An Effective Approach for Semantic Text Analysis using SVM

Swati Gautam
Department of Computer Science
Radharaman Institute of Technology & Science
Bhopal, India
swati.gautam026@gmail.com

Hansa Acharya
Department of Computer Science
Radharaman Institute of Technology & Science
Bhopal, India
hansaacharya@gmail.com

Prof. Anurag Jain
Department of Computer Science
Radharam Institute of Technology & Science
Bhopal, India
anurag.akjain@gmail.com

Abstract — Hence an efficient model for the text analysis is proposed using a supervised learning approach such as SVM for the clustering of text having similar semantics at word level or sentence level. The text for the analysis using semantic models needs tagging based approach and method on words and sentences at the single gram and n gram and then semantic similarity can be calculated along with the co-occurrence between words and a pair of words. Finally these pairs of words are trained and clustered using a supervised learning approach to get classification of sentence polarity i.e. positive or negative. The existing technique implemented doesn't provide effective refinement of lexicons and also doesn't provide higher accuracy and co-relation matrix for the text. The proposed methodology provides an effective model for the analysis of texts such as sentiment words.

The proposed methodology implemented here for the Text Analysis provides higher accuracy as compared to the other existing technique implemented. The result analysis shows the performance of the proposed methodology. The experimental results are performed on different datasets such as ANEW Dataset and BAWL-R Dataset and SemEval2007 Dataset. The results of the proposed methodology is then performed for various seeds values and provides efficient results as compared to the technique implemented. The graphs shows the comparison between existing and the proposed work.

Index Terms— Information Retrieval, Natural Language Processing

I. Introduction

The indispensable confirmation of a textual matching model has to be recognized in unambiguous functional domains that make available methodologies and metrics for assessment. The rising quantity of textual data obtainable in electronic appearance is an main motivation for the search of well-organized techniques in the wide-ranging area of textual data looking are at in particular, Information Retrieval (IR).The most important purpose of IR is to proficiently recognize pertinent documents in a database, assuring an information could do with communicated by a user in a type of a query.

Semantic similarity of the words is used specially in searching operations. The survey of semantic similarity among words has been a part of natural language processing and information retrieval. Accurately measure the linguistics similarity between words is a very important downside in internet mining, data retrieval, and language process. Semantic similarity is defined as the similarity of two concepts as the maximum information content of the concept that subsumes them in the taxonomy hierarchy. The information content of a concept depends on the probability of encountering an instance of the concept in a corpus. The information content is then defined as negative the log likelihood of the probability.

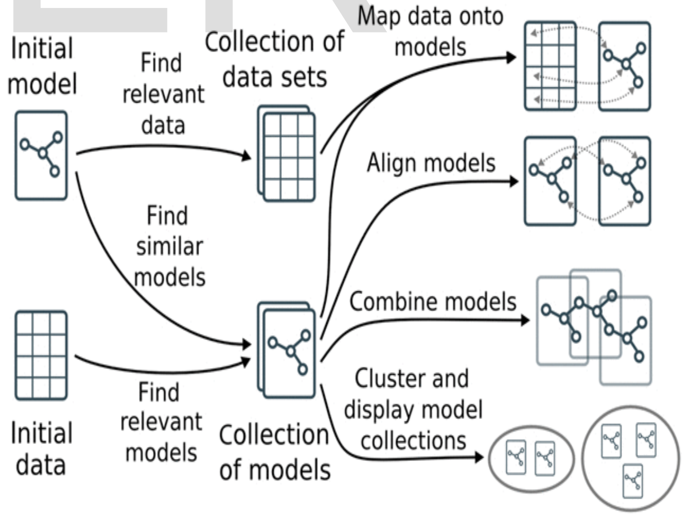


Figure 1. Architecture of the Semantic Model Analysis

Text Mining and Data Mining

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text [1]. The problem, of course, is that the information is not couched in a manner that is amenable to automatic dispensation.

