A Rule And Statistical Based Model For Descriptive Text Analysis Using Natural Language Processing

Indrashis Das, Bharat Sharma, Siddharth S. Rautaray, Manjusha Pandey

Abstract : Data can be called as the most important resource in the Automated world but the challenge we face is that before 80s most of the data was offline written in form of text, picture or paper records. The data present online also is haywire and needs analysis to be done in order to get the information we desired. The need for the text analysis model comes to the rescue where it can be used to analyze the descriptive text. NLP or Natural Language processing can be used for the purpose of analyzing the descriptive text with the help of NLTK. This paper proposes a descriptive text analysis model which can be used for gaining insight or analyzing for the purpose of text analysis.

Keywords: Text Analysis, Data Analysis, Natural Language Processing, Python, NLTK

1. INTRODUCTION

The world we live is of the automation as the Artificial Intelligence and the Machine Learning are a necessity for the society to work efficiently, AI and ML are not just the terms used in the research work these are changing the lives of humans in the real world. For any of the above to work we need data. Data is the new Gold for the upcoming market as data is not just the raw information it is a resource in the world which if used efficiently can be extremely beneficial for the world. Data is any set of characters of numbers which includes text, photographs or audio which can be provide operation or if performed any information can provide information. Data Analysis basically refers to the analysis of the data in order to make conclusion and gain meaningful insights from Data, Data analysis finds the recurring pattern, transforms and model it accordingly to get the meaning out of it. There are basically 4 types of the Data analytics:

1. Descriptive Analysis : This is Data Analysis process where we find answers related to the question for what happened by summarzing all the past data we have accumulated so far and gain the insights on the given data.

2. Diagnostic Analysis : This is the data Analysis process which helps in finding the answer to what will happen, this helps in predicting the outcome of the event in order to gain the future insights, it uses the power of the probability and the Statistical analysis to gain the insights of the future events.

3. Predictive Analysis : This is the data Analysis process which helps in finding the answer to what will happen, this helps in predicting the outcome of the event in order to gain the future insights, it uses the power of the probability and the Statistical analysis to gain the insights of the future events.

4 . Prescriptive Analysis : This is the final step and it combines the information from all the past data analytics in order to predict and survey the actions that needs to be taken, it is a very precise process and AI are the implementation as it utilized the data and then take the decision it is always evolving process. Text analysis is the analysis needed for the text in order to gain the valuable insights as the written document through out the world contain the very important information and there is a lot of issue with the physical document as it can be destroyed, it also needs a maintenance system for it to be preserved also the online documents confronts us with the challenge to sort the document according to our need. Test analysis pushes that barrier as it can be used in order to use and analyze the document according to the need. Text analysis can also be for analyzing the online data as the model can help in tokenizing the data in order to get the desired result from the given string. It can be used for different kinds of analysis scenario for example an online examination system, hospital management system. Natural Language processing helps us in that scenario where the unstructured document or the text can be transformed or integrate into the database in either the role of the HTML tables or can be stored as the file in the database in order to provide the text analysis goal. NLTK or the natural language processing toolkit can be used for the purpose of the text analysis as it generates the space separated tokens with the given string in order to analyze the given string. There are different applications for the model as the examination system conducted in schools are still in the form of pen paper but with the help of our model it can be an online system which can tokenize the written strings and marks can be allotted according to the token in the database. Another example is the hospital management system the diagnosis can be examined for the disease keywords which can be used for grouping patients for either medicine intake or the bed allotment.

2. Literature Survey

Coming down to this section, here one will fine the related study of papers in the similar field of Data Analysis, Text Analysis, Natural Language Processing, application development using Django, and pre-developed test conducting systems. Hence, below are some research works on the above discussed topics and domains in this field. The paper targets the problem of complex text recognition and summarization of the data in the recent world and tries to the counter the problem using the Natural language processing toolkit, it searches for the relationship between the data in order to analyze the complex data, it tokenize the given string and then add tags to the extracted tokens in order to find the relationship between the tags which are specifically attached for this part to run. Then the

