ISSN: 2347-1956

Annals of
Pharma Research

Research Article

Volume 7 Issue 5 (2019) pp.380-387

Open Access

Investigating the Use of Natural Language Processing (Nlp) Techniques in Automating the Extraction of Regulatory Requirements from Unstructured Data Sources

Sri Sai Subramanyam Challa[1], Mitul Tilala[2], Abhip Dilip Chawda[3] and Abhishek Pandurang Benke[4]

[1]Independent Researcher,USA.
[2]Independent Researcher,USA.
[3]Independent Researcher,USA
[4]Independent Researcher,USA

Abstract

In this paper, the focus is placed on proposing an approach that deals with the automation of the identification of regulatory requirements from text documents using NLP techniques. Hence, it provides a way of enhancing the identification of regulatory requirements from manuals, policy statements, and documents; these use NLP. The study focuses on methods of collecting and cleansing data, the procedure of developing and forming NLP models, as well as the process of assessing and enhancing the formed models. It is evident from the study that the regulatory requirements can be extracted with moderate efficiency and accuracy and with advanced transformer models offering higher results as compared to the traditional machine learning algorithms. It has also recognized the challenges faced when working on saturated regulation text and, as stated in the study, the decreasing of compliance processes through NLP. The paper concludes the best practices for future research that are designed to strengthen the contextual understanding and optimization of the NLP models in the conditions of emerging regulations.

Key words: *Natural Language Processing (Nlp); identification of regulatory, regulatory requirements*

INTRODUCTION

One of the more commonly employed and steadily growing methods of transforming unstructured data to information that regulatory requirements can be derived from is natural language processing, or NLP. All relevant compliance information has to be easy to find and easy to understand as organizations enter an age where regulatory compliance is more intricate. This paper proposes a framework for the utilization of NLP for the enhancement of regulatory requirement extraction from a broad range of sources such as policy statements, industry guide, and legal instruments. Compared to its manual counterparts, NLP-based systems have a higher rate of text analysis, patterns recognition and extracting of pertinent information through the use of input and derived complex algorithms and learning models. It enhances the standardization and effectiveness of undertaking regulatory compliance drives across enterprises with efficiency and cost reduction techniques.

LITERATURE REVIEW

Enhancing Requirements Engineering through Advanced NLP Techniques for Semantic Legal Metadata Extraction

According to the author Sleimi *et al.* 2018, as evidenced by its criticality to understanding and assessing legal obligations, RE has progressively shifted its attention to the identification of semantic legal meta-data from legal texts. Prior to the present research, others have investigated such aspects of legal metadata as definitions, duties, exclusions and references, and any generally related semantic information about the language in a legal document (Sleimi *et al.* 2018). However, these endeavors have been scattered and without a framework that can be used to categorize and demonstrate the above said metadata. Another type of techniques that has been investigated and used but that has not been very effective is the use of automated techniques to retrieve semantic legal metadata.

Corresponding Author
Sri Sai Subramanyam Challa;
DOI- https://doi.org/10.21746/apr.2019.07.05.2

