International Research Journal of Computer Science (IRJCS)
Issue 04, Volume 07 (April 2020)
ISSN: 2393-9842
www.irjcs.com

CLASSIFICATIONS OF ARTICLES BASED ON THE SOURCE USING MULTI-CLASS CLASSIFICATION

Shreyas Renga
Computer Science & Anna University, Chennai, INDIA
shreyasrenga99@gmail.com

Abishek Ganapathy
Computer Science & Anna University, Chennai, INDIA
abishekganapathy15592@gmail.com

T. Hasith Ram Varma
Computer Science & Anna University, Chennai, INDIA
hasithram@gmail.com

Manuscript History
Number: **IRJCS/RS/Vol.07/Issue04/APCS10084**
Received: 05, April 2020
Final Correction: 19, April 2020
Final Accepted: 28, April 2020
Published: **April 2020**
Citation: Shreyas, Abishek & Hasith, T. (2020). Classifications of Articles Based on the Source Using Multi-class Classification. International Research Journal of Computer Science (IRJCS), Volume VII, 48-53.
DOI://10.26562/IRJCS.2020.APCS10084
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract: In this paper we focus on, helping editors in the newspaper industry, by making their work easy by processing the huge chunks of data they receive in the form of articles that are given to them by multiple news reporters, from different locations. Usually, all the news industry organizations face a great challenge in processing these kinds of huge data, and in this kind of industry, only those organizations that provide good insights will be successful. This type of data can be processed only by using some important methods like Natural Language Processing (NLP). And our model usually would begin with taking all the articles as input, pre-processing it using NLP, adding patterns, and finally finding the good insights that are needed. The main aim of our model is to have high accuracy as well as maintain its robustness. So, the important attributes can be arranged based on the importance discussed in the article and can be highlighted using Displacy. So that it is easy for the Newspaper industry to maintain high standards by establishing a strong connection to reach people. Thus, it would be better we had these articles in the form of tabular data, which would have all the important attributes and hence being our final output. This paper helps the news industry to properly analyse the articles, and to convert them into an easily accessible format, which makes job easy mainly for editors of that particular industry.

Keywords: Natural Language Processing (NLP); Multi-class classification; Spacy Displacy; Entity-Ruler based classifications; articles; source;

I. INTRODUCTION

Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. Thus, our motive is to extract useful information, leaving out the unnecessary ones and turning text classification in a fast and cost-efficient way to enhance decision-making and automate processes. Text analysis, as a whole, is an emerging field of study[1]. Fields such as marketing, product management, academia, and governance a real ready leveraging the process of analysing and extracting information from textual data.

In layman's terms, text classification is the process of extracting generic tags from unstructured text. These generic tags come from a set of predefined categories. Hence the major task of the project is to convert the given data under some attributes, depending upon their choice. The main idea of building this model is to label those articles that are not labelled so that editors need not waste loads of time reading all the articles [2].