data is made a part of the chunk or a group of the similar type. For the further operations the opinion mining is done on the chunks in order to match and concatenate the similar chunks to be precise tokens and the relationship is established on that. The approach only accepts the 1st person and 2nd person concept while a variety of other factors also affect the opinion mining like the negation terms and the attributes related to a particular word. This paper uses the pros of the Natural Language process for the manipulation of the text or in some way the resulting operation of the speech. Apart from using the NLP it also uses the concepts of the digital signal processing which can be used in order to transmit the data or the text and with the help of the signals and the inbuilt library transmitting them in the outer applications which can be used to convert those signal and produce the audio with the given output which are in simpler terms the words or speech. The solution is robust but it has not been tested in the real world in order to gain the ground conditions which can be a huge disadvantage and can alter the results accuracy. The paper deals with the NLP-NG which is a new concept in the world of the natural language processing it is used for the biomedical text analysis as the biomedical text consist of the key terms, it is different from the bioNLP which is quite useful in the biomedical text mining, the NLP-NG consist of basically three components which are NG-Core which works as the language processing unit, NG-DB which is basically the database unit of the NG and it oversees the functionalities regarding to the database, last one is the NG-SEE which is used for the interactive visualization and for the entry purpose. The NLP-NG is quite useful but is quite confided to the field of the biomedical. It has to be modified for a particular domain in order for it to work. This paper focus on the problem of the Natural language processing as the toolkit offers a huge variety of the advantages but have the disadvantages in case of the prefix, suffix and in some cases has to counter the problem of the new term addition as in the field of the zonal morphological search there are various new terms which are added daily and there is no mechanism in the NLP toolkit which can differentiate between the words having no meaning and the words that makes sense. This paper aims at the encryption use of the natural language processing as this explains an algorithm which is based on the Natural language processing and can be used as the base for the encryption as in the field of the information security. The concept of the text watermarking can be used for the concept of the text encryption and various ideologies are proposed from the synonym substitution to the syntactic transformations everything is broken down and explained in the paper. There is also a text decryption format which can be used in order to recover the text from the encrypted form. The decryption method solely depends on the logic applied while encrypting and can change accordingly to that. Though the algorithm is a new concept but is vulnerable and it also needs the concept of the encryption so it can vary according to that and does not have a proper steps. This paper uses the concept of the natural language processing in order to do the sentiment analysis of the text. It is a combination of the machine learning and the NLP techniques. A data gathering module is used to get the search keywords and then it is processed with the help of the machine learning and the NLP then the output is matched to get the sentiment analysis of the analysed text or the feeling in order to get the public mood or the opinion on a issue. This is a mining based algorithm and it ust extract the features or certain kind of words which is fixed, it does not analyzes the complete paragraph or pages. This comes as a disadvantage as there is no mechanism for searching the entire paragraph instead a search is done to gather the information in order for the sentinel analysis. The paper counters the problem of the text mining in the Chinese domain as the information of the power grid is of Chinese language and automated mining is not possible on that in order to extract the information. Natural language processing helps the system to first extract the power keywords from this , by power keywords it means the keywords which are important specifically for the power grid information. This helps in extracting the information from the old text. Though the method clearly is a great but is generalized for a particular domain of the Chinese language aiming with the keywords which might leave some valuable information behind which might be a big disadvantage while considering the kind of the information these power grid holds. The paper discussed a model to help the text analysis of the Bengali language with the help of the unicode. The system converts the text into the unicode form and then the text analysis part is done on the above, the input text is fed word by word and it changes the input to the unicode after which the text analysis is done. The model surely is helpful in changing the codes and text analysis but it is inefficient as of now with the help of Natural language processing the same task can be done easily with its toolkit which in turn increases the efficiency of the system and is faster in order to convert the complete text and analyze it. Proposed Model for Rule and Statistical Based Descriptive Text Analysis using Natural Language Processing & it's implementationAs we are aware of the fact that Natural Language Processing deals with strings and accordingly related tools, techniques and models to evaluate and analyze sentences or string of characters, this section would talk solely about the architecture involved in designing the descriptive text analysis algorithm. When we talk of Data Analytics, handling data in the form of numbers is easy in comparison to data that exists in strings or sentences. It is so because numbers in the form of discrete or continuous values help one to modify accordingly as per the requirements of the analyst, while if the same is to be in case of a strings or sentence, it would take time as capturing pattern and meaning out of sentences and converting them into numerical data is difficult. Moving to the architecture of the designed model, as the code is written in python, it has always been a more python centred or pythonic approach to code out the model as python is a language that has an understandable implementation of code and does not require much deep dive into the code to understand it as mostly it is like sentences in the English language. Also, when someone approaches towards writing a code in python language, the use of loops should be minimized as python gives a much more faster and efficient looping system apart from loops like for loop and while loop. The libraries or rather the packages used for the purpose of Natural Language Processing include NLTK which is python's Natural Language Tool Kit and the other library used is Pandas which is solely used for the purpose of saving data in a tabular approach. Specifically the Stopword sub-package

