Ensemble Approach for Keyword Extraction

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Bhavneet Kaur

(Roll No. 801532009)

Under the supervision of:

Dr. Sushma Jain
Assistant Professor

Thapar University, Patiala

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "***Ensemble Approach for Keyword Extraction***", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* in Computer Science and Engineering Department of Thapar University, Patiala, authentic record of my own work carried out under the supervision of *Dr. Sushma Jain* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for the award of any other degree of this or any other University.

Bhavneet Kaur
Signature

(Bhavneet Kaur)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Sushma Jain)
Assistant Professor
CSED

ACKNOWLEDGEMENT

No volume of words is enough to express my gratitude towards my guide **Dr.Sushma Jain**, Department of Computer Science and Engineering, Thapar University, Patiala, who has been very concerned and has aided for all the materials essential for the preparation of this thesis report. She has helped me to explore this vast topic in an organized manner and provided me all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. S.S. Bhatia**, Dean of Academic Affairs, **Dr. Maninder Singh**, Head of Computer Science and Engineering Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, for the motivation and encouragement that triggered me for the thesis work.

I would also like to thank my colleagues who were always there and provided with all the help and facilities, whenever I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents and the Almighty for showering the blessings and to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

(Signature)

Bhavneet Kaur

ABSTRACT

Text mining likewise referred to as content information mining, generally comparable to content investigation, is the way toward getting astounding data from content. The data is normally determined through the conceiving of examples and patterns through means, for example, factual example learning. Statistical analysis based techniques has done a broad analysis of the text in perspective of the predicted execution of classification and storage strategies. Analysis of text and techniques occurred when keyword based representation is used to classification of logical content documents. The presentations of content documents with keyword in a concise way can be greatly valuable, since content documents are described by the high dimensionality of highlight space. In this thesis the main emphasis is on the natural language processing based extraction of quality keywords from content. These keywords are used to search the multiple documents in the database or dataset. For the keywords extracted ranking and scoring is given to documents based upon the similarity matching in the documents, similarity matching is based upon the cosine similarity. In the last ensembling of the whole procedure is to be done using the proposed approach. From the results and discussion part it is clear that the accuracy, precision, recall and f-measure parameters are better in case of the proposed approach as compared to the existing system that is statistical analysis based keyword extraction.

TABLE OF CONTENT

LIST OF FIGURE

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Overview

Keyword extraction (KE) is characterized as the undertaking that naturally recognizes an arrangement of the terms that explain the topic of documents in the best way. It is diverse wording, utilized as a part of concentrating the terms that speak to the most appropriate data enclosed in the record: key expressions, key fragments, key terms. Every single recorded equivalent word has a similar capacity – describe the points examined in a report [1]. Separating a little arrangement of units, made out of at least one term, from a solitary record is a significant issue in Content Mining (CM), Data Retrieval and Natural Language Processing (NLP). Keywords are broadly useful toward empowering questions inside IR frameworks as these are anything but difficult to characterize, reconsider, recollect, and share. Contrast with numerical marks these are free of documents and can be connected to different documents and IR frameworks [2]. Keyphrases are additionally remained connected toward enhancing the usefulness of IR frameworks. At the end of the day, applicable removed keywords can be utilized to construct a programme file for a report accumulation or on the other hand can be utilized for record portrayal in order or grouping undertakings [1, 3]. The extractive outline of the report is a center errand of the numerous Information Retrieval and Natural Language Processing applications incorporate programmed ordering, programmed rundown, archive administration, content, record or site arrangement or bunching, cross-class recovery, developing area particular word references, name substance acknowledgment, subject identification, following, and so forth. While allotting keywords to records physically is the expensive, tedious and repetitive errand, and notwithstanding that, the quantity of computerized accessible reports is in developing, programmed keyphrase extraction pulled in the specialist's enthusiasm for the most recent couple of years. In spite of the fact that the watchword extraction applications more often than not chip away at single reports,

catchphrase extraction is additionally utilized for more unpredictable assignment (for example catchphrase extraction aimed at the entire accumulation [4], the whole place or for programmed webpage rundown [5]). The presence of enormous information, building a viable model for content portrayal turns out to be considerably more earnest and requesting in the meantime. Cutting edge methods for KE experience, versatility and sparsity issues. Keeping in mind the end goal to bypass these restrictions, new arrangements are always being proposed.

Keyword Extraction methods divided into two categories as follows:

Assignment: In this method, keywords are selected from the controlled vocabulary and according to corpus content documents are divided into categories.

Extraction: In this method, keywords are selected from the words present in the text. Words present in documents are analyzed and keywords are extracted.

The Following figure demonstrates the classification of the keyword extraction methods

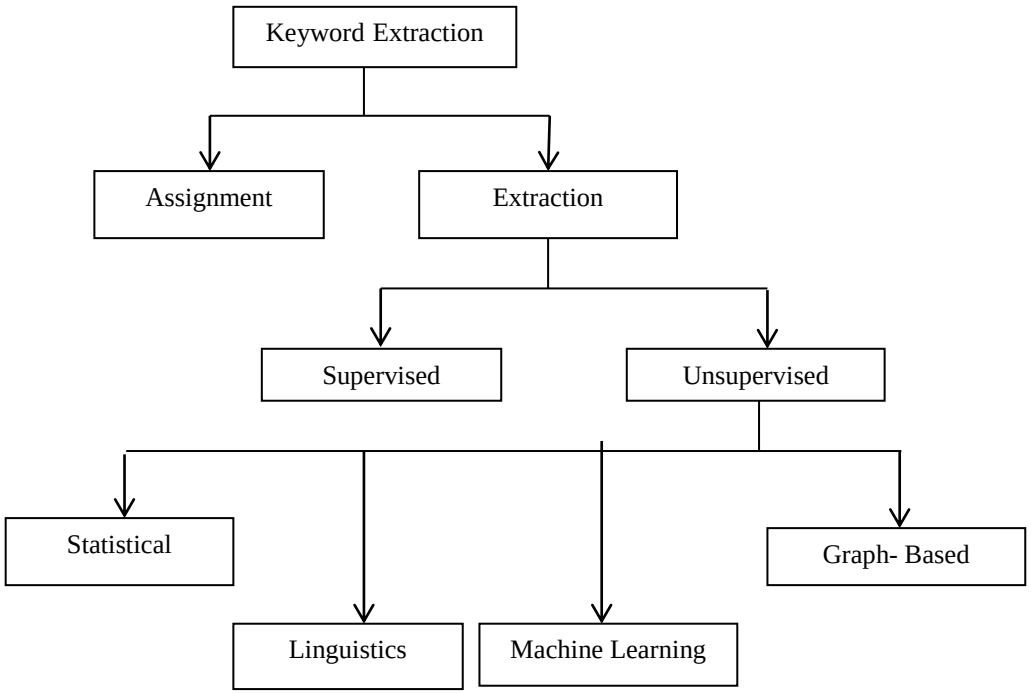

Figure 1.1 Keyword Extraction Methods

Keyword Extraction Approaches

- **Statistical Approaches**

These approaches use the simple techniques which don't require the data to be trained and are domain independent. The statistics of the terms in a document from the whole corpus are utilized to distinguish keywords: n-gram statistics, word reoccurrence, TF-IDF. The drawback of this technique is that in important documents, for example, healthcare and medicine, the most efficient keyword might appear one time in the corpus. Utilization of the statistical techniques will accidentally clean out the important terms.

- **Language-based Approaches**

These approaches utilize semantic features of the terms of corpus, sentences and text document. Morphology, syntax, semantic and talk study are possibly the foremost well-known yet difficult examination.

- **Machine Learning Approaches**

These approaches utilize the supervised or unsupervised learning methods from the illustrations, but mostly supervised approaches are used for keyword extraction. The model that uses the trained keywords that technique is known as Supervised learning. These require a physical comment in learning datasets i.e. to a great degree repetitive and conflicting (in some cases demands predefined scientific classification). This approach contains Decision tree, SVM, Naïve Bayes, C4.5 etc. Thus, these models require the data to be trained and are language dependent.

- **Graph-Based Approaches**

This technique is Graph based model, that empowers study of the connection and fundamental data in an efficient way. In this method, text documents are modeled as the graph where terms of the text document are denoted by vertices and relationship within words are denoted by edges.

1.2 Text Mining

Text mining additionally alluded to as agreeable advice mining, about commensurable to the agreeable investigation, is the way towards accepting alarming abstracts from the content. Abstracts are commonly bent through the conceiving of examples and patterns through means, for example, absolute archetype learning. Agreeable mining added

generally than not includes the way against acclimation the advice agreeable (normally parsing, alongside the amplification of some bent etymological apparatus and the banishment of others, and after accession into a database), answer designs central the organized information, after appraisal and adaptation of the yield. 'High caliber' in agreeable mining as an aphorism alludes to some alloy of pertinence, curiosity, and arresting quality. Ordinary agreeable mining undertakings absorb agreeable arrangement, agreeable bunching, idea/substance extraction, bearing of diminutive accurate classifications, acceptance investigation, address synopsis, and aspect affiliation announcement (for archetype acquirements relations a part of called elements).

The term content examination portrays an arrangement of phonetic, factual, and machine learning methods that model and structure data matter of literary hotspots for business insight, experimental information examination, look into, or investigation. [1] The term is generally identical with content mining; for, sure, Ronen Feldman altered a 2000 depiction of "content mining"[2] to depict "content analytics."[3]. Figure 1.2 shows the primary tasks of text mining.

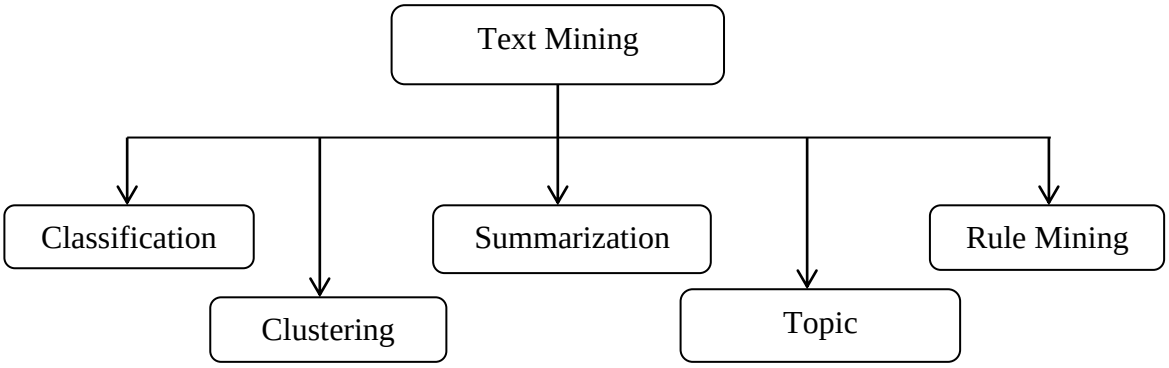

Figure 1.2 Primary Tasks of Text Mining

Information Mining, also known as the discovery of useful information into the databases, introduces towards nontrivial extraction, before unclear and possibly helpful data from the material into databases. While advice mining and acquirements acknowledgment in databases (or KDD) are oftentimes admired as alike words, information mining is quite of the learning disclosure organize. Figure 1.3 shows information mining as a stage in an iterative learning revelation handle [4].