Text summarization

A text summarizer strives to produce a condensed representation of its input, intended for human consumption [2]. It may condense individual documents or groups of documents. Text compression, a related area and are also condenses documents, but summarization differs in that its output is intended to be human-readable. As a field, summarization differs from many other forms of text mining in that there are people, namely professional abstractors, who are skilled in the art of producing summaries and carry out the task as part of their professional life. Studies of these people and the way they work provide valuable insights for automatic summarization.

Text Categorization (TC)

Text Categorization aims to automatically assign most suitable category labels from the available predefined set of labels to the unseen documents. Text categorization (or text classification) is the assignment of natural language documents to pre-defined categories according to their content [3]. The Library of Congress Subject Headings (LCSH) is a comprehensive and widely used controlled vocabulary for assigning subject descriptors. They occupy five large printed volumes of 6,000 pages each—perhaps two million descriptors in all. The aim is to provide a standardized vocabulary for all categories of knowledge, descending to quite a specific level, so that books—on any subject, in any language—can be described in a way that helps librarians retrieve all books on a given subject [4].

II. Problem Statement & Proposed Methodology

Problem Statement

The main problem that includes during the working of the existing technique for the text analysis is as follows:

1. The accuracy for the prediction and analysis of text depends on the number of seed words selected.
2. The technique implemented doesn't provide effective refinement of lexicons.
3. The technique is not implemented for the higher order n-gram ratings.
4. The technique is not implemented for semantically all types of languages.
5. Doesn't provide higher accuracy and co-relation matrix for the text.

Proposed Methodology

The upcoming rehearsal practical here is based in the concept of applying supervised learning approach such as Support vector machine for the Affective Analysis of Texts.

1. **Input Web log dataset:** The input dataset will be a set of words along with some of the sentences and a series of correlated words whose similarity to be measures. Here the analysis is done for a number of datasets. Each of the dataset contains a set of seed words and each of the seed word contains occurrence or valence, supremacy value and normal deviation which are used for the analysis of finding similarity amid words and co-occurrence of words in the document.
2. **Word Level Tagging:** Find the word level tagging of the word whose semantic similarity is to be finding: Here in this methodology the sentences where the input word is detected tagging is done to remove the contents which are of no use to amount the resemblance. Now in this procedure numerous resemblance metrics are functional to degree the comparison amongst words. Here Co-occurrence based similarity and context based similarity metrics is applied.

Word level tagging is used to find the affectivity of the word in the document. By finding word level tagging we compute the valence of the word from a set of seed words. The below formula is used to compute the valence of the word. For computing the valence of the word in the text document we need to find the similarity of the word from a set of seed words

Here the name close category for the disagreements 'W' can be computed using,

$$\breve{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i v(w_i) f(d(w_i w_j))$$

Where, Wj is the word whose tagging is to be done and let us suppose the dataset contains 'N' stone disagreements i.e w1,w2,w3……wN and 'v' be the valence of the expression and d(wi,wj) be the semantic similarity between two words which can be computed using,

$$J(w_i,w_j) = \frac{|D; w_i,w_j|}{|D; w_i| + |D; w_j| - |D; w_i,w_j|}$$

The above formula is used to compute the similarity using jaccard coefficient. Here D denoted document and wi denoted first word and wj signifies second word. Now D;wiwj denoted the words in the whole document and D;wi is the occurrence of the first word in the document and D;wj is the occurrence of word in document and removal of the both words in the document.

Where, J be the semantic similarity between two words using Jaccard Co-efficient.

The semantic similarity can also be computed using,

$$C(w_i,w_j) = \frac{2|D; w_i,w_j|}{|D; w_i| + |D; w_j|}$$

3. After calculating words level tagging of the word 'W' from a set of seed words w1,w2,w3……wN from the document 'D', sentence level tagging can be computed.

The core modification amongst name level tagging and sentence level tagging is that in word level tagging tagging of each word in the text document is done and in sentence level tagging word in the whole sentence of the document is done.

4. Fusion of the word level, sentence level, and multiword level tagging of the words: Likewise sentence level tagging and multi word level tagging is done so that the semantic word can be measured in every aspect. The three tagging techniques are then fused together to get the final semantic words.