and Wordnet sub-package are used for the purpose of the text analytics. Stopwords contains the collection of such words which are extremely common or occur very frequently in the English language. Words like a, an, the, in, is, this and much more such words that are not required while analyzing text, hence such words need to be removed and for this purpose Stopword subpackage is used. While the subpackage termed as Wordnet is used mostly for the purpose to deal with synonyms and antonyms in the entered data by the user. It might happen that the answer entered by the used contains words for which it's synonyms existed in the set of the keywords as set by the examiner for a particular test. In such a case, the answer entered would be correct and would be evaluated with a lower marks. For this, the evaluation of synonyms or antonyms is very important or else there would be an injustice with the student or person appearing for the exam and hence this would turn out to be as a bug in the deployed software or product in production. Also, a direct analysis of text is impossible, hence cleaning out and pre-processing the data is very important as if not done so, the model would give unexpected results or some unexpected run time errors while the test is going on. Hence, proper testing and maintenance needs to be done so that the software or rather the application does not show any bug in the production. Hence what has been done is that the model is made such capable and efficient that it automatically cleans the input data and then it analyzes it and finally allocates marks for the string entered by the user or the person in the provided space. Touching down to the implementation of the algorithm or how it works, this is the section where it is very well defined. In order to understand the entire algorithm the very first thing is that one needs to understand how the test conducting system works for objective type and descriptive type questions and answers. In the case of objective questions and answers or multiple choice questions, the process being that the algorithm marks or credits those questions for which the answer is correct while the algorithm does not credit the answer if the answer is wrong. In such a case, a very simple if-else logic is implemented over a set of questions for which the data is stored in the database. But in the case of descriptive answers, the case is not so easy. It basically deals with a token match while evaluating answers at both token-to-token level and token-to-synonym level. The importance of analyzing synonyms is discussed in the above lines of this section and the pseudocode for implementing the same is also mentioned below. Below is the flowchart or rather the procedural architecture of descriptive text analysis using NLTK, Numpy & Pandas.

3. Architecture for Descriptive Text Analysis

It can be seen from the below algorithm that, how the rule and statistical based descriptive text analytics works. There are multiple important steps or rather components and filters which are discussed in this and the next section with proper mechanism in very detail. The important steps are mentioned clearly in the flowchart of the algorithm while a block diagram of the entire flow is explained below.

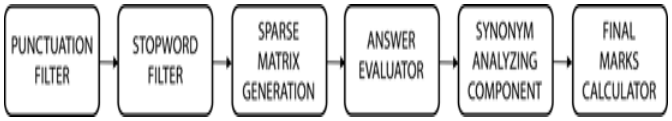

Fig. Block Diagram for the control flow for the rule and statistical based descriptive text analysis algorithm