The iterative procedure comprises of following stages:

1. **Text Preprocessing**: It is a level, in which noisy records and insignificant data are expelled from the gathering.
2. **Text Integration**: At this level, various data resources, frequently heterogeneous, might be joined in typical sources.
3. **Text Selection:** At this level, the facts relevant to the research is selected and recovered from the statistics gathered.
4. **Text Transformation:** At this level, the selected information is normalized into the proper way for mining.
5. **Text Mining:** This is the important level where the normalized information is mined using various models and techniques to extract useful content.
6. **Design Evaluation:** At this level, the mined information is evaluated using performance metrics.
7. **Learning Representation:** This is the last level where the mined information is given to the client and is represented using visualization techniques.

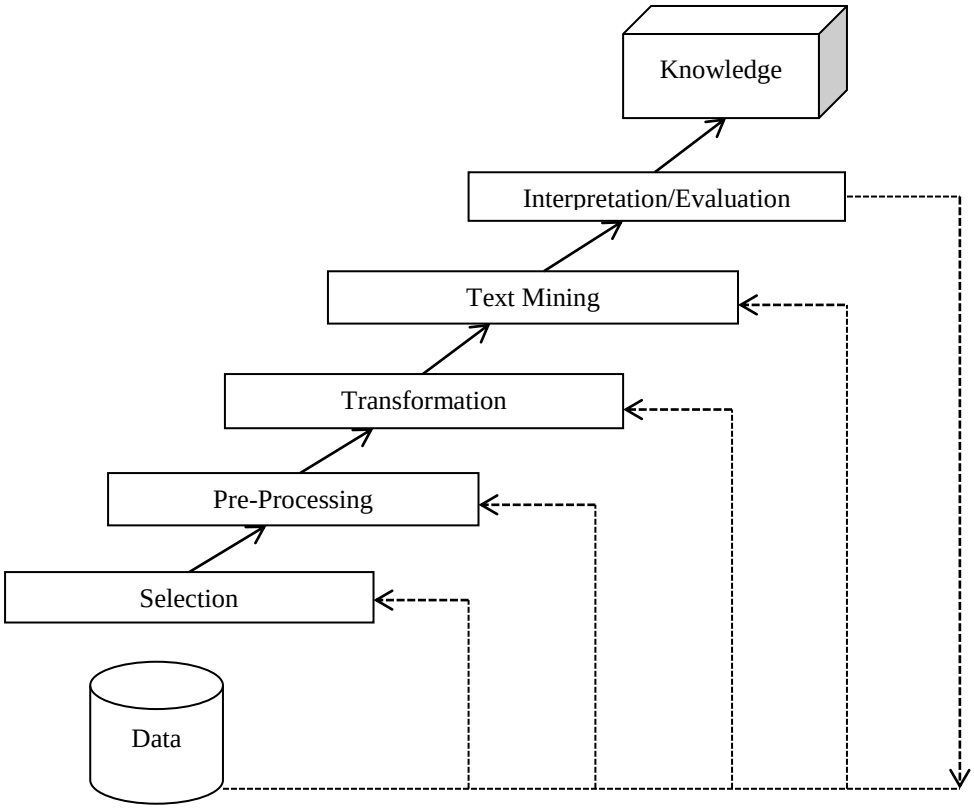

Figure 1.3 Iterative process of Text Mining

It is ordinary to consolidate some of this manner together. For example, statistics cleansing, and information incorporation can be accomplished together as a pre-getting ready stage to create a statistics distribution middle. Information preference and statistics alternate can likewise be joined wherein the union of the records is the effect of the willpower or with appreciating to the example of information distribution centers; the choice is accomplished on changed records. [4].

1.3 Natural Language Processing

NLP is the area of software engineering, man-made brainpower and computational etymology worried with the collaborations amongst PCs and human (normal) dialects, and, specifically, worried with programming PCs to productively handle vast regular dialect corpora.

NLP indicates the Artificial Intelligence methods for language through the insightful frameworks utilized as distinguishing dialect, like English. Formulating of Regular Language is required when we need a smart system like robots that will work according to our instructions. As we require to listen to option after a discourse built proven master framework, and so on. NLP categorized as Regular Language Generation (NLG) and Regular Language Understanding (NLU). At the point when your PC can compose like you, a human, can, that is NLG—customized with assortment and feeling… Understanding the significance of composed content and delivering information which epitomizes this importance is NLU; you have to oversee ambiguities here.

Characteristic dialect handling frameworks take series of words (sentences) as their info and create organized portrayals catching the significance of those strings as their yield. The way of this yield depends vigorously on the job that needs to be done. A characteristic dialect understanding framework filling in as an interface to a database may acknowledge inquiries in English which identify with the sort of information held in the database. For this situation, the significance of the information (the yield of the framework) may be communicated as far as organized SQL questions that could be straightforwardly given into data memory. The primary utilization of PCs to control everyday dialects changed into inside the Fifties with endeavors to computerize interpretation amongst Russian and English [Locke and Booth]. These frameworks were

breathtakingly unsuccessful requiring human Russian-English interpreters to pre-alter the Russian and put up-regulate the English. In view of World War II code-breaking techniques, they took singular words in confinement and checked their definition in a lexicon. They have been of minimal down to earth utilize. Mainstream stories about these frameworks refer to numerous misinterpretations including the expression "pressure driven slam" deciphered as "water gate". In the 1960s characteristic language preparing frameworks begun to look at sentence structure, however frequently in an impromptu way. These frameworks depended on example, coordinating and few inferred portrayals of significance. The most surely understood of these is Eliza [Weisenbaum] However this framework was not the greatest as far as its capacity to concentrate importance from dialect. Genuine upgrades in feature dialect handling came about within the initial and middle 1970s by way of frames useful better massive procedures and enterprise to properly picture the principles of idiom they function with. LUNAR [Woods 1973] presented an English edge towards catalog allotment arguments of concern of moon tremble forms. SHRDLU [Winograd] interface by a digital robotic into a world of portions, bearing English orders towards transporting the squares all over the place and solution explores the state of the area. Meanwhile, that idea present was equivalent developments of mind plus advances that provide the clue to present-day daily language getting ready frames. Look into in PC semantics has given more noteworthy information on sentence structure developments [Gazdar] and Artificial Intelligence scientists have delivered more compelling component for parse regular information and for talking to implication [Allen]. Normal dialect handles framework that is expanded on a strong base of the etymological review and utilizes exceptionally created semantic representation. As of late (amid the 1990s) characteristic dialect frameworks have either centered around particular, constrained spaces with some achievement or endeavored to give broadly useful dialect understanding capacity with less achievement. A noteworthy objective in contemporary dialect handling examination is to deliver frameworks which work with finish strings of talk (with human-like capacities) as opposed to just with separated sentences [Russell and Norvig(a)]. Accomplishments around there are right now restricted.

1.3.1 Segments of NLP

Regular Linguistic Understand (NLU)

Following tasks define the use of NLU

- Map the input value to the language and represent it into useful representation.
- Various parts of the language are analyzed.

Common Language Generation (NLG)

Its method of delivery of major terminologies and sentence as characteristics of language from some input explanation.

It comprises:

Content arrangement − It combines improving the significant material from learning method.

Sentence arrangement − It integrates a selection of useful terms, framed with important vocabularies, set the quality of the sentence.

Content Realization − It map the sentence arrangement in the sentence proper structure.

The NLU is more challenging than NLG.

Entity extraction

Entity extraction includes portioning a sentence to distinguish and extricate substances, for example, a man (genuine or anecdotal), association, topographies, occasions, and so on. NLP APIs utilizes online information from sources like Wikipedia or different storehouses to coordinate these substances. One of the fundamental difficulties is to coordinate distinctive varieties of a substance and bunch it as the same. [32]

For instance, except that there is a substance called Howard Roark. In a given article, the varieties of this substance could incorporate Roark, Mr. Roark, Howard Roark, et cetera. The calculation ought to have the capacity to distinguish and bunch every one of these varieties.

Substance extraction has two imperative segments

Substance sort: Person, put, association, and so forth.

Striking nature: Importance of the centrality of an element on the size of 0 to 1 (these scores demonstrate the significance of the substance to the whole content, with scores more like 1 being more vital than those more like 0).

Syntactic analysis

Syntax alludes to the correct requesting of words. Do the words you've assembled shape a "right" sentence? It manages the auxiliary parts of words in the sentence. And after that, you utilize a parsing calculation to deliver a "tree," which gives you the syntactic connections between the constituents as indicated by setting free sentence structure. (This is a decent video. Caution! Quality not very good). In sentence extraction, content is separated into sentences. In tokenization, the content is separated into tokens (like words or accentuation in common dialect) to which syntactic data are included by regular dialect API. The tokens are placed in this reliance tree you see underneath. This investigation incorporates parts of discourse labeling, piecing, and sentence gathering. [34-38].

Semantic Analysis

After a sentence is parsed to concentrate elements and comprehend the language structure, semantic examination finishes up the importance of the sentence in a setting free frame as an autonomous sentence. The derived significance may not be the genuine plan of the suggested meaning. After a sentence is parsed to concentrate substances and comprehend the punctuation, semantic examination finishes up the significance of the sentence in a setting free shape as an autonomous sentence. The surmised importance may not be the real plan of the inferred meaning. [39-43].

Sentiment Analysis

Once the syntactic and semantic investigation has been finished, we attempt to comprehend the notion behind each sentence. Assessment will incorporate feelings, sentiments, and states of mind. We are talking subjective impressions and not realities. This is likewise alluded to as supposition mining (a capable apparatus in online networking). For instance, to decide if an audit is certain or negative, you utilize size (degree of passionate substance in the content) and scores (general feeling of the content). Extremity esteems for positive substance will be +1 and that for negative substance will be −1. A record with a score of 0.3 and a size of 3.8 will be marginally positive with an obvious level of feeling. In the event that your record is long, the greatness esteem is probably going to be high.

Pragmatic analysis

In the event that you go to your proofreader and request that she recommend a superior sentence structure for a line, her quick question to you will be, "What's the unique situation?" Most of the time, because of the adaptability of the regular dialect, complexities emerge in deciphering the importance of a secluded articulation. Down to earth investigation utilizes the set of expression—when, why, by whom, where, to whom something was said. It manages aims like a reprimand, advises, guarantees, demand, et cetera. For instance, on the off chance that I say "You are late," is its data or feedback? In talk reconciliation, the point is to investigate the announcement in connection with the first or succeeding explanations or even the general section so as to comprehend its significance. Take this one: Chloe needed it. ("It" relies on upon Chloe). Practical investigation deciphers the importance as far as the setting of utilization dissimilar to semantics. [44].

