Powerful Blend of Big Data and Natural Language Processing

Pranjali Nikam[1], Nikita Munot[2], Amit Kadam[3], Vijendra Kadam[4]

Assistant Professor,
[1] Department of Computer Engineering, G. H. Raisoni Institute of Engineering and Technology, Wagholi, Pune, India.
[2,3,4] Department of Computer Engineering Anantrao Pawar College of Engineering & Research, Parvati, Pune, India.
pranjali.amore@gmail.com, nikita.0149@gmail.com, kadamamit1811@gmail.com, vij711@gmail.com

Article Info
Page Number: 16-21
Publication Issue:
Vol. 67 No. 1 (2018)

Article History
Article Received: *15 October 2018*
Revised: *24 November 2018*
Accepted: *18 December 2018*

Abstract

Due to advancement of technology, use of internet and computers is observed in every business and organization. Thus these business transactions generates large volume of text information. The data gathered through various sectors are in variety of languages and variety of form like images, text etc. This data should be analyzed for growth of business. Natural Language Processing (NLP) is fast becoming essential to many new business functions like compliance monitoring, BI, and analytics. Consider all the unstructured and semi-structured content that can bring significant insights – queries, email communications, social media, videos, customer reviews, support requests, etc. NLP tools and techniques help businesses process, analyze, and understand all of this data in order to operate effectively and proactively. Composed of internally stored organizational information such as customer and sales information, transaction data, research, as well as external open source information and social media, this big data is largely unstructured and in a state of constant growth. Natural language processing (NLP) of big data is the next great opportunity. The paper discusses about the use of integrating NLP and big data and its applications.

Keywords: - Cross language information retrieval, Hadoop, International data corporation IDC, Machine learning, Named Entity Recognition, Natural language processing, Part of speech tagging, Spark, Sqoop, Word sense disambiguation.

1.0 Introduction

Big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and Exabyte’s of data. It is used to describe the voluminous amount of unstructured and semi-structured data a company or organization or business generates. There are three dimensions to big data known as Volume, Variety and Velocity. The term "big data" tends to refer to the use of predictive analytic, user behavior, or certain other advanced data analytic methods that extract value from data. The data available to us is 80% raw. Big data comes from information stored in enterprises and big organization. For example information about sales record, social media feedback or review, employees, company purchase, business transaction etc. The data gathered through various sources are in unstructured form like text, images, videos, different languages. Computers are not fully developed to understand the emotions and ambiguous meaning behind the big data. Though human uses language, which is ambiguous and unstructured to be interpreted by computers, yet with the help of natural language processing, this large unstructured data can be harnessed for evolving patterns inside data to know better the information contained in data. Natural language processing is a form of artificial intelligence that helps machines to understand the text written or voice spoken by human. This makes machines as intelligent as human. It includes variety of methods such as semantics, linguistic and machine learning. Various tasks are to be done depending upon the applications such as morphological analysis, tokenization, part of speech h tagging , word sense disambiguation, summarization, categorization, named entity extraction, speech recognition, pattern recognition, machine translation, sentiment analysis, segmentation, reference resolution. NLP can solve significant problems of the business world by using big data for any business of retail, health-care, business, financial institutions, schools/colleges and private and government sectors. Hence there is a scope of research to develop the systems those are capable to understand language and process big data also. So there is a need of integration of natural language processing and big data.

2.0 Literature Review

In this section we cite the research work done earlier. Lot of research is done on NLP and big data. But the integration of both is new field of interest due to use of internet and information overload. Johanna Monti, Mario Monteleone [1] develops a methodology for development of an ontology based cross language information retrieval (CLIR). Extracting relevant information in multilingual context from massive amounts of unstructured, structured and semi-structured data is a challenging task. The outlined research activities are based on Lexicon-Grammar (LG), a method devised for natural language formalization, automatic textual analysis and parsing. Another technology developed [4] a scalable NLP and ML algorithms (classification, clustering and ranking methods) that automatically classify laws into various codes/labels, rank feature sets based on use case, and induce best structured representation of sentences for various types of computational analysis. The development of computational data science techniques in natural language processing (NLP) and machine learning (ML) algorithms to analyze large and complex textual information opens new avenues to study intricate processes, such as government regulation of financial markets, at a scale unimaginable even a few years ago.

Rodrigo Agerri, Xabier Artola , Zuhairz Beloki, German Rigau, Aitor Sarora [5] proposed a new distributed architecture and technology for scaling up text analysis running a complete chain of linguistic processors on several virtual machines. The use of Storm in a new approach for scalable distributed language processing across multiple machines and evaluate its effectiveness and efficiency when processing documents on a medium and large scale. The system has been success for improvement regarding language processing performance when adopting parallel architectures.