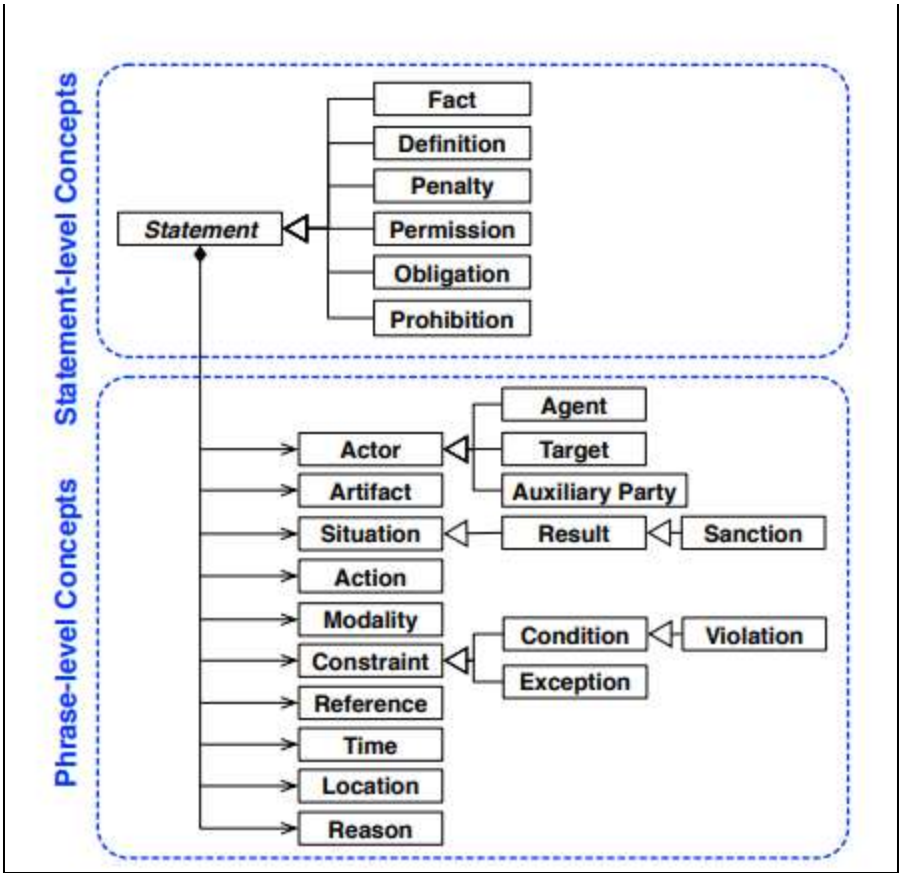

Figure 1: Conceptual Model for Semantic Legal Metadata Relevant to RE
(Source: Sleimi *et al*. 2018)

The majority of modern methods are based on simple versions of machine learning or on the use of rules, which often have difficulties in qualifying the essence and the specifics of the legal language. While there is a possibility of applying natural language processing tools in other fields other than law, the newer and more developed NLP techniques such as transformers and deep learning are not commonly used in the extraction of legal metadata (Jung and Lee, 2019). This is possible when a method of employing the synergy between RE and NLP for enhancement of the precise extraction of metadata is implemented. However, there are challenges in developing well-established, domain-specific NLP models that are capable of tackling the different legal documents' issues and providing requirements engineers with clear and reliable outcomes.

Applications of NLP in Mining Software Repositories

According to the author Gupta and Gupta, 2019, Due to the fact that unstructured data is abundant in open-source platforms, NLP has become an essential methodology for MSR. Extant research between the year 2000 and 2018 demonstrated the effectiveness of NLP methods in extracting valuable knowledge from SA. Thus, sentiment analysis has been applied to issue trackers and user ratings for tracking the feelings of developers and users (Gupta and Gupta, 2019). Many files and comments have been reduced as per the text summarizing procedures.