Morphological Processing

The preparatory stage, which happens before language structure investigation is morphological handling. The reason for this phase of dialect handling is to break the string of dialect contribution to tokenization comparing to the distinct words, sub-word and accentuations frames. A word like "despondently" could be subdivided into three tokens. It's basically concerned fundamentally by perceiving in what way the base words are altered into frame different word by comparative implications however regularly with various syntactic classifications. Alteration ordinarily happens with the expansion of prefixes as well as postfix yet another literary changes can likewise occur. All in all, there are three unique instances of word frame change.

Emphasis: literary portrayals of words change in light of their syntactic parts. In English, for instance, most plural things take - s as a postfix (and may require another adjustment), near and superlative types of general descriptive words take - er - EST additions. [45]

Determination: new words are gotten from existing words. This assembling of words frequently happens in a standard way permitting taking after clear morphological tenets. For instance, in English, a few descriptive words take - ness as a postfix while being utilized to make things (glad → joy). Similar standards apply in most human dialects, however, the guidelines are different Intensifying: new words are framed by gathering

existing words. This happens rarely in English (illustrations incorporate "a migraine" and "toothpaste") yet is generally utilized as a part of different dialects where it is morphologically conceivable to have boundlessly long words.[47]

The way of morphological preparing is vigorously reliant on the dialect, being broke down. In a few dialects, single words (utilized as verbs) contain all the data about the strained, individual and number of a sentence. In different dialects, this data might be spread crosswise over many words. For instance the English sentence "I will have been walking..." where complex tense data is just accessible by looking at the structure of helper verbs.

Linguistic Structure and Semantics:

A language processor must complete various diverse capacities principally based on linguistic structure examination and semantic investigation. The motivation behind punctuation examination is two-overlay: to watch that a series of words (a sentence) is very much framed and to split the sentence into a structured way that demonstrates the syntax connections between the distinctive terms. A syntactic analysis (or parsing) does this utilizing a lexicon of word definitions (the vocabulary) and an arrangement of punctuation principles (the sentence structure). A straightforward dictionary just contains the syntactic class of each word, a basic syntax portrays rules which demonstrate just how syntactic classifications can be consolidated to frame expressions of various sorts [48].

Semantics and Pragmatics

After semantic examination, the following phase of handling manages pragmatics. Sadly, there is no all around concurred refinement amongst semantics and pragmatics. This archive, in the same way as a few different creatures [Russel and Norvig] makes the qualification as takes after: semantic investigation partners importance with disconnected expressions/sentences; down to earth examination translates the consequences of semantic investigation from the point of view of a particular setting (the setting of the discourse or condition of the world and so on). This implies with a sentence like "The vast feline pursued the rodent" semantic investigation can create an expression which implies the substantial feline yet can't complete the further stride of derivation required to recognize the expansive feline as Felix. This would be surrendered over to down to earth

examination. [49] At times, similar to the case, simply portrayed, businesslike examination essentially fits real protests/occasions which exist in a given setting with question references acquired amid the semantic investigation. In different cases, the realistic investigation can disambiguate sentences which can't be completely disambiguated amid the linguistic structure and semantic examination stages. For instance, consider the sentence "Put the apple in the wicker bin on the rack".

1.3.2 Linguistic Features and Methods

Linguistics is the logical investigation of language and includes an examination of language frame, language significance, and language in the setting.

Linguistics generally break down human language by watching an exchange amongst sound and significance. Phonetics is the investigation of discourse and non-discourse sound, and dives into their acoustic and articulatory properties. The study of language significance, then again, it manages the relation between language encoded elements and their properties. The different part of a word to pass it on, processes, and appointed importance additionally oversees and resolution equivalences. While the investigation of semantic ordinarily frets about truth conditions, pragmatics manage how situational setting impacts the creation of importance.

Semantic, syntactic and orthographic components and techniques in NLP are normally utilized as a part of the keyword extraction framework. In this segment distinctive systems for distinguishing semantic elements and techniques, which can be utilized as a preprocessing instrument for keyword extraction strategies, are explained.[50]

- **Parts of Speech Tagging (POS):** In corpus linguistic, grammatical feature labeling (POS labeling or POST), additionally called syntactic labeling or word-classification disambiguation, is the way toward increasing the term in a content (collection) as comparing toward a specific linguistic form, in light of both its definition and its unique circumstance—i.e., its association with contiguous and related words in an expression, sentence, or section.[51] A rearranged type of this is generally educated to class age kids, in the recognizable proof of words as things, verbs, descriptors, qualifiers, adjectives and so on. POS tagging as shown in the example:

Table 1.1 POS tagging

Tokens	Heating	Water	In	Big	Vessels
Tagging	Verb	Noun	Pronoun	Adjective	Noun

- **Stemming:** Records written into regular language that is syntactically right for the most part contain words written in various languages. For instance, word "Organize", that has present form "Organizes", present participle form "organizing" and past participle form "organized" be that as it may they all originate from a similar word. There are likewise grouped of related words, for example, education, educated, educational which don't originate from a similar fundamental word frame yet they are related ought in this way be stemmed to a similar base. To discover the expressions of a report that originates from a similar form or words that are identified with each other a method called stemming can be utilized. Stemming is a strategy that stems the end of the word to decrease Inflectional structures. This method can be utilized to discover how normal or phenomenal a word is to record.
- **Lemmatization**: Lemmatization (or lemmatization) in etymology is the way toward gathering together the curved types of a word so they can be investigated as a solitary thing, recognized by the word's lemma, or lexicon form. In computational etymology, lemmatization is the algorithmic procedure of deciding the lemma of a word in light of its expected importance. Not at all like stemming, lemmatization relies on upon effectively recognizing the proposed grammatical form and importance of a word in a sentence, and in addition inside the bigger setting encompassing that sentence, for example, neighboring sentences or even a whole report. Subsequently, creating proficient lemmatization calculations is an open range of research.
- **Co-occurrence**: Co-occurrence is a linguistic term that can either mean simultaneousness/happenstance or, in a more particular sense, the above-chance successive event of two terms from a content corpus close by each other in a specific request. Co-event in this etymological sense can be translated as a marker of semantic nearness or an informal expression. Corpus semantics and its measurement

investigations uncover examples of co-events inside a dialect and empower to work out average collocations for its lexical things. A co-event confinement is recognized when phonetic components never happen together. Examination of these confinements can prompt disclosures about the structure and improvement of a language.[52]

- **Stop Words:** In figuring, stop words will be words that are sifted through before or in the wake of preparing of regular language information (text).[1] Though stop words ordinarily allude to the most well-known words in a dialect, there is no single widespread rundown of stop words utilized by all normal dialect handling apparatuses, and undoubtedly not all devices even utilize such a rundown. A few devices particularly abstain from expelling these stop words to bolster state seek.

1.4 Ensembling

Ensembling models were utilized to profit by the qualities of various demonstrating systems and techniques. When a person has to make essential decisions, then they prefer to rely on the suggestions of a number of experts to make essential decisions than of their alone opinion or of an individual advisor.in order to find the optimal solution, views of different experts are being combined.

Similarly, in Machine Learning, an observable methodology for making decisions more consistent is to combine the result of various classifiers to form the final output. Several methods in the field attain this by learning an ensemble of classifying models and using them together.

Ensembles, also known as groups, are rightly receiving growing attention and compliment, generating a wealth of research. Theoretic and Experimental readings have confirmed that ensembles frequently rise analytical presentation over a single improper classifier and can be applied to diverse kinds of problems, as numeric forecast (the output is a number between 0 and 1) problems and classification (like binary classification: keyword and not keyword) tasks as well Diverse methods for joining models occur, like bagging, boosting and stacking, presence the first two the most widely used. These mutual models share sometimes though the disadvantage of being difficult to study, once they can include lots or even hundreds of discrete classifiers. Figure1.5 demonstrates the

ensemble knowledge method. Though ensembles achieve well, it is not trivial to recognize in a natural way which factors are paying for the improved results of such methodologies.

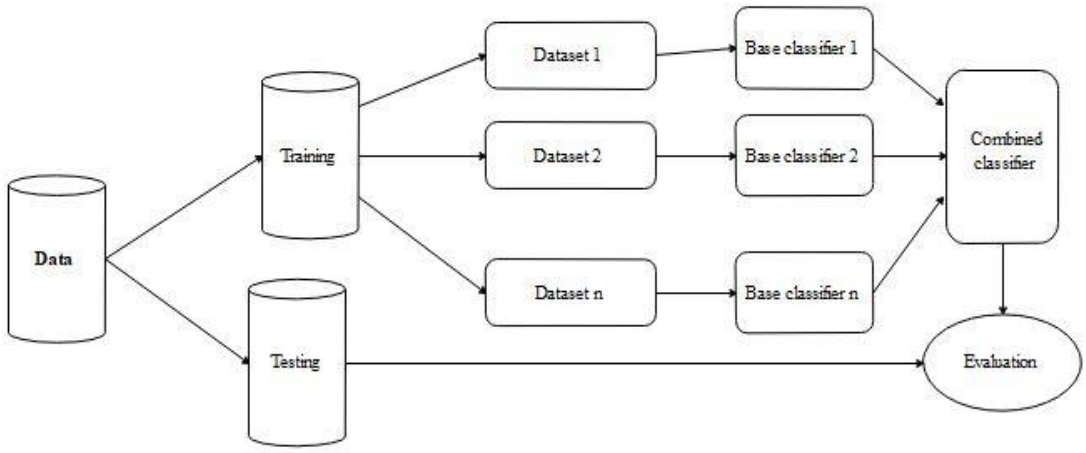

Figure 1.4 Logical View of Ensemble Learning

- **Boosting**

It includes incrementally constructing an ensemble via making each new model happening to highlight the research cases that earlier models misarranged. At times, boosting takes performed to yield better exactness over stacking, yet it also has an inclination to perhaps over-fit the training information. By a long round, the supreme well-known effecting of Boosting is Adaboost, although some additional date intentions are accounted for to achieve better conclusions.

Similar bagging, it is used to combine unbalanced simulations of the similar type, though boosting does that iteratively while in bagging discrete models are made in single (i.e. equal mass is given to each one), in boosting all original model is subjective by the performance of those constructed previously (i.e., models that label properly more examples are assumed more weight, existence this weights directly relative to each model's specific consistency).

- **Stacking**

Sometimes called stacked generalization includes preparing a learning calculation to consolidate the expectations of a few other learning calculations. To begin with, the greater part of alternate calculations are prepared utilizing the accessible information, at

that point a combiner calculation is prepared to make the last expectation utilizing every one of the forecasts of alternate calculations as extra data sources. In the event that a subjective combined calculation is utilized, at that point stacking can hypothetically speak to any of the gathering procedures portrayed in this article, in spite of the fact that by and by, a solitary layer strategic relapse model is regularly utilized as the combiner.

CHAPTER 2

STATE OF ART

2.1 Statistical Approaches

Manchanda and Athavale [1] explained various statistical techniques used in natural language processing. Statistical techniques explained are HMM-based technique, conditional random field-based technique, support vector machine based technique and N-gram based technique. These techniques are used in developing fundamental applications like morphological analyzer, part of speech tagger, phrase chunker and clause boundary identification etc. and advanced applications like grammar checker, spell checker, summarization system etc.