3.0 Need of integrating Big Data and NLP

Big data is gathered from various sources like social media activity, customers online shopping behavior, movies reviews, tweets collected, feedback from users or customers to the products/brands and point of sale data. The real time data that keeps on accumulated from online websites is stored in cloud. The data is mostly in the text form. It is characterized by 3Vs they are volume, variety and velocity at which data must be processed. There is no specific term to decide the volume of big data but it is used often for terabytes. Petabytes and Exabyte's. These data is stored for decades and generates petabytes of data or which is now known as big data. This data which is collected from various sources is in raw form means unstructured. Also data generated from all such sources are in variety of languages. This kind of big data needs natural language processing to analyze its contents.The breakdown of 3Vs of big data is discussed. The voluminous data can come from different sources, such as business sales, records, customer feedback. Data may also exist in variety of forms such as text, images, audio and video. Velocity refers to the speed at which big data must be analyzed. Business cannot boost with traditional techniques and process big data. If we train computer then they can perform tasks faster and efficiently. Only drawback is lack of understanding the language that humans speak. Computers can understand structured and unambiguous data. Researchers are working in order to train the system to understand the languages that humans use and meaning behind it.

Natural language is the sub-domain of artificial intelligence concerned with the task of developing programs having capability of understanding a natural language in order to achieve some specific goal. There is a need to design, develop and test systems that process natural language for practical applications like speech recognition, machine translation, text to speech and speech to text translation. But to design all such systems the basic requirement is to process the natural language in such a way that computers can understand it. And to process natural language there are various tasks to be performed depending upon the need of application. NLP is being as the next big research area in data analysis. Although research is going on from past few decades but still it is in developing phase.There are lot of advantages of integrating big data and NLP. Future technology can predict the human behavior by analyzing the increasing amount of unstructured data like product reviews, emails, voice calls or messages. For example, analyzing the There are lot of applications of NLP in day to day life. In medical sectors the prescription written by doctor might in English but patient if is illiterate then he or she may not understand the schedule of medicines to be taken. There is a need to translate the prescription from English into their regional language. Also if proper analysis is done on patient information then early treatment and diagnosis of disease is possible at earlier stage. The health-care information are usually handwritten by doctors. So data can be categorized and saved into database which can be useful for data mining purpose. We use are smartphone while traveling for navigation. Users provide the address by speaking near mic and the address is searched. This is an application of voice recognition. Identifying the names of people in Facebook posts.

NLP is taking over its place in all applications. India still is a country will literacy rate is low. Experts are working on cross language information retrieval system where user can query in any language and receive the result back in the same language with help of machine translation. There are lot of open source tools available for performing NLP tasks.Big data revolution has been boosted up for handling unstructured data. The technologies like Hadoop and Spark makes it possible to process huge volumes of unstructured data. Organizations are just beginning to understand the enormous potential value stored in all the text we generate on a daily basis, in the form of emails, text messages, social media posts, search queries, medical and legal records, customer's online shopping behavior, social media activity, and internal data logging like point-of-sales data and more. In other words, we can say that big data is in natural language but in the form of strings or words which a normal human being uses in their day to day life which may include slang terms, emotions. Pharmaceutical company produces large volume of clinical data like doctor information, patient history and personal information. These types of information are largely made up of language, natural language processing of big data invokes an opportunity to discover the current trends in market.

Natural language processing for big data can be leveraged to automatically find relevant information and summarize the data gathered from all sources for analysis. Shopping trends today is growing towards e-commerce media where user can read reviews and also provide a feedback for the products. Sentiment analysis can help the organizations to understand what is being said about their brands and products. Also we can discover the customer's preference, pattern of shopping, habits, opinion and predict what does the user needs in future. This information can then be applied to product development, business intelligence and market research.NLP studies the patterns emerging in the text entries in the big data by analyzing the linguistics and semantics through statistics and machine learning and extracts the significant entities and relationships in the context of what the customers are trying to say in their posts. It focuses on sentences as a whole first and then discover about the words or character. In every domain, the data received is in the form of text or set of documents. For example medical, education, offline or online shopping, legal, government and private organization all such applications produces text information. Hence NLP becomes a necessity for analysis of data.

4.0 NLP Tasks and Applications

Following are few of the NLP applications and the tasks needed to perform based upon the requirement. Basic tasks to process natural language are tokenization, part of speech tagging, morphological analysis and stemming, part of speech tagging, named entity recognition, reference resolution.Tokenization is a process of breaking a stream of sentences into tokens. This is done by searching a space after each word. All the generated tokens are saved in a separate file for further processing. End point of a word and beginning of the next word is called as word boundary. It is also called as word segmentation [6]. E.g. what is your name? Expected Output for this should be [What] [is] [your] [name] [?] are recognized ad separate tokens and saved. Morphological analysis is a study of internal structure of words. Finding root words by removing suffixes and inflection of the words is an important task for information retrieval scheme. E.g. Drinking, Drank, Drunk. Expected output is "Drink". Part-of-speech tagging deals with designating any given sentence its appropriate part of speech tag word such as noun, verb, adjective, etc. Whole sentence is parsed to describe each words syntactic meaning and generate POS tag for same. For E.g. “John is playing cricket”, if this sentence applied as input to POS tagging software should produce the output as: John –> Noun, is –> verb, playing –> verb, cricket –> proper noun.Named entity recognition (NER) classifies each given word from the sentence to predefined set of classes such as name of things or person name or organization name into “Person” category. E.g. let’s meet Bob at 6 p.m. in India. Expected output should be Bob –> Name, 6 p.m. –> Time, India –> Location.Sentence segmentation identifies sentence boundaries in the given text, i.e., where one sentence ends and where another sentence begins. Sentences are often marked ended with punctuation mark ‘.’ or special mark ‘? ‘.