II. SYSTEM ARCHITECTURE

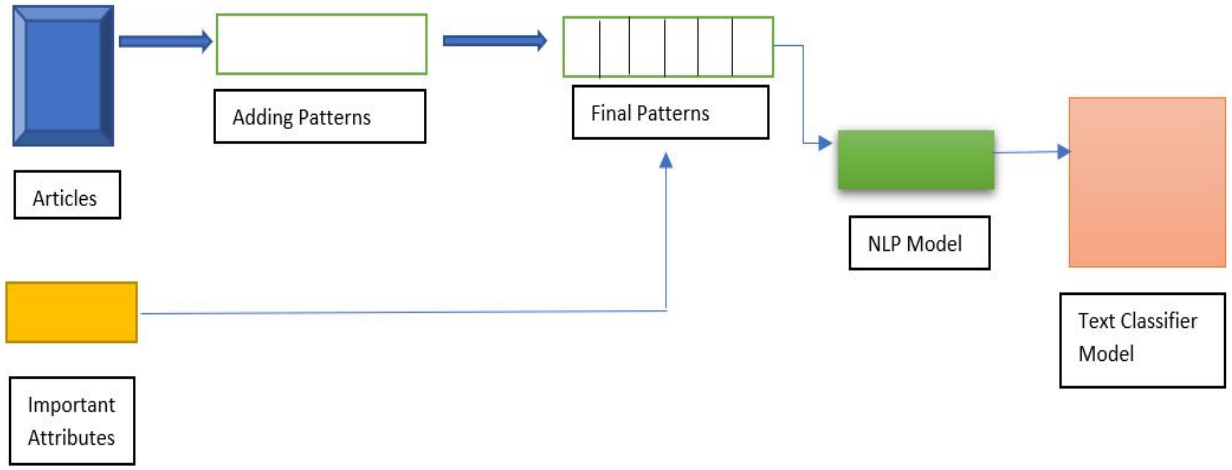

Fig.1 Proposed system block diagram for training

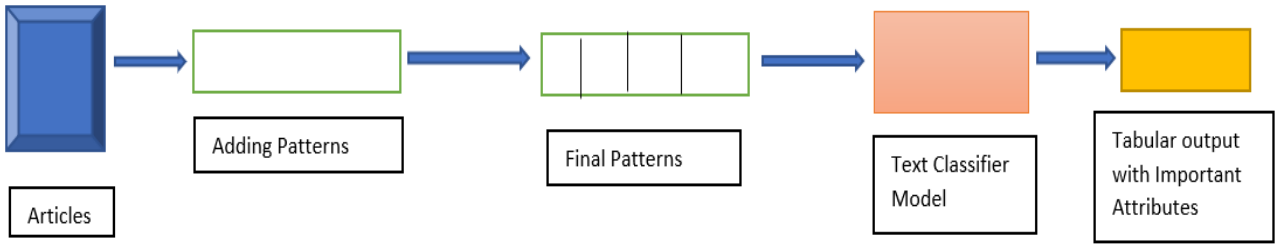

Fig.2 Final proposed flow of project for testing

III. MULTICLASS CLASSIFICATION

Classification and Prediction are two forms of data analysis that can be used to extract models describing important data classes or predict future data trends [4]. Classification problems aim at building an efficient, effective model for predicting class membership of data. There is always a confusion regarding the two terms: multilabel classification and multiclass classification. Multiclass classification refers to a classification task that generally has more than two classes where each sample that is classified can only have one class. Multilabel classification refers to the kind of classification where each sample is assigned multiple target labels. In multiclass classification, it is assumed that (1) only one class label is assigned to each instance, and (2) the class labels are independent of each other. In multiclass classification, accuracy is generally the main evaluation metric. As [5] says, in multiclass the classification performance can be significantly improved by dividing the multiclass problem into intermediate classification problems that are less complex by either reducing its dimensions or by reducing the number of classes. So, in the classification of articles, multiclass classification is used rather than multilabel classification as the words in the articles are supposed to be predicted into multiple classes such as location, names, etc.

IV. PSEUDOCODE FOR THE MODEL

Algorithm for assigning labels for each customer review

1. Start
2. Keep all the articles under one folder name
3. blockList = []
4. Looping through each file:
    a. Replace the words like “Planned”, “Under construction”, and “Completed” with a blank space
    b. Also, replace the source name with just “Source” so as to differentiate between articles
    c. Perform the python Split () operation to split the lines at “Source” and store this list in “articleList”
    d. Now append this articleList to blockList

5. str1 = ‘’.join(str(e) for e in blockList)
6. Create a variable “processed” to store the value returned bynlp objecton str1
7. [(ent.text, ent. label_) for ent in processed. ents] //This is to view the Named Entities in Spacy
8. Create a variable “ruler” that stores the EntityRuler value of the nlp object
9. Now create various patterns for the articles
10. For each pattern created:
    a. Add patterns to “ruler”
11. Add this ruler to nlp object
12. displacy. render (processed, style="ent") //Done to visualize the labels in the text
13. ENT_LIST = []
14. for ent in processed. ents:
    a. ENT_LIST.append((ent.text, ent. label_))
15. ORG_ls = []
16. for ent in ENT_LIST:
    a. if 'ORG' in str (ent [1]):
        i. ORG_ls. append (ent [0])
17. DATE_ls = []
18. for ent in ENT_LIST:
    a. if 'DATE' in str (ent [1]):
        i. DATE_ls. append (ent [0])
19. PERSON_ls = []
20. for ent in ENT_LIST:
    a. if 'PERSON' in str (ent [1]):
        i. PERSON_ls. append (ent [0])
21. QUANTITY_ls = []
22. for ent in ENT_LIST:
    a. if 'QUANTITY' in str (ent [1]):
        i. QUANTITY_ls. append (ent [0])
23. GPE_ls = []
24. for ent in ENT_LIST:
    a. if 'GPE' in str (ent [1]):
        i. GPE_ls. append (ent [0])
25. FAC_ls = []
26. for ent in ENT_LIST:
    a. if 'FAC' in str (ent [1]):
        i. FAC_ls. append (ent [0])
27. CARDINAL_ls = []
28. for ent in ENT_LIST:
    a. if 'CARDINAL' in str (ent [1]):
        i. CARDINAL_ls. append (ent [0])
29. TIME_ls = []
30. for ent in ENT_LIST:
    a. if 'TIME' in str (ent [1]):
        i. TIME_ls. append (ent [0])
31. df=pd. DataFrame () //Used to create a table
32. df['organization’] =ORG_ls
33. df['person’] =pd. Series (PERSON_ls)
34. df['date’] =pd. Series (DATE_ls)
35. df['quantity’] =pd. Series (QUANTITY_ls)
36. df['city’] =pd. Series (GPE_ls)
37. df['location’] =pd. Series (FAC_ls)
38. df ['opening time’] =pd. Series(TIME_ls)
39. This can be made to an excel file by executing the command df.to_excel ('output.xlsx')