Ranjan *et al*. [1] depicted three periods of common dialect preparing specifically, dialect displaying, parts of discourse labeling and parsing, plotting the methodologies utilized that can be utilized. The essential objective of NLP is computerized comprehension of the semi-organized dialect that people utilize. This review stems application in assorted fields like semantic investigation, synopsis, content arrangement and so forth. Space common dialect handling is a youngster area with no solid sign of when it will develop. Contrasted with entrenched areas like "Investigation of Algorithms", NLP is yet in its rising period and from this time forward, there's deficiency in a minimized piece of composing that clarifies on the times of NLP and records unmistakable methods that can be balanced. NLP gets stronger from foundational subjects of concentrate like bits of knowledge, probability theory, and speculation of computation.

Nivre *et al*. [2] Recognized three sorts of strategies that are important to this venture: application techniques, securing strategies, and assessment strategies. Utilizing cases from the present writing, we demonstrate that each of the three sorts of strategies might be factual as they include the idea of the likelihood or different ideas from the measurable hypothesis. Moreover, we demonstrate that these measurable strategies are frequently joined with conventional semantic tenets and portrayals. In perspective of these realities,

we contend that the evidence polarity between "run based" and "factual" strategies is an over-disentanglement, best case scenario.

Liu *et al.* [3] Assessed existing approaches and created frameworks, and talk about how existing techniques can profit the improvement of biomedical ontologies. One critical prerequisite of space ontologies is that they should accomplish a high level of scope of the area ideas and idea connections. Be that as it may, the improvement of these ontologies is ordinarily a manual, tedious, and frequently blunder inclined process. Constrained assets bring about missing ideas and connections and in addition trouble in refreshing the cosmology as learning changes. Systems created in the fields of Natural Language Processing, data extraction, data recovery and machine learning give methods to mechanizing the improvement of a metaphysics from free-content reports.

Chen *et al.* [4] Presented a neural tensor system (NTN) demonstrate which predicts newly related passages that can be added to the database. This model can be enhanced by instating element portrayals with word vectors taken in an unsupervised manner from the content, and while doing this, current relations can even be questioned for substances that were absent in the database.

Zou *et al*. [5] Presented bilingual word embedding: semantic embedding related crosswise over two dialects with regards to neural dialect models. The paper proposed a technique to take in bilingual embedding's from an enormous unlabeled corpus while utilizing MT word plans to urge translational indistinguishable quality. The new embedding's basically out-perform baselines in word semantic likeness. A lone semantic comparability highlight impelled with bilingual embeds includes closing the large portion of a BLEU indicate the consequences of NIST08 Chinese-English machine interpretation assignment.

Pichotta *et al.* [6] Scripts speak to information of cliché occasion arrangements that can help content comprehension. Introductory factual techniques have been created to take in probabilistic scripts from crude content corpora; notwithstanding, they use an extremely ruined portrayal of occasions, comprising of a verb and one award contention. Creator introduced a script learning approach that utilizes occasions with various contentions. Dissimilar to past work, we display the connections between different elements in a script. Probes an expansive corpus utilizing the assignment of gathering held-out

occasions (the "story close assessment") show that displaying multi-contention occasions enhances prescient exactness.

NLP Modelling

Roller *et al.* [7] Redesigned a - dimensional multimodal sort of Idle Dirichlet Designation (Andrews et al., 2009) in different strategies. 1. Beat substance just forms in unmistakable surveys and tried that low-organize noticeable added substances are no ifs and or buts immaculate with the present variant. 2. Demonstrated a particular way to manage be a piece of noticeable components into the LDA demonstrate the utilization of unsupervised social occasions of pix. The social affairs are truly interpretable and improve our assessment tries. Given novel strategies to build the bimodal models to support under the modalities we find that the 3-, Four-, and five-dimensional models essentially defeat models the use of least complex alone or two modalities and that not uncovered modalities each convey separated, disjoint insights that can't be constrained directly into a common, dormant shape**.**

Guadarrama *et al.* [8] displayed an answer by taking a small video clasp besides yields a concise sentence that sums up and about the fundamental movement in the video, for example, to the performing artist, activity, question. Dissimilar to past work, this approach takes a shot at out-of-area activities: it doesn't require preparing recordings of the corrective action. On the off chance that it can't locate an exact forecast aimed at the pre-prepared model, that finds a fewer particular answer that is additionally conceivable from a down to earth outlook. We utilize semantic chains of importance gained from the information to pick a fitting side by side speculation, and prior gained after web-scales common dialect collection to punish impossible mixes the performing artists/activities/objects; we likewise utilize a web-scales dialect models to "fill in" new verbs, for e.g. at the point when the term does not show up in the preparation setup.

Pichotta *et. al.* [9] Tended to the issue of recognizing multiword expressions in a dialect, concentrating on English phrasal verbs. Our multilingual positioning methodology incorporates recurrence measurements from interpreting corpora in 50 distinct dialects. Our test assessment shows that joining factual confirmation from many parallel corpora utilizing a novel positioning focused boosting calculation delivers an extensive

arrangement of English phrasal verbs, accomplishing execution tantamount to a human-curated set.

Bhosale *et al.* [10] Exhibited an approach for recognizing special substance in Wikipedia. By joining stylometric highlights, including highlights in light of n-gram and PCFG dialect models, we exhibit enhanced precision at recognizing limited time articles, contrasted with utilizing just lexical data and meta features.

Garrette *et al.* [11] Developed normal language handling apparatuses for low-asset dialects frequently requires making assets starting with no outside help. While an assortment of semi-directed techniques exists for preparing for deficient information, there are open inquiries with respect to what sorts of preparing information ought to be utilized and what amount is vital and examined a progression of examinations intended to reveal insight into such inquiries with regards to grammatical form labeling.

Krishnamoorthy *et al.* [12] introduced all-encompassing information driven system that produces normal dialect portrayals for recordings. And joined the yield of cutting edge protest and movement indicators with "real world" learning to choose the most plausible subject-verb-question triplet for depicting a video. We exhibit that this adapting therefore mined from web-scale content corpora, updates the triplet decision figuring by giving it significant information and prompts a four-wrinkle to augment in real life recognizing confirmation. Not under any condition like past systems, this approach can remark on subjective recordings without requiring the expensive aggregation and remark of a tantamount planning video corpus.

Raghavan *et al.* [13] considered the issue of learning rational information as first-request rules from fragmented and boisterous common dialect extractions created by an off-the-rack data extraction (IE) framework. A great part of the data passed on in the content must be deduced from what is expressly expressed since effectively inferable actualities are once in a while said. The proposed control learner represents this marvel by learning rules in which the body of the choose contains relations that are for the most part unequivocally communicated, while the head uses a less-frequently determined association that is easily derived. They oversee leaner shapes get ready cases in an online approach to empower it to scale to broad substance corpora. Plus, we propose a novel approach to managing weighting rules using a curated lexical theory like WordNet.

From above writing, study it is inferred that there are many points of interest and the burdens of NLP. Preferences of NLP are as per the following:

- Highly expressive
- Permits an assortment of getting to focus
- Highly adaptable
- Highly illustrative of reality
- Represents (any) many perspectives

Chandola *et al.* [14] Showed that records, mining brings a relationship of mechanical assemblies and methodologies that can be identified with find covered plans that give therapeutic offerings specialists a further wellspring of learning for recognizing. In more detail, gathering the patients that have a similar status permits finding new sickness, however, the correct number of groups is not much of the time self-prominent. This paper most importantly reviews exhibit methods for picking a number of associations for the computation. At that component, a more alluring estimation is presented for picking up learning of alright in the meantime as packing.

Reddy and Sujeeth [15] characterizes that it is a noteworthy issue to recover great sites from the bigger accumulations of sites. As the quantity of accessible page develops, it is turning out to be more troublesome for clients discovering archives significant to their interests. The gathering is the portrayal of an enlightening file into subsets (clusters) so that the data in each subset share some standard characteristic - as often as the possible region as demonstrated by some described evacuate measure. By bundling we upgrade the idea of destinations by social affair relative locales in get-togethers. This paper addresses the employments of data mining, mechanical assembly Weka by applying implies bunching to discover groups from colossal informational indexes and discover the characteristics that administer improvement of web search tools. Unlabeled archive accumulations are ending up plainly progressively normal and mining such databases turn into a noteworthy test.

2.2 Machine Learning Approaches

Agahi *et al.* [16] Grouping information into significant bunches is vital in information mining. K-implies bunching is a quick strategy for discovering groups in information. The essential disparities are a prescient instrument in information seek to utilize imbalances that are particular for Euclidean separation. A stretched out imbalance identified with the Hölder sort for all inclusive indispensable is acquired in a fairly broad shape.

Elgohary *et al.* [17] gave a review about k-imply that part k-means are a powerful technique for information grouping which amplifies the generally utilized k-implies calculation to take a shot at a likeness lattice over complex information structures. The portion k-implies figuring is, however computationally amazingly baffling as it requires the aggregate data grid to be found out and sent away. Further, the kernelized method for the part k-infers count keeps the parallelization of its figurines on current systems for scattered enrolling. This paper has described a gathering of bit based low-dimensional embeddings that think about scaling piece k-suggests on Map Reduce by methods for a profitable and bound together parallelization strategy. A brief timeframe later, we propose two strategies for low-dimensional embedding's that hold quick to our importance of the embedding family.

Zhang and Fang [18] portrays that The customary K-implies calculation is a generally utilized bunching calculation, with an extensive variety of uses. This paper presented the possibility of the K-implies bunching calculation examination the favorable circumstances and detriments of the conventional K-implies grouping calculation expound the technique for enhancing the K-implies bunching calculation in light of enhancing the underlying point of convergence and decide the K esteem. Recreation tests demonstrate that the enhanced grouping calculation is not just more steady in bunching process, in the meantime, enhanced bunching calculation to lessons mineral even stay away from the effect of the clamor information in the dataset question guarantee that the last grouping outcome is more exact and viable.

Ndehedehe *et al.* [19] Described that information mining is the use of specific computations for confining plans from records. Various Information mining approaches

had been connected on sizable volumes of records to discover hid cases and associations pleasing in straightforward authority. This work inquires about the resolute palatable of K-approach, a standard and greatest clear unsupervised acing figuring in Land Utilize Land Cover mapping of the Uyo Capital City. The spatial subset of the requested imagery and the ground reality Records tried for these works of art was a 500m x 500m window. The outcomes have been in like manner insisted with the guide of overlaying the uncommon group packs with various certify Records assets like Orthophoto and digitized vector of a practically identical area. The use of K-way packing exam in arriving user request may moreover furnish us with fundamental disclosures and dependable gathering comes roughly like the directed and machine becoming more acquainted with computations**.**

Sakthi and Thanamani [20] presented that because of the expansion in the amount of information over the world, it ends up being the exceptionally perplexing errand for investigating that information. Classify that information into astounding gathering is one of the essential sorts of perception and learning. This prompts the essential for better data mining methodology. These workplaces are given by a standard data mining framework called Clustering. The key desire of this strategy is to order a dataset into a game plan of gatherings that contains similar data things, as enrolled by some division work. One of the broadly utilized bunching methods is K-Means grouping. K-Means grouping is exceptionally basic and powerful to the cluster. Yet, the fundamental drawback of this procedure is the point at which the vast dataset is utilized for grouping. To defeat this trouble, different analysts concentrate on proposing the better change in K-Means bunching. This paper gives another method to adjust K-Means gathering which can realize better execution. For statement, this paper uses an upgraded version of the Hopfield Artificial Neural Network (HANN) figuring. Furthermore, the Genetic Algorithm (GA) is in combining with K-Means calculation.