In language processing, reference resolution occurs when two or more words, expression in a sentence refer to the same person or thing; means they have same referent. It is about defining the relationship of given the word in a sentence with a previous and the next sentence. For E.g. “John said he is playing cricket so he will be late.” Here the word “John” which is noun and pronoun “he” which is repeated twice in the sentence refers to the same person John.Some of the application of NLP are information extraction, machine translation, spam detection, opinion mining, summarization, question answering and lot many to list. Some of the applications are discussed below.Stock market and trading these days has become fully automatic. Algorithmic trading is a form of financial investing that is controlled by technology. The share market depends on the current news going in English language. It’s the job of NLP to take these stock market news as input and extract the information and make decisions.

Web is loaded with information and is getting overloaded with data each passing day. There needs to be efficient development in the area of information retrieval so the required data can be fetched accurately and efficiently. Web NLP [3] is a system for information retrieval from the webpages using Natural Language Processing (NLP).Machine translation is another application where the meaning of the text should be preserved after translating. We have online Google translator which translates text to many different languages. But still that translator are having drawback. It gives wrong outputs for many inputs. Also Facebook has a feature of translate link.Text categorization helps to fight against spam mails which classifies the spam and not spam mails. Google and other companies provide such service. Text summarization [2] accept the input document and generate a small summary. Due to the use of internet there is lot of information overload. Sentiment analysis is done to analyse the movie reviews are positive or negative.

Sentiment analysis is routinely used by social analytics companies to put numbers behind the feelings expressed on social media or the web in order to generate actionable insights. Marketers use sentiment analysis to inform brand strategies, while customer service and product departments can use it to identify bugs, product enhancements, and possible new features. Marketers are using NLP for sentiment analysis, combining millions of tweets and other social media messages to determine how users feel about a particular product or service. It has the potential to turn all of Twitter or Facebook into one giant focus group. Question answering is an application used by search engines which provides real time responses to customer.Big data techniques used to store and analyse data are Apache Hadoop, Hive, Spark, Sqoop etc. Hadoop is an open source java based programming framework which supports processing and storage of large data sets in a distributed environment. Developed by Apache Software Foundation. It can scale up from single server to 1000 of server machines.Apache spark is a cluster computing framework. It can use hadoop distributed file system. It is a fast engine for big data processing capable of supporting SQL, machine learning. Apache hive is a data warehouse for big data. Whereas sqoop is used for transferring bulk data between hadoop and structured data stores such as relational databases. It can also extract data from hadoop and export it into the structured database. Using such techniques like hadoop and spark, the data stored from variety of sources can be used for natural language processing.

5.0 Conclusion

IDC estimated that by 2020, 44 trillion gigabytes of digital data will be created and this data will be converted to knowledge with the help of NLP. No matter where you apply it, natural language processing for big data will be an essential feature to build into your analysis pipeline in order to capture the value of this information for insight, reduced costs and increased productivity. Hence there is a need to integrate big data and NLP.

6.0 Acknowledgement

We would like to express our sincere gratitude Mr. Atul Marathe (Adjunct Professor), Dr. Ashok Saraf (Innovation Club Member) for their invaluable guidance and support throughout the research process. We also wish to thank Dr. Sunil B. Thakare (Principal, APCOER, Pune) for their support. Finally, we are grateful to all of the research participants who generously gave their time and effort to this project..

7.0 References

[1] Johanna Monti, Mario Monteleone, Maria Pia Di Buono, “Natural Language Processing and Big Data - An Ontology-Based Approach for Cross-Lingual Information Retrieval”, IEEE 2013 International Conference.

[2] Nikita Munot, Sharvari S Govilkar, “Conceptual Framework for Abstractive Text Summarization”, March 2014 International Journal on Natural Language Computing (IJNLC) Vol. 4, No. 1, February 2015, DOI: 10.5121/ijnlc.2015.2014.

[3] Rini John, Sharvari S. Govilkar “A Novel approach for information retrieval technique for web using NLP” ,International Journal on Natural Language Computing (IJNLC) Vol. 6, No.1, February 2017.

[4] Dhabliya, M. D. (2018). A Scientific Approach and Data Analysis of Chemicals used in Packed Juices. Forest Chemicals Review, 01–05.

[5] Sharyn O'Halloran. Sameer Maskey, Geraldine McAllister, “Big data and the regulation of financial markets” , IEEE / ACM 2015 International Conference.

[6] Rodrigo Agerri, Xabier Artola , Zuhairz Beloki, German Rigau, Aitor Sarora “Big data for Natural Language Processing”, ACM Volume 79 Issue C, May 2015 Pages 36-42.