Now Integration is done for all the tagging, means if the features are not present in word level and some features are present in sentence level tagging then after fusion all the features are present in the similarity.

5. Apply SVM based clustering to measure the similar words: these semantic words are then trained clustered using SVM so that that the semantic similarity is measures.

Consider training sample, containing input and output to be performed.

$$W_0^T X_i + b_0 \geq +1, for\ d_i = +1$$

$$W_0^T X_i + b_0 \leq -1, for\ d_i = -1$$

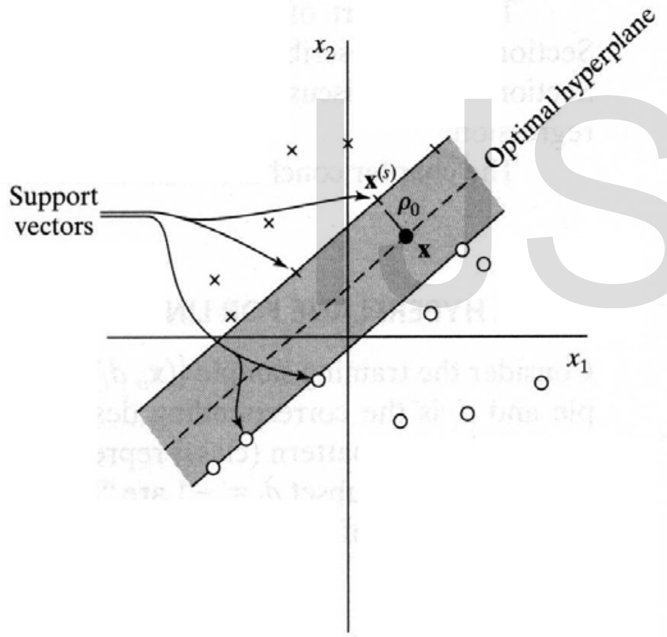


Figure 2 Basic Architecture of SVM

The statistics opinion which is very close is called the margin of separation

The main aim of using the SVM is to find the particular hyperplane of which the margin is maximized.

Optimal hyperplane

For example, if we are choosing our model from the set of hyper planes in P*n*, then we have:

$$f(x; \{w; b\}) = \mathrm{sign}(w \cdot x + b)$$

$$R_{emp}(\alpha) = \frac{1}{m}\sum_{i=1}^{m} l(f(x_i, \alpha), y_i)$$

Support vector Clustering

It is a supervised leaning approach which is used for the grouping of data on the basis of Input ‘X’ and Output ‘Y’.

Annotations Used

Xi – input values

Yi – labels according value of Xi

C – Class Index

G – Gamma Co-efficient

Oi – Optimization Parameter

M(y) – Margin Width

W – word containing matrix

```
1.  Initially Support vector learning consists of input and their distinct indicators (x1, y1), (x2, y2), (x3,y3),……..(xn,yn) and Class C and Gamma Coefficient ‘G’.
2.  Optimization Parameter $O_i \leftarrow \emptyset, \in 1,2,\ldots\ldots n$
3.  Repeat
4.  For i=1 to n do
5.  Compute Margin for linear kernel using
    $M(y) = \Delta(y_i, y) + W^T X(x_i, y) - W^T X(x_i, y_i)$
6.  Compute the maximum marginal width of the Support learning using
    $Y = Max(M(y))$
7.  If M (y) > Y then
8.  Group the value into one cluster using
    $O_i = O_i \cup \{Y\}$
9.  End if
10. End for
11. Untill no $O_i$ has changed during iteration.
```

Here supervised learning algorithm is implemented using SVM based clustering.

III. Flow Chart of Proposed Work

The figure shown below is the flow chart of the proposed methodology. Here in the flow chart first of all web log dataset is taken on which the tagging is done. The tagging applied here is on the basis of Word level, sentence level and Multi-word level based.