Ghalib *et al.* [21] In information mining, the arrangement is a type of information examination that can be utilized to concentrate models portraying imperative information classes. Two of the known learning calculations utilized are Naïve Bayesian (NB) and SMO (Self-Minimal-Optimization). Thus the accompanying two learning calculations are utilized on a Car survey database and in this manner, a model is consequently made

which predicts the normal for an audit remark in the wake of getting prepared. It was discovered that model effectively anticipated accurately about the audit remarks subsequent to getting prepared. Likewise two bunching calculations: K-Means and Self-Organizing Maps (SOM) are utilized and worked upon a Car Database (which contains the properties of a wide range of CARS), and in this manner, the accompanying two outcomes are then analyzed.

Yadav [22] Data mining is the way toward removing from disguised prognostic data from a gigantic measure of databases. It is a persuasive innovation, which encourages organizations to concentrate on imperative data in their information stockrooms. There are distinctive strides in information mining process like Anomaly location, the Association manages to get the hang of, Clustering, Classification, Regression, Summarization. This paper is for the most part worried about bunching which is the technique of sorting out the articles in gatherings whose individuals contains some sort of comparison.

Ghosh and Dubey [23] information mining innovation is considered as helpful resources for recognizing examples and configurations of the huge capacity of information. This methodology is essentially used for the extrication of the obscure example, after the extensive arrangement information for commerce and constant applications. It is a computational insight, teaching which has developed as an important apparatus for information examination, new learning disclosure, and self-ruling basic leadership. The crude, unlabeled information after the substantial capacity of the dataset to be grouped at first in an unofficial manner through utilizing bunch examination that is bunching task of the arrangement perceptions in groups, therefore, perceptions into a similar bunch might be in some sense be dealt with as comparative. The result of the grouping procedure and effectiveness of its space application is for the most part decided through calculations. There are different calculations which are utilized to take care of this issue. In this examination work, two vital grouping calculations to be specific centroid founded K-Means and delegate protest based Fuzzy C-Means bunching calculations to look at.

Oza, Kamat [24] explored the game plan and modification of the subjects in two postgraduate scholastic projects. Information Mining, particularly, information bunching

systems were utilized in mining the connection between the scholarly projects wherein the subject marks (result) were taken as the key measurements.

Kaur [25] Clustering is a basic errand in a Data Mining process that was utilized for a reason to create gatherings and bunches for the information collection in light of the closeness between them. K-Means grouping is the bunching technique that gives informational index that is isolated into numbers of bunches. The paper was planned to provide a presentation around K-implies grouping and calculation of it. Exploratory consequences of K-means grouping and the execution if there should arise an occurrence of execution time are talking about here. Be that as it may, there are sure confinements in K-implies bunching calculation, for example, it sets aside more opportunity aimed at implementation. Thus to decrease the execution time we utilized the Rank Technique and Enquiry Redistribution. Furthermore demonstrated the way grouping was performed in a smaller amount of execution time when contrasted with customary strategy. This work makes the endeavor of concentrate the possibility of the K-means bunching calculation of information taking out utilizing distinctive techniques.

Adhikari and Agrawal [26] proposed a novel weighted outfit plot which astutely joins numerous preparation, calculations to build the ANN conjecture exactnesses. The weight for each preparation, the calculation is resolved from the execution of the comparing ANN show on the approval dataset. Directed analyses on four essential time arrangement delineate that our proposed system accomplishes fundamentally preferable gauge correctnesses over two other well knew actual models. Likewise, it diminishes the specified inadequacies of individual ANN preparing calculations all things considered.

Araque *et al.* [27] developed a deep learning based sentiment classifier using a word embedding model and a linear machine learning algorithm. This class serves as a baseline to compare to subsequent results. Then proposed the two ensemble techniques which aggregate the baseline classifier with other surface classifiers widely used in Sentiment Analysis. Then proposed two models for combining both surface and deep features to merge information from several sources.

2.3 Graph-Based Method

Baliga *et al.* [29] Reviewed the systems and techniques for the errand of watchword extraction. The green review of strategies was gathered which included around an escalated study of current techniques. Business identified with watchword extraction is characterized for overseeing and unsupervised procedures, with a one of a kind compliment on outline based absolutely methodologies. Diverse diagram based procedures are penniless down an idea roughly.

Erkan and Radev [30] presented a stochastic diagram based strategy for figuring relative significance of printed units for Natural Language Processing. We test the system on the issue of Text Summarization (TS). Extractive TS depends on the idea of sentence striking nature to recognize the most vital sentences in an archive or set of reports. Notability is commonly characterized as far as the nearness of specific imperative words or as far as compared to a centroid pseudo-sentence.

Building [31] Analyzed extraordinary centrality measures for graph-based keyphrase extraction. Through trials completed on 3 fashionable datasets of numerous dialects and areas, and reveal that truthful diploma centrality accomplish comes about equivalent to the extensively applied TextRank calculation, and that closeness centrality acquires the satisfactory outcomes on short facts.

Litvak *et al.* [32] Presented DegExt, a diagram based dialect autonomous keyphrase extractor, which broadens the watchword extraction technique portrayed and contrasted DegExt and two best in class ways to deal with keyphrase extraction: GenEx and TextRank.

Ramasubramanian *et. al* [35] approached to make an effective pre-processing steps to keep each space and time necessities with the aid of using progressed stemming set of rules. Stemming algorithms are used to convert the phrases in texts into their grammatical root form.

Several algorithms exist with unique techniques. the most widely used stemming algorithm is “M.F Porter stemming set of rules. However, it nonetheless has certain drawbacks of coping with named entities [36].

Lya Hulliyyatus Suadaa [37] the massive utilization data because of interaction of customers and web may be extracted to be knowledge implemented in diverse application. the primary hassle of web usage mining is the character of facts they address and the dealing with techniques. A survey on web usage mining changed into conducted with systematic literature assessment method to pick out relevant research approximately statistics assets, strategies, applications, and contemporary troubles that could be the important thing of destiny studies path on this vicinity [38-39].

Cristian Moral [40] provided the primary strategies for stemming have been reviewed to extract their major features, advantages, and downsides. Also, papers overseeing stemmers for non-English dialects or with some more present proposition were furthermore counseled and arranged. Ultimately, experimental papers defining the most well-known methods and metrics geared toward evaluating and classifying steamers had been additionally taken into consideration to expose their contributions and outcomes.

Chen *et al.* [41] aimed to complete a know-how base with the aid of predicting extra proper relationships between entities, based on generalizations that can be discerned inside the given information base. we introduce a neural tensor community version which predicts new dating entries that may be introduced to the database.

This model may be progressed by using initializing entity representations with word vectors learned in an unmonitored style from the text, and whilst doing this, existing relations can even be queried for entities that had been now not gift inside the database [42-43].

Pichotta *et. al* [44] tended to the bother of distinguishing multiword expressions in a dialect, that work in English phrasal verbs. our bilingual positioning methodology coordinates recurrence records from interpreted corpora in 50 exceptional dialects.

Shruti Bhosale [46] presented a technique for detecting promotional content material in wikipedia. By using incorporating stylometric functions, together with capabilities primarily based on n-gram and pcfg language models, we reveal stepped forward accuracy at figuring out promotional articles, in comparison to using best lexical facts and meta functions [45].

Garrette *et al.* [47] mentioned a chain of experiments designed to shed light on such questions inside the context of element-of-speech tagging. We obtain timed annotations

from linguists for the low-resource languages Kinyarwanda and Malagasy (as well as English) and examine how the quantities of diverse types of facts have an effect on the overall performance of a skilled pos-tagger [47-50].

Chandola *et al.* [51] supplied an overview of the Minnesota intrusion detection system (minds), which uses a collection of facts mining primarily, based algorithms to deal with distinct aspects of cyber safety. the various components of minds including the test detector, anomaly detector, and the profiling module detect different styles of attacks and intrusions on a pc network. the scanning detector objectives at detecting scans that are the precursors to any network attack.

KondReddy *et al.* [52] addresses the programs of statistics mining device weka through applying adequate method clustering to find out clusters from huge information units and find out the attributes that govern optimization of serps. Unlabeled document collections are becoming increasingly more commonplace and mining such databases come to be a primary mission.

proposed two techniques for low-dimensional embedding that adhere to our definition of the embedding own family. exploiting the proposed parallelization method and furnished scalable.

Elgohary *et al.* [54] map reduce algorithms for kernel ok-way. the kernel ok-means is an powerful approach for statistics clustering which extended the usually-used okay-method algorithm to work on a similarity matrix over complicated statistics systems. the kernel k-method algorithm is however computationally very complex because it calls for the whole statistics matrix to be calculated and saved. further, the kernelized nature of the kernel ok-technique algorithm hinders the parallelization of its computations on current infrastructures for distributed computing.

CHAPTER 3

PROBLEM FORMULATION

3.1 Barriers in previous work

Statistical Analysis for text classification finds applications in a variety of fields, similar to content organization and opinion mining. Henceforth, the improvement of classification algorithms can be helpful in various application fields of content classification. From the perspective of experts, text classification algorithms are a vital device to give valuable knowledge. Alongside the improvement of data innovation, major share of data is put away as text documents. This huge amount of unstructured data provides valuable insights to decision makers with the use of text classification tools.

- In the statistical analysis phase the keywords are extracted from the structured documents but this technique is not efficient.
- In the case of multi-document text summarization, several issues occurs frequently while evaluation of summary such as redundancy, temporal dimension, co-reference or sentence ordering, etc. which makes very difficult to achieve quality summary. Some other issues occurs, such as grammaticality, cohesion, coherence which is harmful for the summary. The above defined problems are not solved or gain importance in case of structural analysis phase.
- To get the quality summary, quality keywords are required for text summarization in case of NLP.
- There is no standard to identify quality keywords within or multiple documents. The extracted keywords are varying for applying different approaches of keyword extraction.
- Multilingual text summarisation is another challenging task.

3.2 Problem Statement

To propose an ensemble approach for quality keyword extraction using natural language processing. As discussed about the multiple barriers like keyword extraction from multiple documents and to do the quality assessment of the keywords, the problem of as

defined above may be extracted in which algorithm has to extract the keyword from the NLP string that is fed as input. This technique had been implemented on the structured forms already, but this technique will extract quality keywords from NLP, so that an accurate information may be extracted which is the main problem.

3.3 Objectives

1. To implement the Statistical approach for keyword extraction.
2. To implement the ensemble approach for keyword extraction.
3. To validate the results of the ensemble approach with statistical approach.