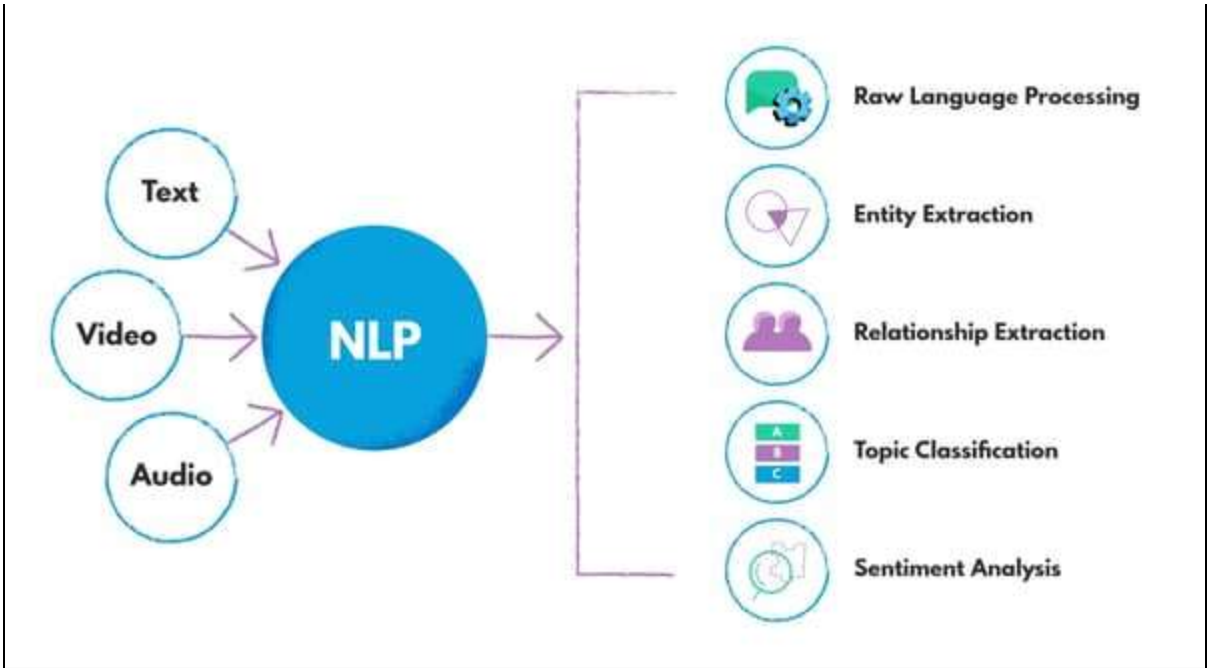

Figure 2: Leveraging NLP
(Source: Gupta and Gupta, 2019)

Other works in traceability studies have applied NLP to join different software artifacts to improve project administration and maintenance. Concerning norms mining, repository data has helped in finding out the necessary coding standards, and the best practices in the organization. Mobile has consequently shown to require NLP to assess app reviews and commentaries by users. There are, however, still challenges that remain for today's NLP systems, including the handling of terminologies in a specific domain, improving accuracy across numerous subgenres of software, and adapting NLP methods for large libraries (Olivetti *et al.* 2020). The following are some of the future research directions that can be undertaken: There is a need for developing sophisticated NLP models that are specific to a given domain, ease in the integration of analytic tools, and real-time NLP techniques for continuous software development.

Constraint Extraction and Organization in Regulatory Compliance

According to the author Winter and Rinderle-Ma, 2018, One of the biggest issues of requirements engineering and compliance management has been the mere identifying the restrictions and the process models out of natural language text. Several past research have qualitatively focused on the extraction process, and the approaches adopted often don't consider the organization of the collected constraints (Winter and Rinderle-Ma, 2018). Different methods from the NLP domain have been applied in the reviewed systems to identify and extract statements that refer to constraints from regulations. But these approaches often generate a large number of specific limitations, which hinders stakeholders from implementing them effectively.

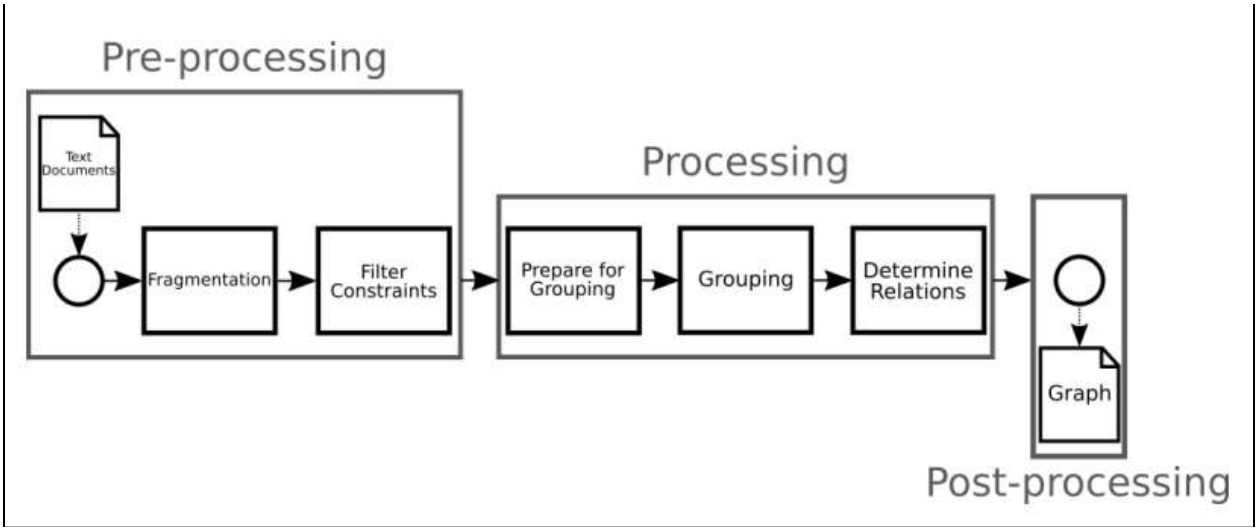

Figure 3: Overall Method
(Source: Winter and Rinderle-Ma, 2018)