Input Web log Dataset

Affective Norms for English Words (ANEW) Dataset: The Dataset contains a number of words along with their valence range and dominance factor. The words used are mostly related with emotions and moods. It was collected by Bradley and Land in 1994 for 1034 words and is used for the text analysis.

Berlin Affective Word List Reloaded (BAWL-R) Dataset: The dataset mainly contains German words. It contains a list of over 2900 German words which are taken from CELEX Database and represents words or sentence polarity of Positive, Negative or Neutral valencies.

SemEval 2007 Dataset: The Dataset contains annotate text for the analysis of emotion words i.e joy, surprise, fear and also used for the sentence polarity orientation Positive or Negative. The Corpus in SemEval 2007 is taken from news websites such as CNN or Google and contains News Headlines.

Word Level tagging

Since the main purpose of text mining is for finding of the text comes into positive or negative category. Hence a valence is decided between [+1, -1] for the very positive to very negative. Word level tagging is used to find the affectivity of the word in the document. By finding word level tagging we compute the valence of the word from a set of seed words. The below formula is used to compute the valence of the word. For computing the valence of the word in the text document we need to find the similarity of the word from a set of seed words.

Here the word level tagging for the words 'W' can be computed using,

$$\breve{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i v(w_i) f(d(w_i w_j))$$

Where,

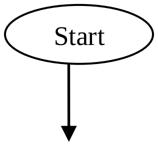
Wj is the word whose tagging is to be done and let us suppose the dataset contains 'N' seed words i.e w1,w2,w3……wN and 'v' be the valence of the word and d(wi,wj) be the semantic similarity between two words which can be computed using,

Sentence Level Tagging

Let us suppose that the document consists of sentence 'S' which contains a number of words w1, w2, w3…..wn.

$$V_a(s) = b_0 + b_1 \frac{1}{N} \sum_{i=1}^{N} v(w_i)$$

where and are trainable weights corresponding to an offset and unigrams respectively. Linear fusion assumes that words should be weighted equally independently of their strong or weak affective content. As a result, a sentence containing only a few strongly polarized terms might end up having low absolute valence (due to averaging).