CHAPTER 4

METHODOLOGY

4.1 Language Modeling for Keyword Extraction

Language grounding a process of mapping natural language to relevant aspects of a surrounding perceptual environment is one approach to this goal. A human child “grounds" language in perceptual contexts via repetitive exposure to the co-occurrence of language and perception. Recent research supports the idea that the human language learning process also happens in a statistical manner. Ideally, language grounding systems should be able to mimic the language learning process of humans.

Frameworks based on machine-learning approaches are more effective than manually created rules as follows:

- Learning strategies utilized in machine learning adjusting naturally focus on widely recognized issues, though composing rules by hand it is frequency not evident in the least wherever trial have to be compelled to be coordinated.
- Automatic learn approaches could make utilization of the measurable derivation approach to create models that are hearty to new information (e.g. containing words or structures that have not been seen earlier) and to incorrect info (e.g. with incorrectly spelled words or words coincidentally discarded). For the most part, dealing with such info smoothly with transcribed guidelines — or all the more, by and large, making frameworks of manually written principles that settle on delicate choices — is greatly troublesome, mistake inclined and tedious.

Frameworks in view of naturally taking in the guidelines can be made more precise, just by providing more info information. Be that as it may, frameworks in light of manually written tenets must be made more exact by expanding the many-sided quality of the standards, which is a substantially more troublesome errand. Specifically, there is a point of confinement to the many-sided quality of frameworks in light of hand-created rules, past which the frameworks turn out to be increasingly unmanageable. Be that as it may, making more information to contribute to machine-learning frameworks essentially

requires a relating increment in the quantity of worker hour worked, by and large without critical increments in the unpredictability of the explanation procedure. Figure 4.1 shows the flow diagram for the keyword extraction process.

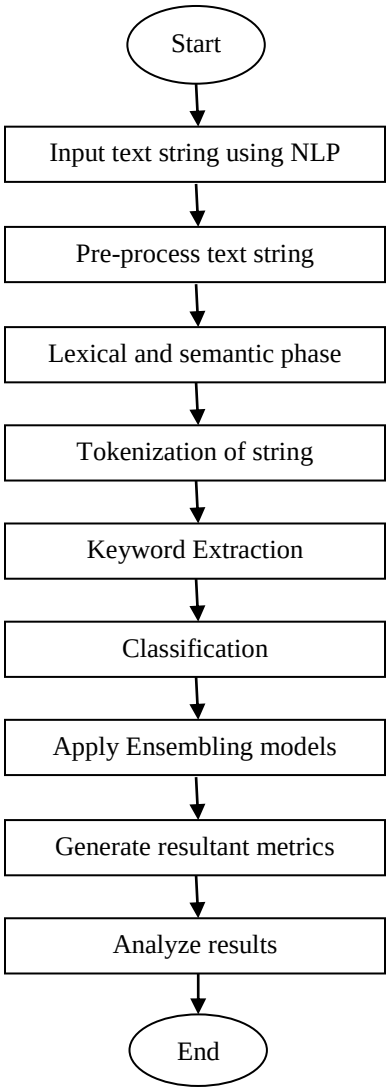

Figure 4.1: FlowChart for Keyword Extraction

4.2 Proposed Ensembled Approach for Keyword Extraction

Input: NL String

Output: Keyword Extraction, Scoring, Classification

Begin

Step 1: S ←Natural Language String

In the very first step a Natural Language String is feed into the algorithm. This string is fed will be pre-processed, i.e. lexical and semantic analysis phase is applied to that string. In this step various data sets and ACM text is considered.

For example: a string may be taken as S ←Keyword Extraction in NLP

Step 2: Token ←Lexical Phase

In the second step the string is processed using lexical phase. Lexical phase is used to convert the string into various tokens. From these tokens the keywords may be extracted.

In this phase the tokens are generated form NLP string. Hence there will be 4 tokens as KEYWORD, EXTRACTION, IN, NLP

Step 3: P ← Preprocess S

- The first step is to clean data for example removing outliers,missing values.
- Then documents are converted into token using tokenization.
- Stop words are removed, Stop words are the words which occur many of the times in the document and have very less importance such as a, an, the, of, into which are just used as helping verbs in the documents.
- Stemming is done, which means to chop off the words which have same meanings but written in various forms. Such as in singular plural or in different tense forms. For eg. system and systems.

Step 4: F← Feature Selection

- After cleaning and tokenization, the important step for document study is feature extraction. It is used to make the vector space that enhances the adapt capability, productivity and the precision of the document classifiers.
- Feature selection is opted to select or choose the subset of components of the documents. It is done by choosing the terms with the highest scores as it implies the effectiveness of the terms.
- It is done for the reason that data is present in high dimensions. By selecting important terms time consumption and cost can be reduced. There are many methods used for feature extraction, such as term re-occurs, gene list, chi-square technique odds ratio of

the terms. Methods used in this work are reoccurrences and term frequency of the terms.

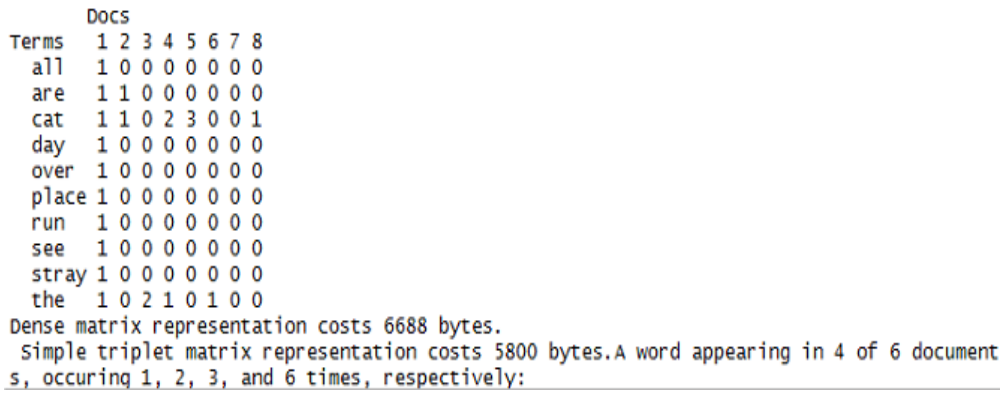

Figure 4.2: Document term matrix

Figure 5.5 show the document matrix of the documents after preprocessing and feature selection.

Step 5: Keywords ← Extract Keywords from Token

Table 4.1 Document Sparsing

Non-/Sparse Entries	22/90
Sparsity	80
Max term Length	7
Weighting	Term frequency

Table 4.1 described the document sparsity and entities in documents that are to be used to calculate the frequency of document. Now from the tokens, the keywords will be extracted by comparing it with the database and the applied algorithm. This step includes lexical phase along with in which the length of document, term length and term frequency is calculated.

Step 6: Apply classification algorithms

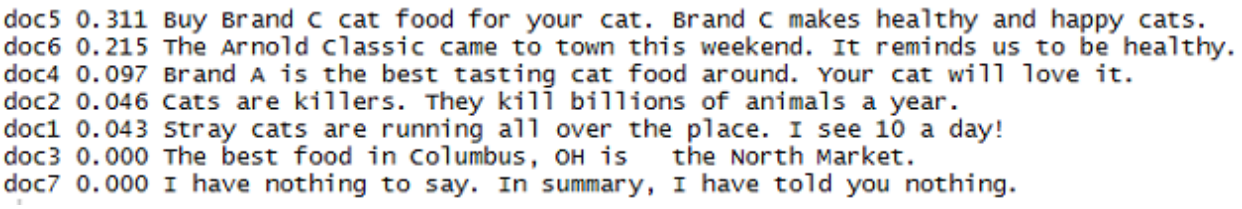

Figure 4.3: Keyword Ranking

Figure 4.3 describe the keywords ranking that is searched in accordance with the natural language inputs. By extracting these ranks or scores of the keywords one may search from the database according to these keywords.

Step 7: Apply ensembling algorithms

In the last steps the classification and ensembling algorithms will be applied and results will be generated.

Following algorithms are used for classification and ensembling:

SVM (Support Vector Machine)

It is a supervised algorithm that is concerned with learning algorithm which examines the dataset that is applied for classification algorithms and regression study. This algorithm creates a model that will put newer examples to one or another division. This algorithm is useful in content and hypertext categorization as by using this algorithm in the applications of content would reduce the requirement of labeling the training data in transductive and inductive situations.

This model is also used in image classification problems. This model gives higher accuracy as compared to other algorithms of machine learning.

MAXENT

This classifier is a probabilistic algorithm that fits the class of exponential models. This algorithm is created on the Norm of Extreme Entropy and as of all the models that fit into the training data, opt for the one and only that has the largest entropy. This algorithm could be used to resolve a bulky variety of content classification issues like language exposure, topic sorting, sentimental exploration and more.

GLMNET

Glmnet is a cluster that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. The algorithm is extremely fast and can exploit sparsity in the input matrix x. It fits linear, logistic and multinomial, Poisson, and Cox regression models. A variety of predictions can be made from the fitted models. It can also fit multi-response linear regression.

Decision Tree

A decision tree is kind of supervised algorithm and mostly used in the problems of classification. This algorithm works both for discrete as well as continuous values. This algorithm is basically used as predictive modeling, statistics, data retrieval as well as machine learning problems.

In this algorithm branches combination of features and leaves represent the outcomes and in this where the outcome is discrete there, this algorithm is known as regression tree.

Following are the combinations used for Ensembling

- SVM and MAXENT
- SVM and GLMNET
- SVM and TREE
- GLMNET and MAXNET
- MAXENT and TREE
- MAXENT and TREE and GLMNET
- MAXENT and SVM and GLMNET

Step 8: Generate and analyze the results with evaluation metrics.

Step 9: End

4.3 Evaluation Metrics

True positives (TP) = Total no. of outputs that are correctly identified are true

False positives (FP) = Total no of outputs that are incorrectly identified are true

True negatives (TN) = Total no. of outputs that are correctly identified are false

False negatives (FN) = Total no of outputs that are correctly identified are false

- **Accuracy**

It is defined as the method to discover the measure of correctly scored documents from the given input. For calculating the accuracy, we need to specify the proportion of true negatives and true positives. The formula to calculate accuracy is shown below:

$$Acc = \frac{True\ P+True\ N}{TP+FN+FP+TN} \quad (1)$$

- **Precision**

This measure is used calculate the quality, it is the most commonly used performance metric, this measure is used to measure the standard deviation which means that how much proportion of defined measures are different from each other. It is used to calculate the statistical probability. The formulae to calculate the precision is given below:

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

- **Recall**

This measure is used to calculate the quantity, it is defined as the proportion of useful instances that are retrieved over the total number of present useful instances. Precision and recall are based upon the fact of measuring relevance. The formulae to calculate Recall is given below:

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negative} \quad (3)$$

- **F-Measure**

This measure is used to calculate the combined value of precision and recall known as the accuracy of both precision and recall. This can also be defined as an evaluation of missing or added value. It calculates the performance of the models and its values are 1 or 0. If calculated values are 1 then it signifies that the model performance is better and if the value is 0 then signifies that performance is worst. The formulae to calculate f-measure is given below:

$$F\ measure = 2.\frac{Precision*Recall}{Precision+Recall} \quad (4)$$

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Dataset Used

1 **Reuters-21578:** The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987.