	organization	person	date	quantity	city	location	opening time
0	Star Saginaw LLC	Desmone Architects	March	140,000 square feet	Saginaw Township, MI	Bay Road	2-4 p.m.
1	CubeSmart	Dave	about 10 years	87,692 square foot	Value City	Gratiot Road	825 3rd Ave
2	Kmart	Ann Brach	the last two years	117,870 square feet	Tamarac	Rock Island Road	The Hour
3	Storage of America	W. Elk	June 2020	5.1 acres	FL	Port Saint Lucie	2500 N. 24th
4	Kmart	E. Linden Ave.	2670	20,000 square feet	Florida	Washington Road	11 a.m. to 4 p.m.
5	Newman Realty Partners LLC	E. Thompson	mid-to-late June	220,000 square foot	Peters Township	2497 Power Road	about 20 minutes

Fig.3 Final output after implementing the algorithm

V. DIVISION OF ARTICLES INTO A SET OF BLOCKS

The division of articles into different blocks is a very important task because the same reporter may write two different articles in the same one, and this would be the case with all the articles, hence dividing them is a very huge task[3]. Hence this process will be starting by storing all the articles written by one reporter in a single variable, we do this separately for each reporter's articles since all the reporters generally write articles on one kind or type(Merger and acquisition, Industry setting up a self-storage industry, etc).

```
para ="""Two vacant buildings in Saginaw Township, MI, will be turned into self-storage facilities. In March, Star Saginaw

A mixed-use project with apartments and self-storage - the latter with a four-story building with 100000 square feet could
EDG Properties is buying Frontier's former site off of Willard Street, and attorney Adam Blank said they're seeking to upda
Planned
The prospective new owners of the plaza off Clay Pond Road, once home to a Grand Union grocery store, plan to build a self-
A map of a multi-phase self-storage development proposed in Green Oak Township, MI.
Green Oak Township, MI officials and Kril Properties project leaders are discussing building a multi-use and storage develo
Quincy, IL, may become home to two new self-storage facilities after the city council approved a special permit. They would
Under construction
Jumbo Self Storage is developing a 120000-square-foot facility in Stoneham, MA.
Jumbo Self Storage LLC is building a three-story, 120000 square foot facility with 1000 temperature-controlled units at 54
Summit Self Storage, based in Augusta, GA, is building a new facility at 4042 Highway 17 in Mount Pleasant, SC, close to Aw
Completed
```

Fig.4 All the articles being stored in one variable

Now after storing all articles into one variable, we start finding common patterns in all the articles, like each article usually begins with a certain type of words, and ends with some particular pattern, hence that must be found.

```
para1 = re.sub(r"(?si)Planned|Under construction|Completed", r"",re.sub(r"(?si)Source:.*?\n", r"Source:", para))
block_list= para1.split('Source:')
print(block_list)
```

Fig.5 Common Patterns found in each article

Thus now by writing a small snippet of code we can divide the articles into blocks.

```
The prospective new owners of the plaza off Clay Pond Road, once home to a Grand Union grocery store, plan to build a self-st
orage facility at the Bourne, MA site. Clay Pond Acquisitions LLC has a purchase-and-sale agreement for the site, currently o
wned by the Claybourne Trust and run by Winslow Property Management in Lexington, MA.
====================
A map of a multi-phase self-storage development proposed in Green Oak Township, MI.
Green Oak Township, MI officials and Kril Properties project leaders are discussing building a multi-use and storage developm
ent to be built in three phases. It would go up on a nearly 13-acre site. Phase One would have 11463 square feet that could i
nclude office, shop or warehouse space. Phase Two would have four 3000 square foot buildings for self-storage plus outdoor RV
storage. Phase Three would have a 5000 square foot building surrounded by a drive and parking spaces.
====================
Quincy, IL, may become home to two new self-storage facilities after the city council approved a special permit. They would b
e located at 6411 Broadway and 2500 N. 24th St. An ordinance will be written, then have three readings for the council's fina
l approval.
====================
```

