iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57762

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Natural Language Processing Based Classification of Publication Data

Anshul Das[1], Prachi Goel[2], Apurva Jain[3]
[1]*CSE Department, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India*
[2, 3]*CSE Department, Assistant Professor, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India*

Abstract: ***In the ever-expanding landscape of scholarly publications, the need for efficient and accurate methods of classifying and organizing vast amounts of information has become imperative. This research explores the application of Natural Language Processing (NLP) techniques to enhance the classification of publication data. By leveraging advanced linguistic and machine learning approaches, we aim to automate and optimize the categorization of diverse publications, thereby facilitating streamlined access to relevant knowledge.The proposed methodology involves the extraction of key features from textual content, such as abstracts, titles, and keywords, using state-of-the-art NLP algorithms. These features serve as input for a robust classification model that is trained on a diverse dataset of publications spanning various domains. The model's performance is fine-tuned through iterative processes, ensuring adaptability to the nuances and evolving trends within different research fields.***
Furthermore, we explore the integration of domain-specific ontologies and semantic analysis to enhance the precision and granularity of classification. This allows for a more nuanced understanding of the relationships between publications, enabling users to navigate through knowledge landscapes with increased contextual relevance.The study's significance lies in its potential to revolutionize the way researchers, academics, and professionals access and organize vast amounts of information. The proposed NLP-based classification system not only promises efficiency in information retrieval but also lays the groundwork for developing intelligent recommendation systems tailored to individual user preferences and research interests.Ultimately, this research contributes to the evolving field of information science by presenting a novel approach to publication data classification that aligns with the accelerating pace of information creation and dissemination in today's knowledge-driven society.
Keywords: Natural Language Processing, Publication Classification, Machine Learning, Information Retrieval, Semantic Analysis, Ontologies, AutoGluon.

I. INTRODUCTION

Natural Language Processing (NLP) is a revolutionary technology that enables computers to decipher and understand human speech . This study provides an indepth look at the application of NLP in the distribution of biological sciences in India. The aim is to supp ort artificial intelligence to better understand the contents of the texts in the DBT Apex BTIC database by dividing them into differe nt classes.Starting from the DBT Apex BTIC Preprocessing library, the data is divided into 11 classes, carefully organizing and reco rding the data, different for the better. The importance of creating subcategories. The next step involves data preparation, statistical a nalysis, and classification using AutoGluon, an Auto MlLibrary.

This study examines NLP classification results by name and general content. The results reveal many successes, highlighting the im portant role of quantitative and qualitative data. AutoGluon's role in the development of machine learning techniques is significant a nd forms the basis for detailed investigations into the interaction between product features and NLP results. This study concludes by demonstrating the usability of all codes used, clarifying and encouraging further research into NLP in the classification of biological sciences.

Initially, a total of 333 records were taken from the above given database. The classes which didn't have a certain amount of records (atleast 10), we neglected those classes as shown in the table below which are highlighted in red. So in order to make the data easy to recognized by the program we gave them a label which made the data more classified as per there classes. Further to make the data more classified in their respected classes we even divided the classes into sub classes which might help in differentiating the new data into their specific classes.

Under classes table we had certain columns such as title, sub class, label and reference. From the database we visited every site which was given under our records but some of them were not published yet (highlighted in yellow), some of them were not reaching the source of the data (highlighted in red) and at last the most favorable and helpful data which would help the program to distinguish different classes, the published data. From the table you might get a better understanding of these data.

©IJRASET: All Rights are Reserved

As we have discussed before the classes which does not meet the minimum requirement of records were neglected so we made 8 more tables similar to Table S1. For testing and training the data for title and label we put the label in front of every title as per their respected classes and after assigning the labels in front of each and every title we had around 331 data sets. After that we shuffled the data sets and divided them into 1:3 ratio. So for testing we took 83 data sets and whereas for training we took 248 data sets. For the full text based data set we only considered those data which were currently published and their pdfs were available in their respected links. After downloading all those pdfs we then labeled them according to their classes.And in this case we had a total of 353978 lines from which we considered 265483 lines for training the data and the rest 88495 for testing data (1:3 ratio).

II. DATA STATISTICS

The distribution of records into various classes is shown in Table 1.

Data set preparation

At first in order to test our NLP program we took the title and label from each class and also in order for the program to be more specific with differentiating the data we put the label at the end of each title so that the program have a good note of which class the title is from as stated above. Although the results were bad, we at least got the rough idea of how to make it work with more data and to get a better result than before. Also the program works with tsv file so in order to load the data within the program we first had to convert the class data into tsv format. So at first we shuffled the data and put 75% of the data to train in tsv format and then the rest of the data in test also in tsv format. Later we gave these data to the program to train and test and we got the estimated score between 0.19-0.24 (roughly 19-24 %).