2 **ACM Digital Library:** To evaluate the results abstracts of various research papers from ACM digital library are used.

5.2 Results of the Evaluation Metrics

In this section statistical and ensembled approach is compared with each other to validate the results of ensembled algorithm that it is accurate as compared to the existing one.

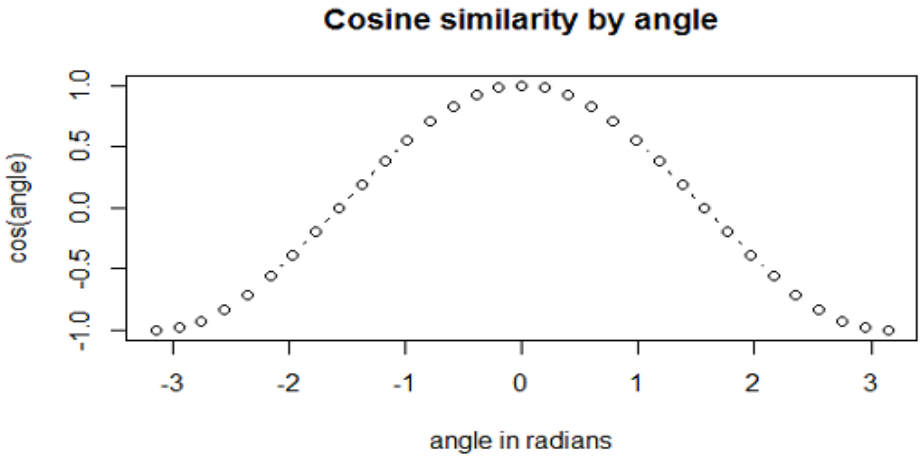

Figure 5.1: Cosine Similarity

Figure 5.1 described the cosine similarity in the document and from this, the keyword ranking may be calculated.

Table 5.1: Comparative Study for Statistical analysis and Ensembling approaches

No of Keywords	Accuracy (Statistical Approach)	Accuracy (Ensembling)	Recall (Statistical Approach)	Recall (Ensembling)
50	83	90	0.89	0.96
40	85	90	0.86	0.94
30	88	92	0.83	0.92
20	90	96	0.8	0.91
10	93	98	0.76	0.9

Table 5.2: Comparative Study for Statistical analysis and Ensembling approaches

No of Keywords	Precision (Statistical Approach)	Precision (Ensembling)	F-Measure (Statistical Approach)	F-Measure (Ensembling)
50	0.9	0.94	0.87	0.92
40	0.82	0.92	0.86	0.9
30	0.82	0.92	0.83	0.9
20	0.81	0.91	0.8	0.87
10	0.8	0.9	0.76	0.8

Table 5.1 and 5.2 is the comparative study for the statistical analysis and ensembling approaches in NLP. In the ensembling approaches four algorithms are used namely SVM, MAXNET, GLMNET and TREE. From the table it is clear that the ensembling values of results are better than that of statistical analysis. From table 5.1 it is clear that the values of accuracy, recall, precision and F-Measure are better in case of ensembling approaches that are nearly 84, 0.92, 0.93 and 0.87 approx.

Table 5.3: Comparative Study for SVM,MAXENT and ensembling

Parameters	SVM and MAXENT (ensemble)	SVM	MAXENT
Recall	0.92	0.83	0.84
Precision	0.82	0.92	0.92
Accuracy	95	90	91
F-Measure	0.90	0.83	0.84

In table 5.3 a comparison between the ensemling and non ensembling of SVM and MAXENT is shown that validates that the ensemble results are more accurate than non ensembling.

Table 5.4: Comparative Study for SVM,GLMNETand ensembling

Parameters	SVM and GLMNET(emsembe)	SVM	GLMNET
Recall	0.94	0.83	0.82
Precision	0.83	0.92	0.91
Accuracy	95.3	90	91.5
F-Measure	0.91	0.83	0.82

In table 5.4 a comparison between the ensemling of SVM and GLMNET is shown. From the values of the defined parameters it is clear that the ensembling provide the better results.

Table 5.5: Comparative Study for SVM,TREE and ensembling

Parameters	SVM and TREE(ensemble)	SVM	TREE
Recall	0.90	0.83	0.82
Precision	0.81	0.92	0.90
Accuracy	94.2	90	91
F-Measure	0.89	0.83	0.82

In table 5.5 a comparison between the ensemling of classification algorithm and non-ensembling of classification algorithm is done. From the values of the defined parameters it is clear that the ensembling provide the better results because it emerges the benefits of two classification algorithms SVM and TREE also defined above.

Table 5.6: Comparative Study for MAXENT,GLMNET and ensembling

Parameters	MAXENT and GLMNET	MAXENT	GLMNET
Recall	0.92	0.84	0.82
Precision	0.83	0.92	0.91
Accuracy	95.4	91	91.5
F-Measure	0.88	0.84	0.82

In table 5.6 a comparison between the ensembling and non ensembling of algorithms i.e. MAXENT,GLMNET is showm. The results verify that the emsembled results are more accurate than non ensembling.

Table 5.7: Comparative Study for MAXENT,TREE and ensembling

Parameters	MAXENT and TREE(ensemble)	MAXENT	TREE
Recall	0.87	0.84	0.82
Precision	0.80	0.92	0.90
Accuracy	94.4	91	91
F-Measure	0.87	0.84	0.82

In table 5.8 a comparison between the ensemling of classification algorithms i.e. MAXENT AND TREE and non-ensembling of classification algorithm is done. The results of evaluation metics shows that the ensemble results are better than non ensembling.

Table 5.8: Comparative Study for MAXENT,TREE,GLMNET and ensembling

Parameters	MAXENT,TREE and GLMNET(ensemble)	MAXENT	GLMNET	TREE
Recall	0.92	0.84	0.82	0.82
Precision	0.85	0.92	0.91	0.90
Accuracy	95.4	91	91.5	91
F-Measure	0.88	0.84	0.82	0.82

In table 5.8 a comparison between the ensembled approach of classification algorithms i.e.MAXENT,GLMNET,TREE is done. From the evaluation results it is clear that the ensembling provide the better results because it emerges the benefits of two and more classification algorithms.

Table 5.9: Comparative Study for MAXENT,SVM,GLMNET and ensembling

Parameters	MAXENT, SVM and GLMNET(ensemble)	SVM	MAXENT	GLMNET
Recall	0.93	0.83	0.84	0.82
Precision	0.87	0.92	0.92	0.91
Accuracy	95.5	90	91	91.5
F-Measure	0.90	0.83	0.84	0.82

In table 5.9 a comparison between the ensembling of classification algorithms i.e. MXENT,SVM,GLMNET and non-ensembling of classification algorithm is shown with the evaluation metrics. From the values of the defined parameters it is clear that the ensembling provide the better results.

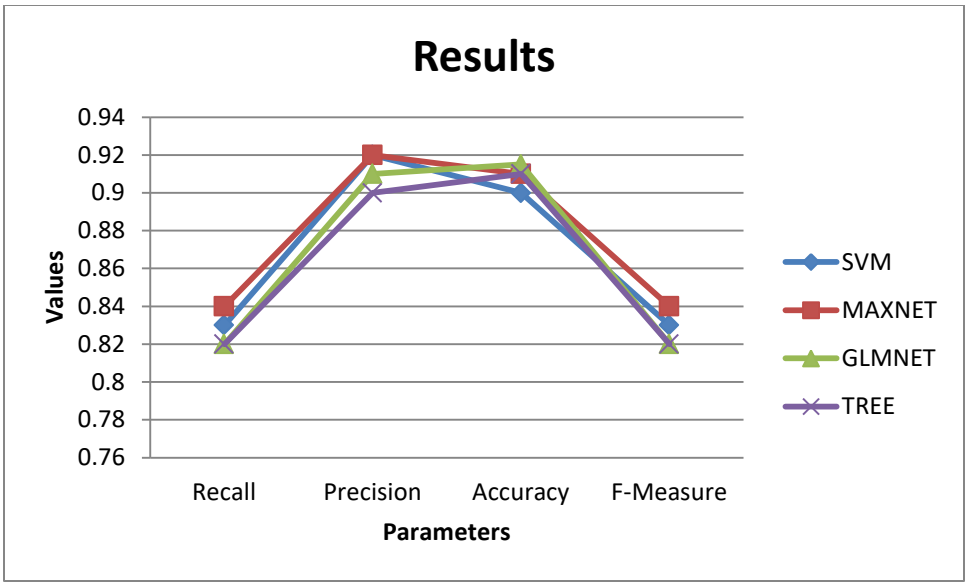

Figure 5.2: Comparative Study without ensembling of algorithms

In figure 5.2 a comparison between the classification algorithm is done. From the values of the defined parameters it is clear that the SVM and MAXENT performed better than that of other algorithms.

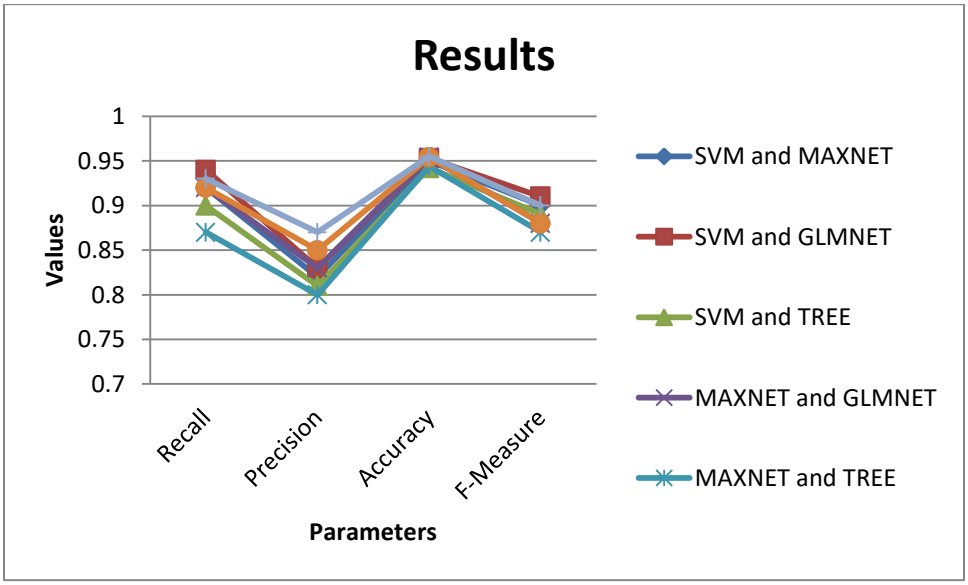

Figure 5.3: Comparative Study with ensembling of algorithms

In figure 5.3 a comparison between the ensembling of classification algorithm is done. From the values of the defined parameters it is clear that the ensembling of MAXENT, SVM, GLMNET performed better than that of other combination of algorithms.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Text mining is one of the quickest developing fields today. The ensembling approaches in the text document for keyword extraction approach are far better than that of statistical approaches. With the proposed ensemble approach for keyword extraction the results for evaluation metrics that are accuracy, recall, f-measure and precision are improved.. It is also easy in the case of proposed ensemble approach to extract the keyword from the natural language processing(NLP) and find the data from the text documents on the behalf of extracted ketwords which are extracted from the NLP using ensembling approach.