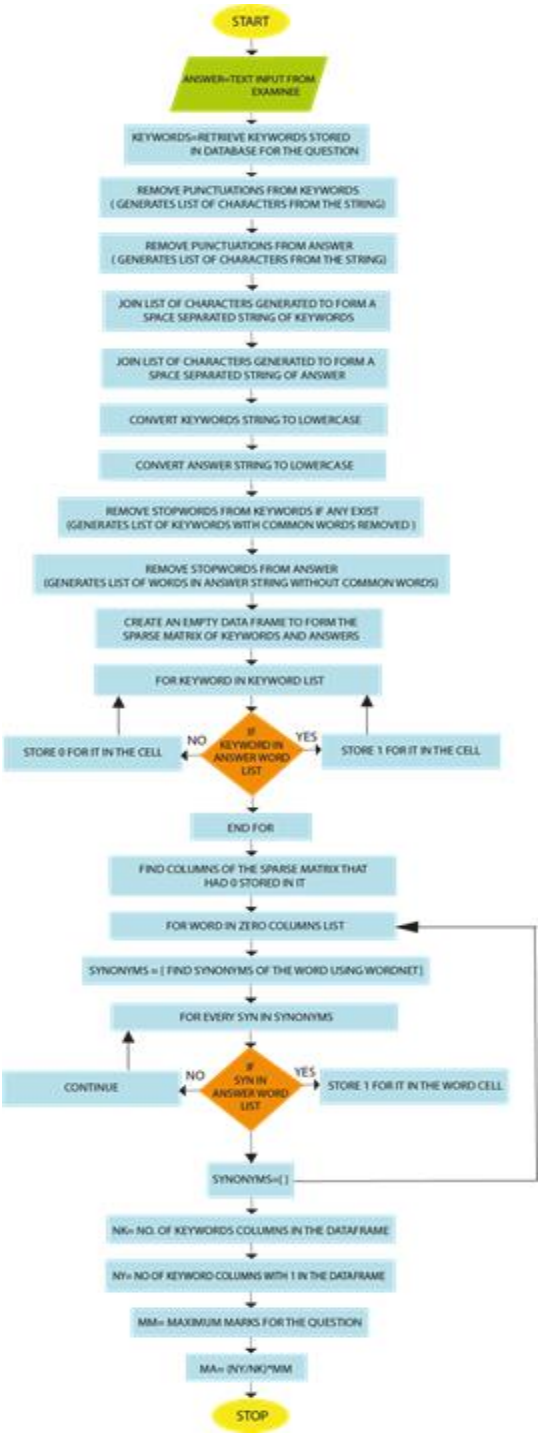

***Fig. Proposed** Model / Algorithm for Rule and Statistical Based Descriptive Text Analytics using Natural Language Processing*

4. Result and Analysis

As it can be seen below in the given figure that there are nine answers to the question which asks that “What is Machine Learning?”. Further the answers are passed via the autonomous marks allotment algorithm for the particular testing question. The answers were collected via web scraping and extracting definitions over multiple web pages. Also, in this section there will a step by step analysis of the entire algorithm as discussed above for a single answer, and then it will be discussed on how the algorithm works when there are multiple answers in a queue to be cleared.

	ID	Answer
0	1	Machine learning is an application of artifici...
1	2	Artificial intelligence (AI) has received incr...
2	3	Machine Learning is the field of study that gi...
3	4	Machine learning is the science of getting com...
4	5	Machine learning (ML) is the scientific study ...
5	6	Machine learning is a large field of study tha...
6	7	Machine learning (ML) is a category of algorit...
7	8	Well, Machine Learning is a concept which allo...
8	9	Machine learning is a hot topic in research an...

Fig. *Pandas Dataframe showing all the answers for various answer ID*

Now, it’s time to analyze just a single answer from the pool of answers. Let us assume that an answer is analyzed with ID to be 6 from the above pandas dataframe. So the very first thing that is done is that the comma plus space (“, ”) separated keywords that are stored in the database for that particular question is fetched along with the answer typed by the examinee and further it is cleaned or rather the punctuation are removed. Hence let us assume that the input to the algorithm is as follows,

keywords='Machine, learning, scientific, algorithm, statistical, models, artificial, intelligence'

(1)

answer='Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.'

(2)

Further on cleaning the above sentences, the output of the cleaning agent or rather the punctuation and stopword filter, the output turns out to be a list of important words or tokens that are filtered from the sentences which looks something like this,

cleaned_keywords = ['statistical', 'models', 'artificial', 'learning', 'algorithm', 'machine', 'intelligence', 'scientific']

(3)

cleaned_answer = ['statistical', 'allows', 'predicting', 'become', 'input', 'new', 'analysis', 'outputs', 'applications', 'outcomes', 'available', 'data', 'without', 'predict', 'learning', 'accurate', 'explicitly', 'output', 'build', 'receive', 'algorithm', 'premise', 'software', 'ml', 'programmed', 'category', 'basic', 'algorithms', 'updating', 'becomes', 'use', 'machine']

(4)

The next step that is to be applied to the cleaned_keywords and cleaned_answer is that with the keywords a fresh pandas dataframe needs to be built such that it can act as a sparse matrix and can help in evaluating the answer. Also, after the dataframe is made, a system is iterated over each word in cleaned_answer with respect to the cleaned_keywords such that it can be checked for it’s presence or not. If it is present, then the column is marked 1 while the rest columns auto fill numpy.nan i.e “Not a Number”.