Fig.6 Final Division of articles into blocks

VI. DISPLACY AND ENTITY RULER

The EntityRuler lets you add spans to the Doc.ents using token-based rules or exact phrase matches. It can be combined with the statistical EntityRecogniser to boost accuracy, or used on its own to implement a purely rule-based entity recognition system. After initialization, the component is typically added to the processing pipeline using nlp.add_pipe. For usage examples, see the docs on rule-based entity recognition.

```
ruler = EntityRuler(nlp)
```

Fig.7 Entity Ruler for pattern evaluation

DisplaCy is able to detect whether you're working in a Jupyter notebook, and will return mark-up that can be rendered in a cell straight away. When you export your notebook, the visualizations will be included as HTML[6].

```
displacy.render(do, style="ent")
```

Two CARDINAL vacant buildings in Saginaw Township, MI GPE , will be turned into self-storage facilities. In March DATE , Star Saginaw LLC ORG bought the former Value City GPE furniture store building (140,000 square feet QUANTITY) on Bay Road FAC , which has been vacant for about 10 years DATE . Across the street is a CubeSmart ORG site that formerly was a Kmart ORG . Also, Storage of America ORG has bought another former Kmart ORG building on Gratiot Road FAC , and will create a self-storage business there. Newman Realty Partners LLC

Fig.8 Displacy is used to render the text using colour fields

VII. USAGE OF SPACY IN INFORMATION EXTRACTION

spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. If you're working with a lot of text, you'll eventually want to know more about it [6]. For example, what's it about? What do the words mean in context? Who is doing what to whom? What companies and products are mentioned? Which texts are similar to each other? It is designed specifically for production use and helps you build applications that process and "understand" large volumes of text.

It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning. It provides a variety of linguistic annotations to give you insights into a text's grammatical structure. This includes the word types, like the parts of speech, and how the words are related to each other. For example, if you're analysing text, it makes a huge difference whether a noun is the subject of a sentence, or the object – or whether "google" is used as a verb, or refers to the website or company in a specific context.

During processing, spaCy first tokenizes the text, i.e. segments it into words, punctuation and so on. This is done by applying rules specific to each language. For example, punctuation at the end of a sentence should be split off [6].

VIII. CONCLUSION

The Classification of articles based on the source is a very important thing in recent times. And hence, by using Natural Language Processing it helps us to solve a real-world problem of storing, manipulating, extracting and retrieving data from large sources. If done manually then it proves to be costly and time taking. This is because a significant amount of effort and cost is involved in obtaining large labelled data sets. By applying various machine learning techniques we are successfully able to extract useful information leaving out the unnecessary ones turning text classification a fast and cost-efficient way to enhance decision-making and automate processes. Thus our active learning not only provides a way classify but also to predict what is best for an existing company.

ACKNOWLEDGMENT

I would like to thank Ms.R.Thirumahal (Mentor), PSG College of technology for supporting my work, especially the faculties of the department of Computer Science and engineering and its staff; students and my colleagues who helped me in publishing my work.

REFERENCES

1. Menaka.Text Classification using Keyword Extraction Technique,Corpus ID: 212463857, June 2014.
2. Yong-Bae Lee andSung-Hyon Myaeng.Text genre classification with genre-revealing and subject-revealing features.Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, January 2002.
3. Wei-Liang Liu, Hui-Shih Leng,Chuan-Kuei Huang and Dyi-Cheng Chen.A Block-Based Division Reversible Data Hiding Method in Encrypted Images.In Symmetry 9(12):308, December 2017.
4. ChaitraP.C, Saravana Kumar. A Review of Multi-Class Classification Algorithms. International Journal of Pure and Applied Mathematics, Volume 118, No. 14, 2018, 17-26
5. Daniel Silva-Palacios, C`esar Ferri, and Mar´ıa Jos´e Ram´ırez-Quintana. Improving Performance of Multiclass Classification by Inducing Class Hierarchies. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland
6. X. Schmitt, S. Kubler, J. Robert, M. Papadakis and Y. LeTraon,A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate,2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 338-343.

