Vol.05 Issue-02, (February, 2017) ISSN: 2321-1776
International Journal in IT & Engineering (Impact Factor- 6.341)

AN OVERVIEW OF NATURAL LANGUAGE PROCESSING AND BIG DATA ANALYTICS

[1]Nitasha Varshney, [2]Dr. ManojKumar

Department of Computer Science and Engineering

[1,2]Shri Venkateshwara University,Gajraula (Uttar Pradesh) – India

ABSTRACT

Natural language Processing (NLP) created because of yet a third issue displayed by big data. A significant part of the information that is customarily critical in capital markets is unstructured, which means it is arranged and intended for people, not PCs, for example, Management's Discussion and Analysis (MD&A) revelations, money related references, and oral divulgences. Applying machine-learning systems to talked and composed language, NLP calculations process these bits, and different sources, and figure out how to peruse and translate language calculations would now be able to naturally supplement organized budgetary revelations from Securities and Exchange Commission (SEC) filings with information from the literary divulgences without an expert really perusing the content and physically changing a model. What's more, as NLP has created, calculations have progressed from simple content recovery to programmed arrangement and point modeling, with the end goal that NLP calculations would now be able to recover filings, monetary reports, public statements, news and so forth. In the following research paper, we will study about the scope and role of natural language processing in Big Data Analytics.

1. INTRODUCTION

The main modern insurgency endeavored to make machine's that could supplant man's physical power. Truth be told, we have machines that can beat human creatures. Industrialization has changed the society absolutely and brought prompt emergency, later improvement. Artificial Intelligence is endeavoring to make machines that can supplant man's psychological power and accordingly may change the society, might be past creative energy. The improvements in Artificial Intelligence will directly affect the advancement of any country.The meaning of the term 'Artificial Intelligence' is a profoundly questionable one. In spite of the fact that, there is a general consent to the importance of artificial as something 'man made' or which is 'not natural', there is no accord with regards to the significance of the term intelligence.

1.1 Developments in AI

In spite of the fact that AI is generally another subject, a great part of the work which later established the framework of AI can be followed back to the most recent century. Principal among them, is crafted by George Boole (1815-1804) on Boolean variable based math in which he presented the coherent meanings of 'and', 'or' and so forth. Another vital commitment was made by Alan Turing (1912-1954), considered as the dad of AI.The advancements in the field of linguistics, particularly the

commitment of Noam Chomsky in formal syntaxes served to a substantial degree the improvements in Natural Language Processing, a branch of Artificial Intelligence. In any case, many think about 1956 as a point of interest ever of.

2. NATURAL LANGUAGE PROCESSING

The zone that endeavors to influence computers to comprehend natural language was called Automatic Language Processing. Be that as it may, the term Natural Language Processing (NLP) is at introducing generally utilized. Natural language processing (NLP) is a sort of artificial intelligence that empowers machines "to scrutinize" message by imitating the human ability to comprehend language. NLP strategies consolidate an assortment of techniques, including linguistics, semantics, experiences and machine figuring out how to remove elements, connections and comprehend setting, which empowers a comprehension of what's being said or created, widely. Instead of understanding single words or mixes of them, NLP empowers PCs to comprehend sentences as they are talked or created by a human. It uses different methods to disentangle ambiguities in language, including programmed rundown, grammatical feature labeling, disambiguation, substance extraction and relations extraction, and what are more disambiguation and natural language comprehension and recognition [1].

Prior endeavors, the original of NLP work, in the 50's in the programmed language processing finished in dissatisfaction as they had the most aggressive objective i.e. the Machine Translation (MT) Presently it is very much understood that natural language is one the most complex ancient rarities of human personality. The present day look into is for the most part with constrained objectives, and attempts to discover arrangements at various levels viz. morphological, lexical, syntactic, semantic and pragmatic.