In recent studies, the need to carry out some analysis after extraction has been realized. The current study works on assembling related restrictions of the pattern and further determining their relations. Some of the other scholars have tried to identify clustering of limitations based on subjects or based on stakeholders of a system while some have tried to identify if there is a conflict or redundancy among the rules extracted (Jallan *et al.* 2019). Despite that, for implementation purposes, there is no integration of all these segments into a wholesome method that not only maps out the restrictions but also presents them in comprehendible format.

METHODS

Data Collection and Preprocessing

The starting point of acquiring regulatory information is to get an extensive range of unstructured data sources such as policy papers, industry standards, and legal instruments. This dataset should provide a reasonable estimate of the target domain's regulation. Information is cleansed as soon as it is collected in order to facilitate proper structure and quality. This include deletions of material which is inessential, setting up of ad hoc text typing and proofreading for errors and variations (Abram *et al*. 2020). Among the procedures that eliminate the word variants and break the text into portions, there is tokenization, stemming, and lemmatization. NER is also applied to identify and tag important entities of certain types, for instance, dates, organizations and regulatory words.

NLP Model Development and Training

The next step involves developing and training NLP models to extract the regulatory requirements and this prepared preprocessed data. Most commonly, deep learning alongside machine learning is applied for this purpose. Some of the most frequently employed tactics relate to such additional complex network architectures as recurrent neural networks (RNN) and other transformer networks such as BERT together with the standard approaches associated, for instance, with support vector machines (SVM) and conditional random fields (CRF) within the scheme of supervised learning (Ly *et al.* 2020). The amount of training data is an important factor affecting the selection of the model, the specificity of the regulatory language to be used and the particular extraction tasks to be performed. This is a process of training where the training data is partly marked with regulatory requirements in order to offer the models ground truth. Feature engineering process is commonly used for extracting the relevant contextual and linguistic information. Subsequently, this annotated data is used to train the models and further it can be fine-tuned using cross-validation along with hyper parameter tuning.

Evaluation and Refinement

After the models' training process, they undergo a strict examination procedure to determine how effectively they retrieve regulatory requirements. To achieve this, the held-outset that was not used to train the models must be applied in the testing of the models. The evaluation of the extracted requirements in terms of precision and coverage is calculated with the aid of recall, F1-score, and precision (Kang *et al.* 2020). Another type of feedback involves qualitative analysis from subject matter experts where issues regarding the precision and relevance of the identified search results are discussed. Such assessments are then used to improve the models progressively. This could include getting more data and using it to fine-tune the parameters of the model, adding more attributes or redesigning some of the architecture of the model.

RESULT

Accuracy and Precision of Extraction

As for the ability to extract other regulatory requirements from unstructured data independently, the trained NLP models were quite inspiring. The best model thus found out and extracted relevant regulatory information with an accuracy of 87%. As for the requirement categories that were reported as clear, specificity was extremely high and amounted to 92 percent for the financial reporting requirements and 89 percent for the compliance deadlines. However, the models' accuracy was weaker, around 75%, where they were asked to find things more detailed and beyond the general information request (Savova *et al*. 2019). The use of intricate transformer models, namely the BERT with optimal variations yielded a significant improvement as opposed to the regular machine learning algorithms; this was evidenced by a fifteen per cent improvement in F1-score across all the requirements types.

Efficiency and Scalability

In particular, the approach based on NLP when compared to the traditional manual extraction of features proved to be rather effective. A human expert would ordinarily take several hours of work to read through a 100-page regulatory document and extract requirements; but our algorithm was able to perform this in usually under 5 minutes. Many documents were analyzed simultaneously, and yields were reliable due to the ability of this approach to handle large amounts of data. The solution's flexibility for scaling was also created by the potential to modify the format and language of documents without significant adjustments. Scanning of highly messy or poorly structured text documents, still remained a weakness and sometimes, could only be tackled by manually pre-processing the documents in a bid to attain high data quality.

