**Semantic Analysis On Email**

Asfina Vayani

Computer and Information Science

University of Massachusetts , Dartmouth

North Dartmouth, MA, USA

asfina.vayani@umassd.edu

**Dataset:** **Personal Gmail Dataset**

***Abstract:* Due to the rapid evaluation of the technology in every field, data becoming the more valuable and important thing by the time despite of the fact that in which shape and category the data fits in, whether it is image dataset, numerical dataset, tabular dataset, textual dataset etc. However with the early rise of machine learning, these data has given some sort of power to the developers that no one has thought of before. Using machine learning algorithms, we can develop such features and functionality that does not need any conditioning, it is all based on state of art algorithms. Here, we are trying semantic analysis on email dataset. The main aim for us is to find the similarities between the emails and their context through various NLP pipelines and Latent Semantic Analysis.**

***Keywords***

Semantic Analysis, Unsupervised Learning, Natural Language Processing, Text Data, K-means Clustering, Email Dataset

1. INTRODUCTION

The advancement of calculations that empower computers to consequently handle text and characteristic language has consistently been one of the extraordinary difficulties in Artificial Intelligence. As of late, this exploration heading has progressively picked up significance, last not least because of the approach of the World Wide Web, which has intensified the requirement for keen content and language handling. The interest for PC frameworks that oversee, channel, and search through colossal vaults of text documents has made a totally different industry, as has the interest for shrewd and customized interfaces. Thus, any generous advancement in this space will strongly affect various applications going from data recovery, data sifting, and astute operators, to discourse recognition, machine interpretation, and human-machine collaboration.

The fundamental test of a machine learning system to address establishes in the qualification between the lexical level of "what really has been said or expressed" and the semantical level of "what was expected" or "what was alluded to" in a book or an expression. The subsequent issues are twofold: (I) polysemous, i.e., a word may have various faculties and numerous kinds of use in an alternate setting, and (ii) equivalents and semantically related words, i.e., various words may have comparative importance, they may at any rate in specific settings signify a similar idea or - from a more fragile perspective - allude to a similar point.

Semantic analysis is derived from the WordNet database where each term is associated with each other. This database is of English words which are linked together. If two words are close to each other, they are semantically similar. More specifically, we are able to determine synonyms like similarity. We map terms and examine their relationship in the ontology. The key task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences. Thus it is helpful to show the polarity of the sentiment for the users.

For example, in the sentence”I am happy” the word ‘’happy’’ being an adjective gets selected and is compared with the stored feature vector for synonyms.

Latent semantic investigation (LSA) [1] is a notable method that partially addresses these inquiries. The key thought is to plan high-dimensional feature vectors, for example, the ones emerging in vector space representations of text documents [2], to a lower-dimensional portrayal in a so-called latent semantic space. As The name recommends the objective of LSA is to discover an information planning that gives data well past the lexical level and uncovers semantical relations between the elements of interest. Because of its over-simplification, LSA has demonstrated to be an important investigation device with a wide scope of uses (for example (3, 5, 8, 1]). However, its hypothetical establishment stays to an enormous degree unacceptable and fragmented.

In this paper, we will evaluate Latent Semantic Analysis techniques. Latent Semantic Analysis (LSA) is a hypothesis and technique for removing and speaking to the logical utilization importance of words by factual calculations applied to a huge corpus of text. LSA is a data recovery strategy that investigates and recognizes the example in an unstructured assortment of text and the connection between them. LSA itself is an unsupervised method of revealing synonyms in an assortment of documents.

Here, we are trying to extract textual data from our personal email id through gmail API. After the extraction of email data, we will do the pre-processing on our textual features such as subject and content. This preprocessing thing involved NLP on our textual dataset to convert the string format data into numerical format so that when we feed our data into machine learning algorithms then they can understand. After the pre-processing, we will get the feature matrix of our dataset through the TF-IDF method and then evaluate that on unsupervised algorithm techniques.

1. LITERATURE SURVEY
2. METHODOLOGIES

IV DATASETS

V EXPERIMENTS AND RESULTS

VI DISCUSSIONS

VIII CONCLUSION

ACKNOWLEDGEMENTS

REFERENCES

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, Landauer. T. K., and R. Harshman. Indexing by latent semantic analysis. Journal of the American

Society for Information Science, 41, 1990.

[2] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.