2.1 NLP and Information Retrieval

The connection amongst linguistics and Information Science has been a controversial one. Various researchers among them Lancaster(9) and Salton and McGill(lO), noted in the 1970s and mid 2000s that there had been a push toward improvement in information retrieval and that there might in certainty be no requirement for the level of detail and modernity gave by semantic analysis [2]. In any case, the survey writing gives the feeling that there is an extensive shared opinion between etymological hypothesis and information science and this is additionally upheld by the restored enthusiasm for linguistics as a major aspect of natural language processing application in information retrieval.

2.2 Automatic Indexing

The endeavors in automatic watchword recognizable proof can be extensively assembled into two methodologies. One is utilizing statistical techniques, started by H. P. Luhn and the other approach is utilizing NLP techniques. The real issue with absolutely statistical strategies is that they don't make any endeavor at semantic analysis of the idea substance of a given document. In natural language, we ordinarily utilize anaphora which incorporates every one of the pronouns

and words like 'former', 'latter', 'first', 'second' and so on. As anaphora refer to one of the things expressed before (for the most part watchwords), it adds up to a specific catchphrase seeming more than once. Be that as it may, as the statistical techniques don't have any mechanism to supplant anaphora by the watchword it refers to, they increment the tally of anaphora occurrence as opposed to the occurrence of the catchphrases [3].

3. MACHINE LEARNING TECHNIQUES APPLIED TO NLP

Machine learning is the strategy of procuring knowledge from a situation in a computational manner. Diverse examinations over the recent decades have seen the mixing of ML and NLP turn out to be progressively normal. The accessibility of huge corpora, powerful computing assets and a more noteworthy demand for NL based applications, made ML extremely critical. The announced works converge to various territories which include: Machine Learning Technology (Symbolic Methods and Statistical), Computational Models (Logic allegation, Learning, reasoning and assessment), NLP tasks (Stemming, Parsing, POSTagging, and WSD), Data Driven Technology (IR, IE, and NLIDB), and data Mining (Knowledge Discovery, Clustering). NL content processing is normally referd to Data Intensive Approach (DIA) or Corpus Based Approach (CBA). Various ML strategies and techniques are situated in to NLP issue like grammatical feature (POS) labeling, Word Sense Disambiguation (WSD), Propositional Phrase Attachment Disambiguation (PP-AD), Automatic Text Summarization (ATS), Grammatical Inference (GI), Structural Parsing (STP), Information Extraction (IE), Information Retrieval (IR) and Machine Translation (MT). The statistical methodologies in computational linguistics can likewise be fused with the demonstrating of NL [4].

4. SCOPE OF NATURAL LANGUAGE PROCESSING

Theoretical linguists are basically inspired by creating an auxiliary teach of natural language. They are not intrigued on parsing or language age from basic depictions. A noteworthy target of theoretical linguists is to create hypotheses that hold great crosswise over languages. As it were, they endeavor to portray the general sorting out principles that underlie every single human language and don't as a rule think looking at a specific language. Clinicians then again are occupied with the way that people really create and appreciate natural language. A semantic theory, for them is just helpful to the degree that it clarifies real behavior. Thusly, Psycholinguists are occupied with both the portrayals of etymologist's structures and the procedures by which a man can create such structures from real sentences [5]. The essential instrument that is utilized is — experimentation in which real estimations are made from people as they create and understand language, including how much time a man needs to peruse each word in a sentence, how much time a man needs to choose whether a given thing is a legitimate word or not, what kinds of mistakes people make as they perform different phonetic tasks and so on. Exploratory data is utilized to approve or dismiss a particular speculation about language, which are often taken from the hypotheses that linguists and

computational linguists propose. Natural language understanding requires knowledge of how the words are shaped, how the words thus frame clauses and sentences. Likewise, to effectively understand an arrangement of sentences in a given setting, it ought to have larger amount knowledge.