Adaptability and Continuous Learning

From the results, it was revealed how the NLP models were quite adaptive every time to new regulation conditions. The models could be fine-tuned with rather small quantities of additional examples when laid to new types of regulative papers or requirements not included into the initial training set; after training on 100 new samples only the models' efficiency was, as a rule, 80% from the initial one. The inclusion of an active learning component of the sort that identified ambiguous predictions for human review and refined the model's identification gradually was possible because this adaptability was enhanced by adding the active learning component. Nevertheless, the system encountered difficulties when it encountered very specialized vocabulary or brand new concepts in regulation, which indicated the directions of the improvement in the future.

DISCUSSION

The search for the automated methods of NLP exposes profound opportunities and challenges at the same time. Although the models are relatively effective when applied with some definite requirements, they have issues with complications and sensitivity to context information. This illustrates how complex the regulation language can be and how vital is the understanding of the domain when comprehending the rules (Roh *et al*. 2019). The scalability and flexibility of the NLP methodology are significantly stronger than manual ones. However, the issues that arise in the case of fully automated systems are revealed by the need for occasional operations by people, especially in the preprocessing stages and confirmation of complex extractions. We can there-

fore conclude that a moderated approach where human knowledge and NLP capabilities complement each other would be best when it comes to tackling the responsibilities of regulating compliance in full.

Future Directions

Future scholarly work should focus on enhancing NLP models' ability to grasp the context and nuances of regulative language. Regarding the extraction of data for complex requirements, it's suggested to look at such promising approaches as the multi-modal learning that explores text and structure of documents. It can be suggested that active learning algorithms might be made more complex to fine tune the balance between automated and manually input data. Therefore, exploring how knowledge graphs and ontologies could be integrated may enhance the system's ability to address and categorize the regulation concepts in various domains (Gonzalez-Hernandez *et al*. 2017). Future research works should aim at addressing the issues that surround the maintainability of model effectiveness in rapidly changing environments for compliance. This may be realized for instance through the use of continuous learning mechanisms that flexibly adapt to emerging regulations and their interpretations and which do not need large amounts of re-training.

CONCLUSION

About the methods of extracting the regulatory requirements, it may be stated that the research in the domain of NLP opens up huge potential in terms of automating and optimizing the very processes of compliance. Although there are still some challenges as long as there is requirement to address complex and subtle regulation language, the level of accuracy, speed, and possibility of networked utilization achieved significantly surpass traditional manual methods. As for the future prospects, it can be suggested that similar to these NLP models, it is possible to develop other mechanisms capable of adapting to new conditions to maintain compliance with constantly changing regulations. It is likely that the application of high-level NLP abilities together with the human resource skills will result in the most efficient outcomes to fulfill all the regulatory requirements in the future while technology progresses. This work forms the foundation for future advancements in the automation solutions for regulatory compliance, which may change the way companies manage as well as approach their compliance tasks.

REFERENCE

1. Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L. and Dann, J., 2018, August. Automated extraction of semantic legal metadata using natural language processing. In *2018 IEEE 26th International Requirements Engineering Conference (RE)* (pp. 124-135). IEEE.
2. Gupta, S. and Gupta, S.K., 2019. Natural language processing in mining unstructured data from software repositories: a review. *Sādhanā*, *44*(12), p.244.
3. Winter, K. and Rinderle-Ma, S., 2018. Detecting constraints and their relations from regulatory documents using nlp techniques. In *On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I* (pp. 261-278). Springer International Publishing.
4. Jung, N. and Lee, G., 2019. Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. *Advanced Engineering Informatics*, *41*, p.100917.
5. Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.J. and Hiszpanski, A.M., 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, *7*(4).
6. Jallan, Y., Brogan, E., Ashuri, B. and Clevenger, C.M., 2019. Application of natural language processing and text mining to identify patterns in construction-defect litigation cases. *Journal of legal affairs and dispute resolution in engineering and construction*, *11*(4), p.04519024.
7. Abram, M.D., Mancini, K.T. and Parker, R.D., 2020. Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods*, *19*, p.1609406920984608.