6.2 Contribution

In this research, we attempted to introduce an outline of content mining approach with its strategies, instruments and applications. In this approach the keyword extraction is done based upon the NLP techniques to fetch the data from the multiple documents. In this approach keywords that are extracted from the NLP text or can say string are compared with the multiple documents in the database based upon the cosine similarity. On the basis of frequency of keyword in the document and cosine similarity the ranking and score are given to documents from which the data is to be extracted. In the end ensembling on keywords is achieved based upon the classification algorithms that are SVM, MAXNET, GLMNET and TREE.

6.3 Future Scope

In future, one can target following direction in the field of summarization:

- Text summarization in low resourced languages especially in Indian language context such as Telugu, Hindi, Tamil, Bengali, etc.
- This work can also be extended to multi-lingual text and multimedia summarization.

References

[1] Blossom Manchanda, Vijay Anant Athavale, “Various Statistical Techniques Used in NLP,” International Journal of Computer Applications & Information Technology, Vol. 9 9(1), pp. 172-176, 2016.

[2] Nihar Ranjan, Kaushal Mundada, Kunal Phaltane, Saim Ahmad, “A Survey on Techniques in NLP,” International Journal of Computer Applications, Vol. 134(8), pp. 6-9, 2016.

[3] Joakim NIVRE, “On Statistical Methods in Natural Language Processing,” School of Mathematics and Systems Engineering.

[4] Blossom Manchanda, Vijay Anant Athavale, “Various Statistical Techniques Used in NLP,” International Journal of Computer Applications & Information Technology, Vol. (1), pp. 172-176, 2016.

[5] Danqi Chen, Richard Socher, Christopher D. Manning, Andrew Y. Ng, “Learning New Facts From Knowledge Bases With Neural Tensor Networks and SemanticWord Vectors,” Proceedings of the International Conference on Learning Representations (ICLR, Workshop Track), 2013.

[6] Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, “Bilingual Word Embeddings for Phrase-Based Machine Translation,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013.

[7] Karl Pichotta, Raymond J. Mooney, “Statistical Script Learning with Multi-Argument Events,” Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014.

[8] Stephen Roller, Sabine Schulte im Walde, “A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp.1146-1157, 2013.

[9] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, “YouTube2Text: Recognizing and Describing Arbitrary Activities Using

Semantic Hierarchies and Zero-shot Recognition", Proceedings of the 14th International Conference on Computer Vision , pp.2712--2719, 2013.

[10] Karl Pichotta, John DeNero, “Identifying Phrasal Verbs Using Many Bilingual Corpora", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 636-646, 2013.

[11] Shruti Bhosale, Heath Vinicombe, Raymond Mooney, “Detecting Promotional Content in Wikipedia,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1851-1857, 2013.

[12] Dan Garrette, Jason Mielens, Jason Baldridge, “Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages,” Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.583--592, 2013.

[13] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, Sergio Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge,” Proceedings of the NAACL HLT Workshop on Vision and Language, pp.10-19, 2013.

[14] Sindhu Raghavan, Raymond J. Mooney, “Online Inference-Rule Learning from Natural-Language Extractions,” Proceedings of the 3rd Statistical Relational, 2013.

[15] Varun Chandola, Eric Eilertson, Levent ErtÄoz, GyÄorgy Simon and Vipin Kumar, “Data Mining for Cyber Security,” Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.

[16] Mahesh Kumar Kond Reddy, Sujeeth .T, “Data Mining Tool using Clustering Technique on Exploration Engine Dataset,” Int. Journal of Engineering Research and Application, Vol. 3(5), pp.2032-2036, 2013.

[17] Hamzeh Agahi, A. Mohammadpour, S. Mansour Vaezpour, “Predictive tools in data mining and k-means clustering: Universal Inequalities,” Vol. 63(3), pp. 779-803, 2013.

[18] Ahmed Elgohary, Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray, “Embed and Conquer: Scalable Embeddings for Kernel k-Means on MapReduce,” 2013.

[19] Chunfei Zhang, Zhiyi Fang, “An Improved K-means Clustering Algorithm”, Journal of Information & Computational Science Vol.10, pp.1193–199, 2013.

[20] Christopher Ndehedehe, Ogunlade Simeon,1Akwaowo Ekpa, “Spatial Image Data Mining Using K-Means Analysis: A Case Study of Uyo Capital City, Nigeria,” International Journal of Advanced Research, Vol. 1(7), pp.6-15, 2013.

[21] M.Sakthi, Antony Selvadoss Thanamani, “An Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm,” International Journal of Recent Technology and Engineering (IJRTE), Vol.2(3), 2013.

[22] Muhammad Rukunuddin Ghalib, Shivam Vohra, Sunish Vohra, Akash Juneja, “Mining On Car Database Employing Learning and Clustering Algorithms,” International Journal of Engineering and Technology (IJET),Vol.12, 2014.

[23] Asmita Yadav, “A Survey Of Issues And Challenges Associated With Clustering Algorithms,” International Journal for Science and Emerging, Technologies with Latest Trends, Vol.10 (1), pp.7-11, 2013.

[24] Soumi Ghosh, Sanjay Kumar Dubey, “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms,” International Journal of Advanced Computer Science and Applications, Vol. 4(4), 2013.

[25] Kavita OZA, Rajanish KAMAT, “Applying Data Mining for Framing of Computer Science Curriculum”, Proceedings of the IETEC’13 Conference, Ho Chi Minh City, Vietnam, 2013.

[26] Manpreet Kaur, Usvir Kaur, “A Survey on Clustering Principles with K-means Clustering Algorithm Using Different Methods in Detail”, IJCSMC, Vol. 2(5), pp. 327 – 331, 2013.

[27] Ratnadip Adhikari, R. K. Agrawal, “A Homogeneous Ensemble of Artificial Neural Networks for Time Series Forecasting,” International Journal of Computer Applications, Vol. 32(7), pp.1-9,2011.

[28] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, Carlos A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” Expert Systems with Applications, Vol. 77, pp. 236–246, 2017.

[29] Slobodan Beliga, Ana Meštrović, Sanda Martin, “An Overview of Graph-Based Keyword Extraction Methods and Approaches,” JIOS, Vol. 39(1), pp. 1-20, 2015.

[30] Gunes Erkan, Dragomir R. Radev, “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization,” Journal of Artificial Intelligence Research, vol.22, pp. 457-479, 2004.

[31] Florian Boudin, “A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction,” International Joint Conference on Natural Language Processing, pp.834-838,2013.

[32] Marina Litvak, Mark Last, Hen Aizenman, Inbal Gobits, Abraham Kandel, “DegExt – A Language-Independent Graph-Based Keyphrase Extractor”, Vol. 12, pp.856-870, 2007.

[33] Vairaprakash Gurusamy, Subbu Kannan: “Preprocessing Techniques for Text Mining,” 2014.

[34] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya: “Preprocessing Techniques for Text Mining – An Overview,” International Journal of Computer Science & Communication Networks, Vol. 5(1), pp.7-16, 2010.

[35] C. Ramasubramanian, R.Ramya: “Effective Pre Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2(12), 2013.

[36] Parth Suthar, Prof. Bhavesh Oza: “A Survey of Web Usage Mining Techniques,” International Journal of Computer Science and Information Technologies, Vol. 6(6), pp.5073-5076, 2015.

[37] “Data Preprocessing Techniques for Data Mining,” Winter School On “Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets, 2012.

[38] Vikram Singh and Balwinder Saini “An Effective Pre Processing Algorithm For Information Retrieval Systems,” International Journal of Database Management Systems (IJDMS) Vol.6(6), 2014.

[39] Moral, C., de Antonio, A., Imbert, R. & Ramírez, J. (2014). “A survey of stemming algorithms in information retrieval/ Information Research,” Vol.19(1) paper 605. [Available at http://InformationR.net/ir/191/paper605.html]

[40] Shilpa Dang, Peerzada Hamid Ahmad, “A Review of Text Mining Techniques Associated with Various Application Areas,” International Journal of Science and Research, Vol. 4(2), pp. 2461-1466, 2015.

[41] Danqi Chen, Richard Socher, Christopher D. Manning, Andrew Y. Ng, “Learning New Facts From Knowledge BasesWith Neural Tensor Networks and SemanticWord Vectors,” Proceedings of the International Conference on Learning Representations (ICLR, Workshop Track), 2013.

[42] Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, “Bilingual Word Embeddings for Phrase-Based Machine Translation,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.

[43] Karl Pichotta, Raymond J. Mooney, “Statistical Script Learning with Multi-Argument Events,” Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014.

[44] Stephen Roller, Sabine Schulte im Walde, “A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp.1146--1157, 2013.

[45] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, “YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition,” Proceedings of the 14th International Conference on Computer Vision , pp.2712--2719, 2013.

[46] Karl Pichotta, John DeNero, “Identifying Phrasal Verbs Using Many Bilingual Corpora,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp.636--646, 2013.

[47] Shruti Bhosale, Heath Vinicombe, Raymond Mooney, “Detecting Promotional Content in Wikipedia,” Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp.1851--1857, 2013.

[48] Dan Garrette, Jason Mielens, Jason Baldridge, “Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages,” Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.583--592, 2013.

[49] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, Sergio Guadarrama, “Generating Natural-Language Video Descriptions Using Text-Mined Knowledge,” Proceedings of the NAACL HLT Workshop on Vision and Language, pp.10-19, 2013.

[50] Sindhu Raghavan, Raymond J. Mooney, “Online Inference-Rule Learning from Natural-Language Extractions,” Proceedings of the 3rd Statistical Relational, 2013.

[51] Varun Chandola, Eric Eilertson, Levent Ertaoz, Gyaorgy Simon and Vipin Kumar, “Data Mining for Cyber Security,” Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.

[52] Mahesh Kumar Kond Reddy, Sujeeth .T, “Data Mining Tool using Clustering Technique on Exploration Engine Dataset,” Int. Journal of Engineering Research and Application, Vol. 3(5), pp.2032-2036, 2013.

[53] Hamzeh Agahi, A. Mohammadpour, S. Mansour Vaezpour, “Predictive tools in data mining and k-means clustering: Universal Inequalities,” Vol.6 (4), pp. 779-803, 2013.

[54] Ahmed Elgohary, Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray, “Embed and Conquer: Scalable Embeddings for Kernel k-Means on MapReduce,” 2013.

Publications

Bhavneet Kaur and Dr. Sushma Jain, “Keyword Extraction Using Machine Learning Approaches,” *ICACCA- International Conference on Advances In Computing,Communication & Automation,* 2017 [Accepted].