International Journal of Innovative Research in Engineering
Volume 5, Issue 2 (March-April 2024), PP: 315-318.
https://www.doi.org/10.59256/ijire.20240502043
www.theijire.com
ISSN No: 2582-8746

Prediction of Certain examinations on AI procedures for Text rundown

Sivakumar Nagarajan
Technical Architect, I & I Software Inc, 2571 Baglyos Circle, Suite B-32, Bethlehem, PA-18020, USA.

How to cite this paper:
Sivakumar Nagarajan, "Prediction of Certain examinations on AI procedures for Text rundown", IJIRE-V5I02-315-318.

Abstract: *Making text synopses from huge volume of unstructured text like client surveys, web log posts and online entertainment posts is a significant assignment in text mining applications. These synopses uncover the valuable data which depicts the whole report or surveys. Rundown errand could be performed utilizing two significant methodologies: Extractive and abstractive methodology.*

Extractive methodology distinguishes the huge piece of the report that uncovered the whole happy and removes it to shape synopsis. Abstractive approach makes synopsis in light of catchphrases and semantics from the archive, which makes it troublesome when contrasted with the other methodology. The extractive synopsis task is perplexing because of overt repetitiveness, enormous volume of text, changeability and semantics of regular language in the text. The wide pertinence as well as trying nature of the errand has enlivened dynamic research in the space by both scholar and industry specialists. This exploration centers around the plan of AI based frameworks for two center regions connected with text mining, specifically, Component based text synopsis and text comparability location. Existing element positioning and outline frameworks utilize an assortment of strategies including inactive semantic ordering; Innocent Bayes' and other semantics based approaches. Because of the intricacy of the undertaking there is a requirement for creating proficient frameworks. In this proposal, an element positioning framework in light of client inclinations have been created. Three different AI approaches have been embraced for highlight based miniature level extractive text synopsis development.

I.INTRODUCTION

Aspects are significant features in customer reviews that are essentialfor analysing the reviews to serve in various business decisions. These reviewsare often unstructured and do not imply any meaning in the text. Summarizingthe core characteristics of these aspects becomes important for commercial purposes. Both aspect ranking and summarization had been accomplished by different machine learning techniques. This research focuses on utilizing customer preferences for ranking the aspects from the reviews in order to improve the business decisions from the stake holders. This facilitates to improve aspect ranking from the customers point of view. Summarization had been explored using machine learning and optimization with parallel and large scale analytic strategies. This enables to process large volume of customer reviews and improve the quality of the summary generated for the aspects. Employing machine learning and parallel algorithms will enable to improve the quality of text summarization systems. Current researches in the area of text mining deals the problems of text representation, classification, clustering, text summarization and modeling of hidden patterns.

Text mining is an area where large amount of unstructured text is analysed to gain some actionable intuitions. Natural language processing and text mining could be viewed as artificial intelligence technologies to enable transforming key contents from large text to quantitative information. This could be used for further analysis and would help in business processes. Alternate terms for text mining are text data mining and text analytics. In the present age, unstructured text is found in huge volume with internet as key source of data. Most of these unstructured data is generated from millions of customer reviews Organizations would be able to make better decisions when these reviews are quantitatively analysed. Identification of key features and summarizing the key content into meaningful form are the two thriving factors for the wide application of text mining. These are accomplished by feature extraction and text summarization. When opinions in the text were added to these tasks, they gain deep insights in to operational challenges faced by the prospective customers. There is a need to enhance the efficiency of automatic text summarization systems and to handle huge volume of data generated from the web today.

II. LITERATURE REVIEW

Aspect ranking framework is significant to identify the important aspects from numerous consumer reviews posted in various domains like hotel, movie and product etc. They could be broadly classified into supervised and unsupervised approaches broadly. Supervised methods rely on semantic knowledge bases. These are found to be effective for ranking compared to conventional approaches. These methods available in the literature are discussed in detail. Next, this review focuses on the extractive summarization systems, in which the summary is generated by picking a sub-set of sentences from the related text. Extractive summarization systems that utilize machine learning, optimization and map reduce framework are explained elaborately. This is due to the efficiency of these techniques reported in the comprehensive works available for text summarization. A literature covering text similarity discovery methods employing text, semantic information and graph based systems are presented in detail at the end of the chapter. Among these graph based methods play a vital role in current field of the research.

