic oscillations* and *ventricular fibrillation* are the common observations from EKGs while *fever* and *hyphothermia* are the most common vital signs.

Disease syndrome disorder	Communicable disease	Symptom	Drug ingredient	Procedure	Vital sign findings	EKG findings
Infectious diseases	HIV	Cough	Oseltamivir	Resuscitation	Fever	Low VT
Sepsis	H1N1	Inflammation	Biological agents	Cardiac surgery	Hypothermia	Cardiogenic oscillations
Influenza	Tuberculosis	Critically ill	VLPs	Tracheostomy	Hypoxia	Significant changes
Septic shock	Influenza	Necrosis	Antibiotics	CPR	Respiratory failure	CO reduces oxygen transport
Asthma	TB	Bleeding	Saline	Vaccination	Hypotension	Ventricular fibrillation
Pneumonia	Hepatitis viruses	Lesion	Antiviral	Bronchoscopy	Hypercapnia	Significant impedance increases
COPD	Measles	Cell swelling	Quercetin	Intubation	Tachypnea	Ventricular fibrillation
Gastroenteritis	Pandemic influenza	Hemorrhage	NaCl	Transfection	Respiratory distress	Pulseless electrical activity
Viral infections	Seasonal influenza	Diarrhea	Ribavirin	Bronchoalveolar lavage	Hypoxaemia	Mildmoderate hypothermia
SARS	Rabies	Toxicity	Norwalk agent	Autopsy	Pyrexia	Cardiogenic oscillations

local mode and in cluster mode: and found out that tokenization is 20x faster while the entity extraction is 3.5x faster on the cluster, compared to the single machine run.

3. The impact to industrial and academic collaborations

As the creator of Spark NLP, John Snow Labs company has been supporting the researchers around the globe by distributing them a free license to use all the licensed modules both in research projects and graduate level courses at universities, providing hands-on supports when needed, organizing workshops and summits to gather distinguished speakers and running projects with the R&D teams of the top pharmacy companies to help them unlock the potential of unstructured text data buried in their ecosystem. Spark NLP already powers leading healthcare and pharmaceutical companies including Kaiser Permanente, McKesson, Merck, and Roche. Since Spark NLP can also be used offline and deployed in air-gapped networks, the companies and healthcare facilities do not need to worry about exposing the protected health information (PHI). The detailed information about these projects and case studies can be found at [20], [21], [22] (see Table 1).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank our colleagues and research partners who contributed in the former and current developments of Spark NLP library. We also thank our users and customers who helped us improve the library with their feedbacks and suggestions.

References

[1] J.A.T. da Silva, P. Tsigaris, M. Erfanmanesh, Publishing volumes in major databases related to Covid-19, Scientometrics (2020) 1–12.
[2] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al., CORD-19: The Covid-19 open research dataset, 2020, ArXiv.
[3] X. Cheng, Q. Cao, S. Liao, An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation, J. Inf. Sci. (2020).
[4] A. Liede, R.K. Hernandez, M. Roth, G. Calkins, K. Larrabee, L. Nicacio, Validation of international classification of diseases coding for bone metastases in electronic health records using technology-enabled abstraction, Clin. Epidemiol. 7 (2015) 441.
[5] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA 309 (13) (2013) 1351–1352.
[6] G. Perera, M. Khondoker, M. Broadbent, G. Breen, R. Stewart, Factors associated with response to acetylcholinesterase inhibition in dementia: a cohort study from a secondary mental health care case register in London, PLoS One 9 (11) (2014) e109484.
[7] A.R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236.
[8] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[9] Y. Zhang, Y. Zhang, P. Qi, C.D. Manning, C.P. Langlotz, Biomedical and clinical english model packages in the stanza python nlp library, 2020, arXiv preprint arXiv:2007.14640.
[10] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, 2019, arXiv preprint arXiv:1902.07669.
[11] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, 2019, arXiv preprint arXiv:1910.11470.
[12] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.
[13] D. Tzitzivacos, International classification of diseases 10th edition (ICD-10):: main article, CME: Your SA J. CPD 25 (1) (2007) 8–10.
[14] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.
[15] S. Liu, B. Tang, Q. Chen, X. Wang, Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries, Information 6 (4) (2015) 848–865.

[16] V. Kocaman, D. Talby, Biomedical named entity recognition at scale, 2020, arXiv preprint arXiv:2011.06315.
[17] J.P. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Trans. Assoc. Comput. Linguist. 4 (2016) 357–370.
[18] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, J. Han, Cross-type biomedical named entity recognition with deep multi-task learning, Bioinformatics 35 (10) (2019) 1745–1752.
[19] V. Kocaman, D. Talby, Improving clinical document understanding on COVID-19 research with spark NLP, 2020, arXiv preprint arXiv:2012.04005.
[20] J.S. Labs, Apache spark NLP for healthcare: Lessons learned building real-world healthcare AI systems, 2021, https://databricks.com/session_na20/apache-spark-nlp-for-healthcare-lessons-learned-building-real-world-healthcare-ai-systems, (Online; Accessed 22-Jan-2021).
[21] J.S. Labs, NLP case studies, 2021, https://www.johnsnowlabs.com/nlp-case-studies/, (Online; Accessed 22-Jan-2021).
[22] J.S. Labs, AI Case studies, 2021, https://www.johnsnowlabs.com/ai-case-studies/, (Online; Accessed 22-Jan-2021).