Now we used more data to get better result so in order to do that we took the published articles and their labels of their respected classes and downloaded theirpdfs. But as we knew that the NLP program works only under tsv or csv file so the data collected in the form of pdf was not enough, so in order to make it work we then made a python program to convert the pdf into text format and then write thantext file into a tsv file which ultimately gave us the right file format for our NLP program.

Moreover the NLP program which we used for testing and training the title-label data and ultimately getting a score ultimately concludes that Accuracy and time may be varying according to the applied preset.

Even in the full text based case we first shuffled the data and then divided them into training and testing sets. But as we knew that their were just title and label in the first case so labeling the label in front of them never seem very important whereas in this full text based case we need to label each and every line of the pdf with their respected labels as of their classes or else the program will not consider the lines which are not labeled. But we did it while converting pdf to tsv by just writing those labels in front of every-line by starting a loop which read each and every-line of the pdf.

But in this case we got a better result and thus it clearly shows that having more data will generally create better results. The result we got in the full text-label case was between 0.96-0.97 (roughly 96-97%).

We can also chose presets according to our preference.

III. EXPERIMENTAL METHOD OR METHODOLOGY

Auto Gluon, it is a new open source Auto Ml library that automates deep learning (DL) and machine learning (ML) for real world applications involving image, text and tabular data-sets. Under Auto Gluon we used Text Predictor is similar to AutoGluon's Tabular Predictor. We format NLP data sets as tables where certain columns contain text fields and a special column contains the labels to predict, and each row corresponds to one training example. Here, the labels can be discrete categories (classification) or numerical values (regression). In fact, Text Predictor also enables training on multi-modal data tables that contain text, numeric and categorical columns and also support solving multilingual problems.

IV. RESULTS AND DISCUSSION

In the case of title-label, the results were pretty bad but it atleast gave us the idea that we are going in the right path, whereas in the case of full text-label the results were average as the data was pretty vast and was in a good amount that's why the program didn't had to loss much of its data while epoching the data sets (around 0.0001-0.1) but in the case of title the loss was very high (over 2.0). Loss also matters a lot in the results the lower the loss the higher the result, you'll get to know it when you will run the program. But more data played a major role as the results which we had in the full text based data set were over 4 times better then the title based data set where the data was very low. Also Auto Gluon played a major role in completing the program, it is also very useful in many ways such as:

- Quickly prototype deep learning and classical ML solutions for your raw data with a few lines of code.

- Automatically utilize state-of-the-art techniques (where appropriate) without expert knowledge.
- Leverage automatic hyper parameter tuning, model selection/ensembling, architecture search, and data processing.
- Easily improve/tune your bespoke models and data pipelines, or customize Auto Gluon for your use-case.

Nowadays NLP is in great demand and Auto Gluon provides you with easy and simple ways to learn and master it in few steps. It is very useful for the programmers who are just starting to learn it as complexity can be overcomed by practicing new programs but the basic is what people struggle to understand.

Figures and Tables

Sr.No.	Class Name	No. of sub classes	No. of records	Label
1	Bibliographic	11	45	1
2	Genomics	13	58	2
3	Human Disease & Infections	7	30	3
4	Immunology	4	23	4
5	Metobolic & Signaling Pathways	1	1	11
6	Protein	7	56	5
7	RNA	8	43	6
8	SNP	3	11	7
9	Software	8	16	8
10	Structure	5	49	9
11	Taxonomy	1	1	10
	Total		333	

Figure 1. Different classes with there respected labels

Based On	Total data trained	Total data tested	medium_qual	High_qual	Best_qual
Full Text	265483	88495	0.9679	0.9692	0.971
Title	248	83	0.2409	0.1927	0.2168

Figure 2. Results acquired with different presets

V. CONCLUSION AND FUTURE SCOPE

We get to know that the NLP program is not so successful with the title based data set or you can say where the data is limited and very low whereas in the case of full text based data set the program was quite successful or in other words where the data is good and in large amount the program worked very well with the larger data. We also get to know that larger the data lesser the loss and lesser the data larger the loss which implies data is inversely proportional to loss in the above NLP program and the time period also plays a major role if you limit the time limit to 60 secs than the result or the accuracy of the data may differ, it generally affect theaccuracy of larger data but in the case of lesser data such as the title base data the 60 sec is more than enough as it gave the same result in both the time periods (second time we set the time limit as None). Later we can use the created model for classifying the unclassified data in our datasets. And in future to make a web interface regarding this model so as to make it easier for others to ccalssify there respected research papers just on the basss of their title and abstract data.