III. ASPECT BASED RANKING FOR CUSTOMER REVIEWS USING ONTOLOGY

In today's internet world, customer reviews are increasingly posted for any service or product. Most of the customers read these reviews before the opt for a product or a service. They believe in online reviews as much as personal recommendations. These reviews serve as a good source of information foridentifying important features or aspects with respect to a product or a service.

Any service or product may have numero us aspects. Some of them can strongly influence the service based on the choice of the customers. So identification of features and their ranking becomes essential to determine the prominent aspects or features. The goal of aspect ranking systems is to extract and identify the significant aspects as well as to rank them from customer reviews. Customer review analysis becomes difficult since there are language based lexical issues such as words with multiple meanings, slang words etc. within the same domain.

These reviews from different customers vary in syntax, structure and opinions. Every reviewer or author writing reviews have different opinions about a product or a service. These opinions are helpful in capturing the choice or preference of the reviewers. Inclusion of these opinions or preferences of authors would be most useful in feature ranking systems. This would clearly demonstrate the choice of authors for any domain in which their reviews are posted. Ranking of features based on the choice of the authors would enable the feature ranking systems to give precise ranking depending on the customers who add value for decision making.

IV. PROPOSED REVIEW RANKING USING ASPECT BASED PREFERENCE AGGREGATION APPROACH

The proposed aspect ranking system uses customer reviews as inputdata. This data is pre-processed for noise removal and features are extractedusing dependency parsing algorithm. Then domain specific ontology is constructedfor reviews and scores based on author preferences, term frequencies for all thereviews are computed. These scores are annotated in the ontology and pairwise ranking algorithm is used for ranking the features. A new feature score formula including significant score and author preference score has been employed for scoring. This is effective in ranking as it includes customer preferences. These scores are annotated in the ontology and pairwise ranking algorithm is used for ranking the features. These features will be very helpful for analysis as they are most preferred by customers. So they can be used for generating micro level summaries.

Architecture of the Feature Ranking System

The overall Architecture of the proposed system is shown in Figure 3.1. There are three main stages designed in the system. They are pre-processing, feature extraction and feature ranking.

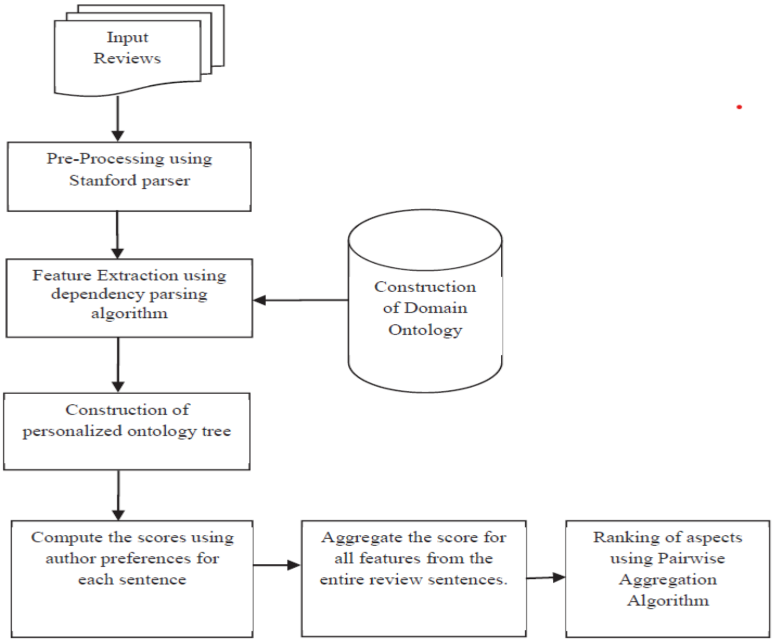

Input Reviews
Pre-Processing using Stanford parser
Feature Extraction using dependency parsing algorithm
Construction of Domain Ontology
Construction of personalized ontology tree
Compute the scores using author preferences for each sentence
Aggregate the score for all features from the entire review sentences.
Ranking of aspects using Pairwise Aggregation Algorithm

Figure 3.1 Overall Architecture of the Proposed Review Ranking System

Pre-processing

Pre-processing is an essential task in text analytics. This is a process which transforms the original data into the data required for another processing format. The study about the implications and the importance of pre-processing techniques are found to be little among researchers. Nevertheless the choices of various pre-processing techniques are found to have major impact in the results obtained after the analysis. Text can be represented efficiently in terms of both time and space for improving the performance. The two methods used for pre-processing in the feature ranking system are detailed in the next section.