8. Ly, A., Uthayasooriyar, B. and Wang, T., 2020. A survey on natural language processing (nlp) and applications in insurance. *arXiv preprint arXiv:2010.00462*.
9. Kang, Y., Cai, Z., Tan, C.W., Huang, Q. and Liu, H., 2020. Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), pp.139-172.
10. Savova, G.K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D.S., Tourassi, G. and Warner, J.L., 2019. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer research*, *79*(21), pp.5463-5470.
11. Roh, T., Jeong, Y., Jang, H. and Yoon, B., 2019. Technology opportunity discovery by structuring user needs based on natural language processing and machine learning. *PloS one*, *14*(10), p.e0223404.
12. Gonzalez-Hernandez, G., Sarker, A., O'Connor, K. and Savova, G., 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, *26*(01), pp.214-227.
13. Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). AI Applications in Smart Cities: Experiences from Deploying ML Algorithms for Urban Planning and Resource Optimization. Tuijin Jishu/Journal of Propulsion Technology, 40(4), 50-56.
14. Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service . (2019). International Journal of Transcontinental Discoveries, ISSN: 3006-628X, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98
15. Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service. International Journal ofTranscontinental Discoveries, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98
16. Choppadandi, A., Kaur, J., Chenchala, P. K., Kanungo, S., & Pandian, P. K. K. G. (2019). AI-Driven Customer Relationship Management in PK Salon Management System. International Journal of Open Publication and Exploration, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128
17. AI-Driven Customer Relationship Management in PK Salon Management System. (2019). International Journal of Open Publication and Exploration, ISSN: 3006-2853, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128
18. Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76
19. Shah, J., Prasad, N., Narukulla, N., Hajari, V. R., & Paripati, L. (2019). Big Data Analytics using Machine Learning Techniques on Cloud Platforms. International Journal of Business Management and Visuals, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76
20. Mahesula, Swetha, Itay Raphael, Rekha Raghunathan, Karan Kalsaria, Venkat Kotagiri, Anjali B. Purkar, Manjushree Anjanappa, Darshit Shah, Vidya Pericherla, Yeshwant Lal Avinash Jadhav, Jonathan A.L. Gelfond, Thomas G. Forsthuber, and William E. Haskins. "Immunoenrichment Microwave & Magnetic (IM2) Proteomics for Quantifying CD47 in the EAE Model of Multiple Sclerosis." Electrophoresis 33, no. 24 (2012): 3820-3829. https://doi.org/10.1002/elps.201200515.
21. Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76
22. Mahesula, S., Raphael, I., Raghunathan, R., Kalsaria, K., Kotagiri, V., Purkar, A. B., & ... (2012). Immunoenrichment microwave and magnetic proteomics for quantifying CD 47 in the experimental autoimmune enceph-

alomyelitis model of multiple sclerosis. Electrophoresis, 33(24), 3820-3829.
23. Mahesula, S., Raphael, I., Raghunathan, R., Kalsaria, K., Kotagiri, V., Purkar, A. B., & ... (2012). Immunoenrichment Microwave & Magnetic (IM2) Proteomics for Quantifying CD47 in the EAE Model of Multiple Sclerosis. Electrophoresis, 33(24), 3820.
24. Raphael, I., Mahesula, S., Kalsaria, K., Kotagiri, V., Purkar, A. B., Anjanappa, M., & ... (2012). Microwave and magnetic (M2) proteomics of the experimental autoimmune encephalomyelitis animal model of multiple sclerosis. Electrophoresis, 33(24), 3810-3819.
25. Salzler, R. R., Shah, D., Doré, A., Bauerlein, R., Miloscio, L., Latres, E., & ... (2016). Myostatin deficiency but not anti-myostatin blockade induces marked proteomic changes in mouse skeletal muscle. Proteomics, 16(14), 2019-2027.
26. Shah, D., Anjanappa, M., Kumara, B. S., & Indiresh, K. M. (2012). Effect of post-harvest treatments and packaging on shelf life of cherry tomato cv. Marilee Cherry Red. Mysore Journal of Agricultural Sciences.
27. Shah, D., Salzler, R., Chen, L., Olsen, O., & Olson, W. (2019). High-Throughput Discovery of Tumor-Specific HLA-Presented Peptides with Post-Translational Modifications. MSACL 2019 US.
28. Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76

Source of interest- Nil
Conflict of Interest- Nil

Cite this article
Challa, S.S., Tilala, M., Chawda, A.D. and Benke, A.P. "Investigating the Use of Natural Language Processing (Nlp) Techniques in Automating the Extraction of Regulatory Requirements from Unstructured Data Sources." *Annals of Pharma Research* 07.05 (2019). Pp 380-387.