When all is said in did, the knowledge that will be utilized as a part of natural language understanding is separated in to the accompanying:

- MORPHOLOGICAL: These arrangements with the morphological structure of words, similar to the word root, prefix, postfix and infixes. The fundamental unit in a composed word is a morpheme. In this way, this level gives knowledge of word arrangement.
- LEXICAL: This level manages thesaurus look into, spell redresses, acronyms and abbreviations and so on.
- SYNTACTIC: Syntax manages the structure and validity of input sentences, how a correct mix of words in a specific arrangement constitutes a legitimate sentence.
- SEMANTICS: Semantics manages the significance of words and that of sentences.
- PRAGMATICS: Pragmatic level manages sentences in a specific setting. This requires a larger amount knowledge which identifies with the employments of sentences in various settings [6].
- WORLD KNOWLEDGE: keeping in mind the end goal to do compelling communication, both the communicator and the communication ought to have foundation knowledge either to send or to get a message with no clamor. This back ground knowledge is considered as the world knowledge of a specific area.

4.1 Natural Language Interfaces Systems

Any search engine (IR system) acknowledges a client ask for through an interface. It is to be utilized by the search engine and converted into questions. In its most regular shape, the translation yields an arrangement of catchphrase (or record terms) which abridges the portrayal of the questions require. A data retrieval system recovers objects which fulfill unmistakably characterized conditions, for example, Regular expressions (RE) and Relational Algebraic (RA) Expressions NL from a structured constant stockpiling (Database). IR system dependably manages content, which is dependably non-structured and could be semantically ambiguous. This investigation distinguishes the complexities of client association with the system by tolerating a NL query. The client is presenting a NL query to a QP system which will be handled utilizing NLP techniques to acquire keywords [7]. In the following stage these keywords are submitted to the system as query to the IR system which

will recover positioned significant documents. On account of structured database NL query will be made an interpretation of in to RE or RA articulation which can be effortlessly moved in to a SQL query. This will recover the output in a relational model.

4.2 Text Query Processing through Various Levels on NLP

Processing of textual information accessible in electronic shape and recovering information shrewdly because of clients queries has developed as one of the considerable test in IR systems. There are a few related levels of analysis for NLP, which is communicated in a synchronized model. The levels of language processing at which subjective linguistics estimate humans understand or extricate meaning. The accompanying clarification will give an unmistakable picture about the significance dispersing in a content lump and the different NLP techniques required in relating levels. The agent indented to ponder the possibility of NLP.Human languages display surprising comparative patterns or principles are called universals. This can be can be extricated from different semantic levels. Phonological universals: for instance, consonants are recognized by the area of their generation, which is based on the different organs of the vocal tract. It gives itemized information one would now be able to refer to each consonant by its area and manner of explanation, for instance, is a voiceless, labiodentals fricative [8].

5.BIG DATA ANALYTICS AND NLP

Big Data is planned as nonspecific stage to determine the issues of volume, speed, assortment, veracity and incentive in data analytics (IBM, 2012). The data is gathered from various sources, for instance, daily logs, social media, and business transactions. Big Data demands the capacity to store a lot of data. With cutting edge stockpiling advances, the data size could be as high as terabytes (1012 bytes), petabytes (1015 bytes) and exabytes (1018 bytes). Moreover, research on NLP often covers with research on or the utilization of Big Data platforms. Big Data handles a wide range of arrangements which are structured and unstructured. The structured data is lucid and very much composed that is normally put away in customary relationship databases. Unstructured data does not have a predefined arrange. This type of data can be found in, for instance, emails, images, and videos. Big Data is broadly utilized as a part of business gauges, logical research, and analysis of social issues, human services, and meteorology. While advancing advances in NLP, it is additionally essential to know about moral issues around the potential misuse and double utilization of big data and NLP tools. A portion of these issues can be extremely intriguing for information innovation understudies to discuss and take in more about. Big data are pointless in a vacuum. Its potential regard is opened exactly when used to drive fundamental initiative. To empower such affirmation based fundamental administration, organizations require gainful methodology to turn high volumes of speedy moving and grouped data into imperative encounters [9].