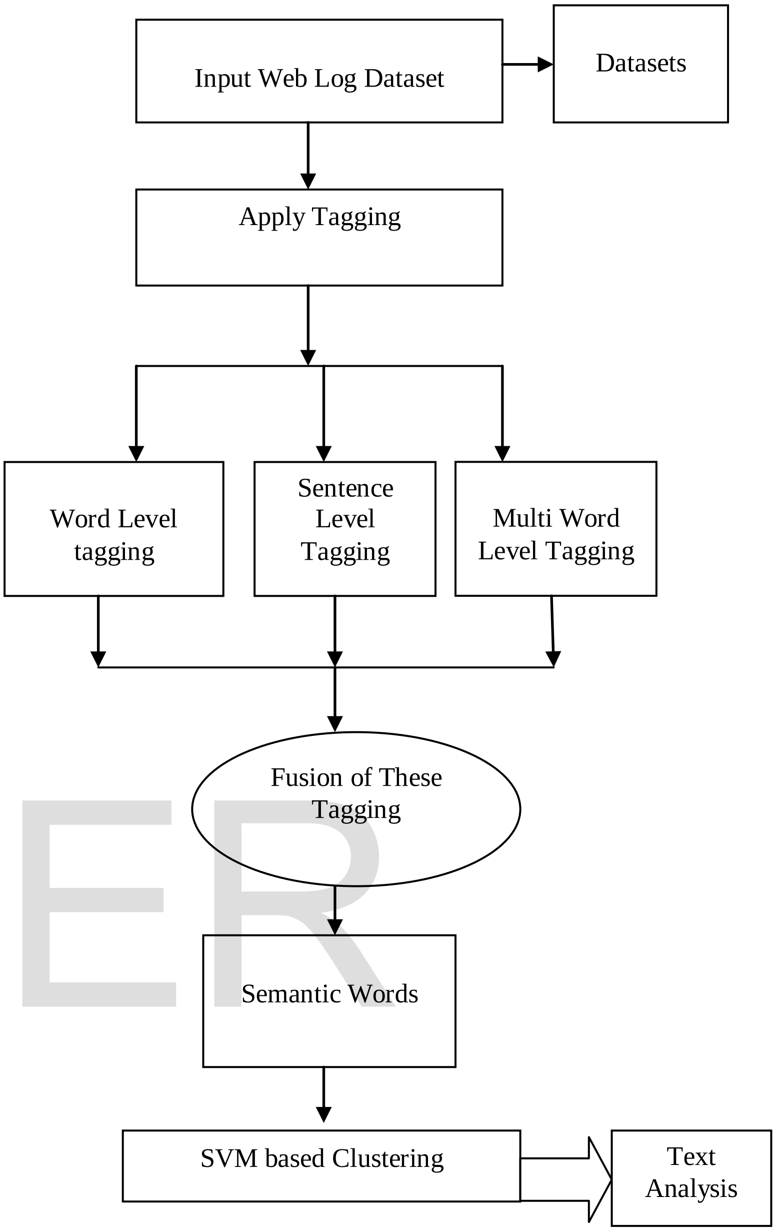


Figure 3 Flow Chart of the methodology

Multi Word Level tagging

Since the main purpose of text mining is for finding of the text comes into positive or negative category. Hence a valence is decided between [+1, -1] for the very positive to very negative. Multi Word level tagging is used to find the affectivity of the word in the document. By finding word level tagging we compute the valence of the word from a set of seed words. The below formula is used to compute the valence of the word. For computing the valence of the word in the text document we need to find the similarity of the word from a set of seed words.

Here the Multi word level tagging for the words 'W' can be computed using,

$$\check{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i v(w_i) f(d(w_i w_j)) + a_1 + \sum_{i=1}^{N} a_i v(w_i) f(d(w_i w_j))$$

Where,
Wj is the word whose tagging is to be done and let us suppose the dataset contains 'N' seed words i.e w1,w2,w3……wN and 'v' be the valence of the word and d(wi,wj) be the semantic similarity between two words which can be computed using,
Finally after applying tagging on the input web log dataset fusion is done and generate a set of semantic words and clustering is done using support vector clustering for the affective text analysis.

IV. Result Analysis

The table shown below in 1 is the Comparative analysis of Accuracy on BAWL-R dataset. The proposed methodology implemented here provides high accuracy as compared to the existing technique. BAWL-R dataset mainly contains German words. It contains a list of over 2900 German words which are taken from CELEX Database and represents words or sentence polarity of Positive, Negative or Neutral valencies.

No. of Seeds	Accuracy Existing Work	Accuracy Proposed Work
100	86	89
200	88	91
300	89.56	92.48
400	90	93.58
500	90.12	94
600	90.56	95
700	91.35	95.29
800	92	96

Table 1 Comparison of Accuracy for BAWL-R Dataset

The table shown below in 2 is the Comparative analysis of Accuracy on ANEW dataset. The proposed methodology implemented here provides high accuracy as compared to the existing technique. ANEW Dataset contains a number of words along with their valence range and dominance factor. The words used are mostly related with emotions and moods. It was collected by Bradley and Land in 1994 for 1034 words and is used for the text analysis.

No. of Seeds	Accuracy Existing Work	Accuracy Proposed Work
100	82.45	88
200	83	90
300	85	91
400	87	92
500	87.56	94
600	88	95.42
700	89.36	96
800	90	96.53

Table 2 Comparison of Accuracy for ANEW Dataset

The table shown below in 3 is the Comparative analysis of Accuracy on SemEval 2007 dataset. The proposed methodology implemented here provides high accuracy as compared to the existing technique. SemEval 2007 Dataset contains annotate text for the analysis of emotion words i.e joy, surprise, fear and also used for the sentence polarity orientation Positive or Negative. The Corpus in SemEval 2007 is taken from news websites such as CNN or Google and contains News Headlines.