Stop word Elimination

Stop words are words that are not meaningful as single words. Reviews collected from movie domain and hotel domains are used for stop wordremoval and lemmatization. Stop words are considered from nltk (naturallanguage tool kit) library and some common words considered for pre-processingare: is, was. The list of stop words used for pre-processing has been providedin.

Consider an example sentence “The restaurant down stairs is always busy and has nice toasted sandwiches and cakes they call it the German bakery”from the hotel domain review.When subjected to stop word removal using the nltk library, result in the sentence given below, after removing the stop words, “restaurant down stairs always busy nice toasted sandwiches cakes they call German bakery”, then lemmatization is performed.

Lemmatization

Lemmatization is an approach to eliminate inflections in the natural language text. Inflections in a natural language refer to morphological variants of a verb. These variants are mapped to a same base word or root word using a stemming algorithm. Thus Lemmatization process reduces the words with same meaning to its correct base forms of words using Stanford stemming algorithm . Stanford NLP toolkit is used for lemmatization.

Feature Extraction

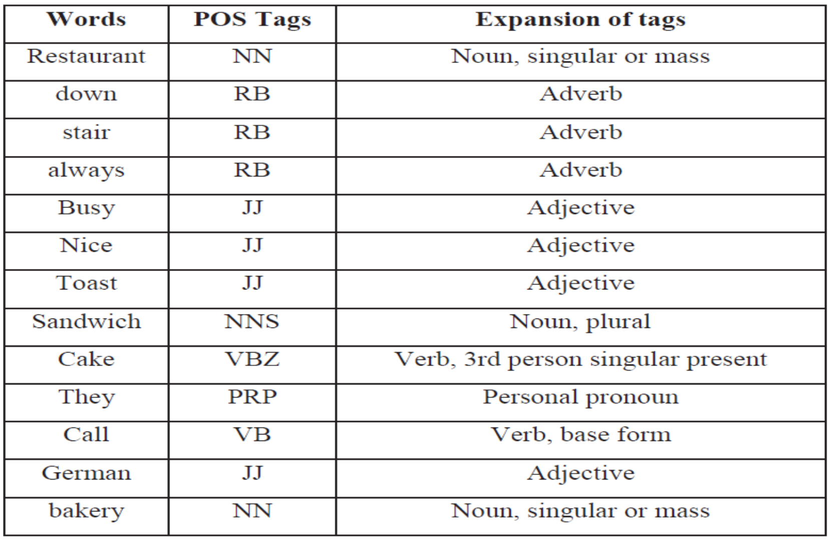
The pre-processed sentences are considered for feature extraction. These review sentences are split up and then each sentence is parsed using the Stanford parser

Sentence Parsing

Stanford parser uses natural language processing to parse a sentence into parts of speech tagging. It identifies the different parts of the sentence into nouns, verbs and adjectives. The process of categorizing words into their parts of speech and labelling them accordingly is known as part-of-speech tagging, POS-tagging, or simply tagging. Common POS tags include verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections.
An example for POS tagging for the input review sentence is given below.

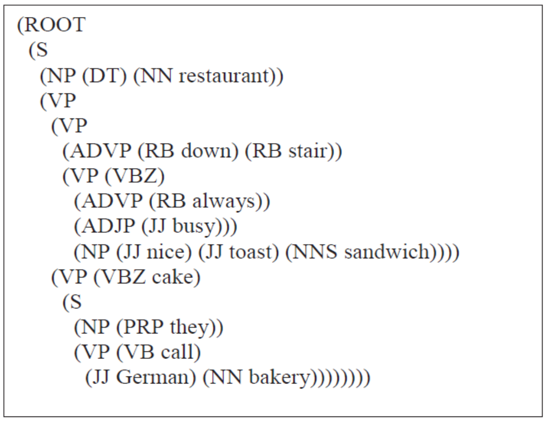
The parser tags the sentence along with its parts of speech identifiers and outputs the tagged sentence as shown in Table 3.1.