	Answer	statistical	models	artificial	learning	algorithm	machine	intelligence	scientific
0	Machine learning (ML) is a category of algorit...	1	NaN	NaN	1	1	1	NaN	NaN

Fig. Pandas *Dataframe showing the evaluated answer with numpy.nan*

Then the cells in the evaluated answers dataframe, where numpy.nan is filled, gets filled with “0” completing the evaluation at sparse matrix level.

	Answer	statistical	models	artificial	learning	algorithm	machine	intelligence	scientific
0	Machine learning (ML) is a category of algorit...	1	0	0	1	1	1	0	0

Fig. *Pandas Dataframe showing the evaluated answer 0 instead of numpy.nan*

Once, all these steps are done, then the data Is to be passed via the synonym filter or analyzing component. Here in this step the corpus included Wordnet is used that consists of all the words and it’s synonyms. Hence, now only those words are to be evaluated for which there exists a zero in the columns or rather the words for which a match was not found in the answer. As any match for such words were not found in the answer, it might be so that there were synonyms for it which did not get evaluated.

zero_cols = ['models', 'artificial', 'intelligence', 'scientific']

(5)

So, for the above columns in the dataframe and in as words in the keywords list did not match with the answer words as separated out or tokenized out from the answer string. They need to be passed via the synonym analyzing component.

	Answer	statistical	models	artificial	learning	algorithm	machine	intelligence	scientific
0	Machine learning (ML) is a category of algorit...	1	0	0	1	1	1	0	0

Fig. *Pandas dataframe with words analyzed using synonym analyzing component*

Now it could be seen that after passing it via the synonym analyzing component, none of the words in the keywords list

turned out to be as a synonym that could have been present in the answer and hence there is no change for this particular use case. Hence, after doing all these operations and calculations via punctuation and stopword filter, sparse matrix generation and synonym analyzing component, it was the time of a ruled based formula using which the final marks for the answer would be calculated.

So, the formula applied as discussed in the above section is as below,

marks_allotted = (# of columns with 1 in the sparse matrix / # of keyword columns) * maximum_marks

(6)

So, according to the formula let the maximum_marks be equal to 5, then,

# of columns with 1 in the sparse matrix = 4
(7)
# of keyword columns = 8
(8)
marks_allotted = (4 / 8) * 5 = 2.5
(9)

Hence, the final marks allotted were to be 2.5 for the entered answer.

Now, on running the same algorithm for a pool of answers, the output looks something like this,

	ID	Answer	Answer Words List	machine	learning	scientific	algo
0	1	Machine learning is an application of artifici...	[machine, learning, application, artificial, i...	1.0	1.0	0.0	
1	2	Artificial intelligence (AI) has received incr...	[artificial, intelligence, ai, received, incre...	0.0	0.0	0.0	
2	3	Machine Learning is the field of study that gi...	[machine, learning, field, study, gives, compu...	1.0	1.0	0.0	
3	4	Machine learning is the science of getting com...	[machine, learning, science, getting, computer...	1.0	1.0	0.0	
4	5	Machine learning (ML) is the scientific study ...	[machine, learning, ml, scientific, study, alg...	1.0	1.0	1.0	
5	6	Machine learning is a large field of study tha...	[machine, learning, large, field, study, overl...	1.0	1.0	0.0	
6	7	Machine learning (ML) is a category of algorit...	[machine, learning, ml, category, algorithm, a...	1.0	1.0	0.0	
7	8	Well, Machine Learning is a concept which allo...	[well, machine, learning, concept, allows, mac...	1.0	1.0	0.0	
8	9	Machine learning is a hot topic in research an...	[machine, learning, hot, topic, research, indu...	1.0	1.0	0.0	

Fig. *Final pool of evaluated and marks allotted answers in a single pandas dataframe*

CONCLUSION

Hence, in the world of digitalization where mostly all the data is stored digitally, it should be such that exams should be conducted over digital media only as the cost of manufacturing paper and writing medium is very high and the machinery costs a lot. Also, with such efficient mediums or systems in the market it would get easier for educational institutions to conduct exams for a large number of students. Also, with such systems for conducting exam, a lot of time and resources can be saved over time which can be applied to some other industry where really time, money, effort and energy is much more required in comparison to digitally maintaining the data.

FUTURE WORK

Moving or flowing towards other advancements in these products or software based applications, a lot many up-gradations can be done. An efficient deep learning based algorithm can be applied such that it can track if the students are doing a fraud in the exam or not via keeping the camera on and recording video of the person appearing for the exam. This can make the system more efficient in terms of functionalities.