REFERENCES

Research Papers and Journals:

[1] Gupta, P., & Khatter, K. (2018). Natural Language Processing: A Comprehensive Review. In 2018 4th International Conference on Computing Sciences (ICCS) (pp. 196-201).

[2] Singh, A., & Gaharwar, R. (2017). Machine Learning and Natural Language Processing: An Overview. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7).

[3] Gupta, D., & Lehal, G. S. (2009). A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence, 1(1), 60-76

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems.

[5] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1408.5882.

[7] Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11), 613-620.

[8] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys (CSUR), 34(1), 1-47.

[9] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[10] Wang, P., & Zhang, J. (2018). A Review of Transfer Learning for Natural Language Processing. Journal of Computer Science and Technology, 33(6), 1176-1192.

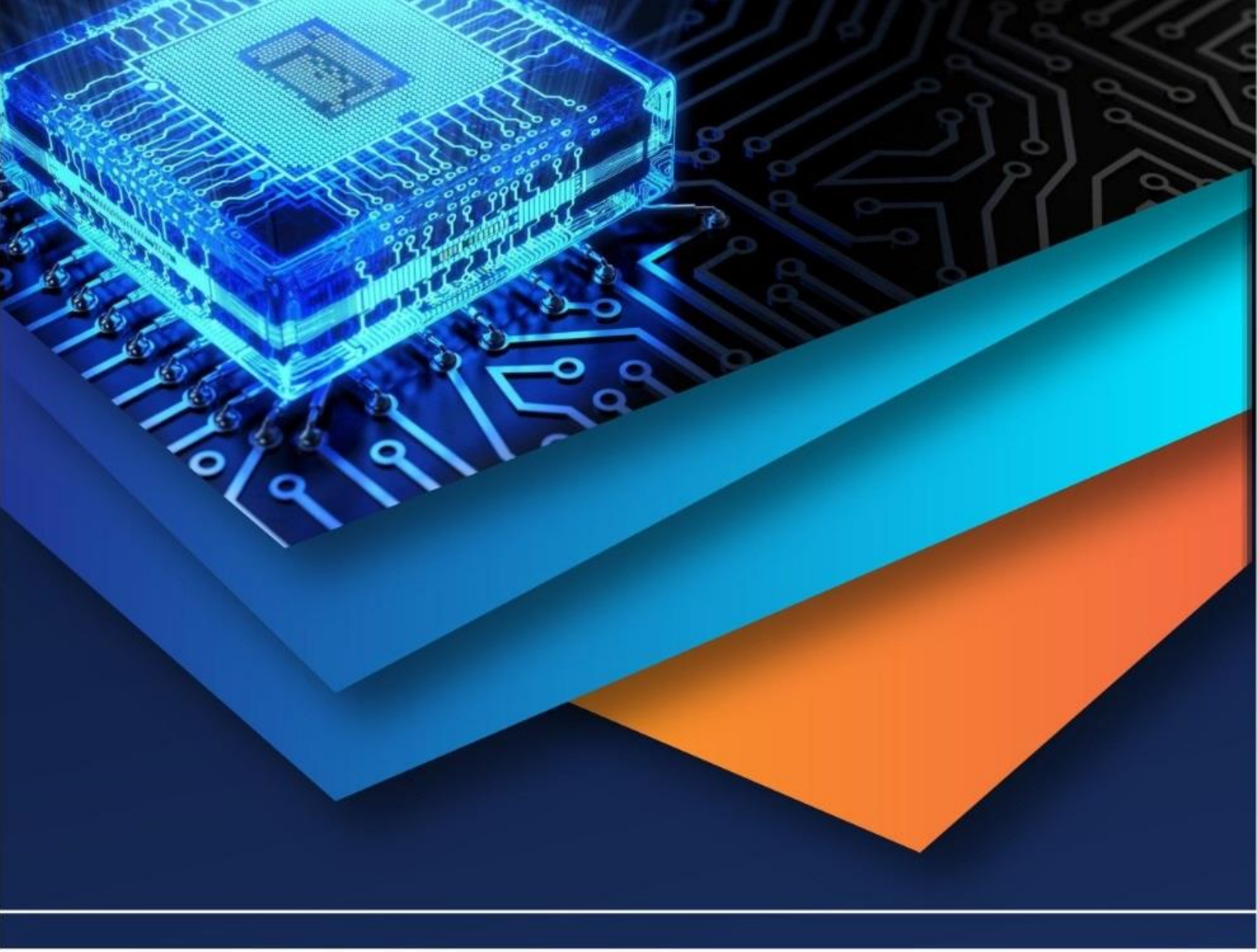
[11] Smith, A., & Wang, X. (2018). Transfer Learning in NLP. In Advances in Neural Information Processing Systems.

Books:

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

[13] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.

[14] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)