Sentence: restaurant down stairs always busy nice toast sandwich cake they call German bakery



Words	POS Tags	Expansion of tags
Restaurant	NN	Noun, singular or mass
down	RB	Adverb
stair	RB	Adverb
always	RB	Adverb
Busy	JJ	Adjective
Nice	JJ	Adjective
Toast	JJ	Adjective
Sandwich	NNS	Noun, plural
Cake	VBZ	Verb, 3rd person singular present
They	PRP	Personal pronoun
Call	VB	Verb, base form
German	JJ	Adjective
bakery	NN	Noun, singular or mass

Each tag represents part of speech like noun, verb, preposition etc. Syntactic features can be extracted from these POS tags. The list of POS tags and its expansion are given in appendix II. The parser generates basic parse tree based on the POS tags annotated. A natural language parser is a program that generates the grammatical structure of sentences. This structure could be evolved by identifying which groups of words go together as "phrases" and which words form subject or object of a verb and so on.



```
(ROOT
  (S
    (NP (DT) (NN restaurant))
    (VP
      (VP
        (ADVP (RB down) (RB stair))
        (VP (VBZ)
          (ADVP (RB always))
          (ADJP (JJ busy)))
          (NP (JJ nice) (JJ toast) (NNS sandwich))))
      (VP (VBZ cake)
        (S
          (NP (PRP they))
          (VP (VB call)
            (JJ German) (NN bakery)))))))
```

Figure 3.2 Basic Parse Tree Generated from Stanford Parser

V.EXPERIMENTAL RESULTS

Ontology Construction

The ontology build using protégé tool (Musen 2015) for reviews from movie domain is shown in Figure 3.5 with top two levels. This tool provides an open source, user friendly interface to build ontologies. This ontology was constructed from the standard sample ontology available.

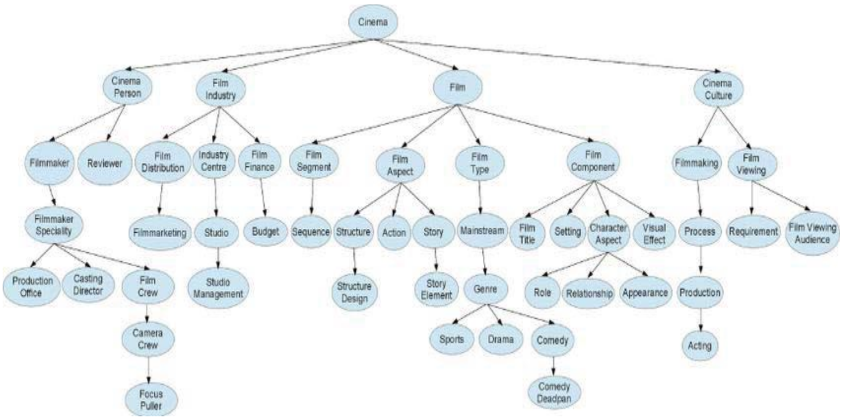


Figure 3.5 Ontology Constructed for reviews from movie domain using protégé

VI. CONCLUSION

The proposed author specific preference aggregation technique has been studied for identification of significant features. The approach combined ontology based domain knowledge with author preferences for all the features. The methodology shows an effective performance in terms of Mean absolute errors and mean squared errors along with ranking loss to compute rank for the identified feature. Domain specific ontology construction and score based on the customer preferences improved the ranking of features. The assessed rank using the system is highly correlating with standard rank computed using the ground truth and domain knowledge. Inclusion of author preferences supported the reviews in reducing these errors and improves the rank of a feature. The method works well for tourism as well as movie domain reviews. The ranks obtained using the system, represent the important features from the domain based on customer perspective. Top ranked features are used for extracting text summaries which enable stakeholders to improve their business decisions.

Reference

1. *Friedrich S, Groß S, König IR, Engelhardt S, Bahls M, Heinz J, et al.. Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. Eur Heart J Digit Health 2021;2:424–436.*
2. *Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.*
3. *van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142–145*
4. *Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al.. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:65–69.*
5. *Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, et al.. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. Eur Heart J 2021;42:2885–2896.*
6. *Tokodi M, Schwertner WR, Kovács A, Tősér Z, Staub L, Sárkány A, et al.. Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. Eur Heart J 2020;41:1747–1756.*