Big Data Analytics refers to the way toward group, sorting out, dissecting vast

data sets to find distinctive patterns and other valuable information. Big data analytics is an arrangement of advancements and techniques that require new types of incorporation to reveal vast hidden qualities from extensive datasets that are not quite the same as the typical ones, more mind boggling, and of a substantial huge scale. It essentially focuses on taking care of new problems or old problems in better and successful ways. The fundamental objective of the big data scientific is to assist association with making better business choice, future prediction, analysis extensive quantities of transactions that done in association and refresh the type of data that association is utilized. Case of big data Analytics are big online business site like Flipkart, snapdeal utilizes Facebook or Gmail data to see the client information or behavior. Investigating big data permits investigators, researchers, and business clients to settle on better and quicker choices utilizing data that was already out of reach or unusable. Utilizing progressed analytics techniques, for example, content analytics, machine learning, predictive analytics, data mining, measurements, and natural language processing, businesses can break down already undiscovered data sources autonomous or together with their current endeavor data to increase new bits of knowledge bringing about altogether better and speedier choices [10]. It causes us to reveal hidden patterns, obscure connections, market patterns, client inclinations and so on. It drives us to more viable marketing, income openings, better client benefit and so forth. Big Data can be analyzed through predictive analytics, content analytics, statistical analytics and data mining.

6. CONCLUSION

NLP is an algorithm that can help secure your protection giving understanding into the blogs and posts. Nonetheless, it isn't intended to supplant human instinct. In social media conditions, NLP slices through commotion and concentrate immense measures of information to help comprehend customer recognition, and hence, to decide the most key reaction. The difficulties to creating helpful utilizations of substance examination of Blogs/Posts, particularly inside the setting of automated investigations, are substantial. Be that as it may, the benefits of an investigation are considerably more substantial, incorporating more prominent trust in learning and the ability to foresee future outcomes. Due to the inherent complexity of natural languages, many natural language tasks are ill-posed for mathematically precise algorithmic solutions. To circumvent this problem, statistical machine learning approaches are used for NLP tasks. The emergence of Big Data enables a new paradigm for solving NLP problems — managing the complexity of the problem domain by harnessing the power of data for building high quality models.

REFERENCES

[1] Natural Language Processing." Natural Language Processing RSS. N.p., n.d. Web. 23 Mar. 2017.

[2] "Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication

Therapy Management Research." AMIA ... Annual Symposium proceedings. AMIA Symposium. U.S. National Library of Medicine, n.d. Web. 19 Mar. 2017

[3] [Elkan C. Log-Linear Models and Conditional Random Fields. 2008. http://cseweb. ucsd.edu/welkan/250B/cikmtutorial .pdf (accessed 28 Jun 2011). 62. Hearst MA, Dumais ST, Osman E, et al. Support vector machines]

[4] [Srihari S. Machine Learning: Generative and Discriminative Models. 2010. http:// www.cedar.buffalo.edu/wsrihari/C SE574/Discriminative-Generative.pdf (accessed 31 May 2011).]

[5] Ahonen, H., Heinonen, O., Klemettinen, M., &Verkamo, A. I. (1998, April). Applying data mining techniques for descriptive phrase extraction in digital document collections. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on (pp. 2-11). IEEE.

[6] Allen, N. and A. Lascarides. 2007. Intentions and information in discourse. In Proceedings of the 32nd annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 34-41.

[7] Alshawi, H. (1992). The core language engine. MIT press.

[8] Anderson, B. and Moore, A. Active learning for hidden markov models: Objective functions and algorithms. In Proceedings of the International Conference on Machine Learning (ICML), pages 9-16 (2005).

[9] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., &Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.

[10] Angluin, D. Queries and concept learning. Machine Learning, 2(4):319-342 (2008)