No. of Seeds	Accuracy Existing Work	Accuracy Proposed Work
100	76.12	78.43
200	78.19	79.41
300	80.91	81
400	82.183	84.73
500	83.66	86.13
600	85	87.16
700	88.12	89.14
800	89	92.22

Table 3 Comparison of Accuracy for SemEval 2007 Dataset

The table shown below in 4 is the Comparative analysis of Co-relation on BAWL-R dataset. The proposed methodology implemented here provides high Co-relation as compared to the existing technique. BAWL-R dataset mainly contains German words. It contains a list of over 2900 German words which are taken from CELEX Database and represents words or sentence polarity of Positive, Negative or Neutral valencies.

No. of Seeds	Correlation Existing Work	Correlation Proposed Work
100	0.76	0.82
200	0.79	0.84
300	0.8	0.82
400	0.83	0.86
500	0.87	0.91
600	0.88	0.92
700	0.89	0.93
800	0.91	0.95

Table 4 Comparison of Co-relation for BAWL-R Dataset

The table shown below in 5 is the Comparative analysis of Co-relation on ANEW dataset. The proposed methodology implemented here provides high Co-relation as compared to the existing technique. ANEW Dataset contains a number of words along with their valence range and dominance factor. The words used are mostly related with emotions and moods. It was collected by Bradley and Land in 1994 for 1034 words and is used for the text analysis.

No. of Seeds	Correlation Existing Work	Correlation Proposed Work
100	0.73	0.79
200	0.77	0.82
300	0.78	0.83
400	0.81	0.84
500	0.84	0.86
600	0.86	0.9
700	0.89	0.92
800	0.9	0.93

Table 5 Comparison of Co-relation for ANEW Dataset

The table shown below in 6 is the Comparative analysis of Co-relation on SemEval 2007 dataset. The proposed methodology implemented here provides high Co-relation as compared to the existing technique. SemEval 2007 Dataset contains annotate text for the analysis of emotion words i.e joy, surprise, fear and also used for the sentence polarity orientation Positive or Negative. The Corpus in SemEval 2007 is taken from news websites such as CNN or Google and contains News Headlines.

No. of Seeds	Correlation Existing Work	Correlation Proposed Work
100	0.72	0.74
200	0.74	0.78
300	0.77	0.79
400	0.79	0.82
500	0.81	0.85
600	0.83	0.87
700	0.87	0.89
800	0.88	0.91

Table 2 Comparison of Co-relation for SemEval 2007 Dataset

The figure shown below in 3 is the Comparative analysis of Accuracy on ANEW dataset. The proposed methodology implemented here provides high accuracy as compared to the existing technique. ANEW Dataset contains a number of words along with their valence range and dominance factor. The words used are mostly related with emotions and moods. It was collected by Bradley and Land in 1994 for 1034 words and is used for the text analysis.

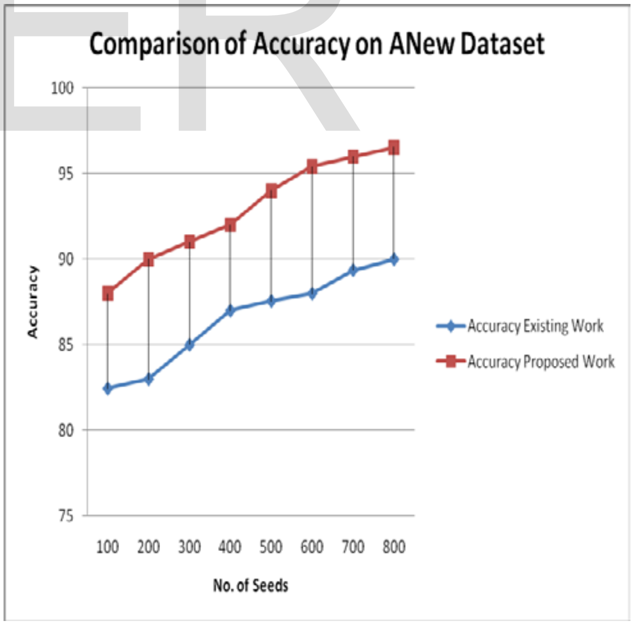


Figure 4 Analysis of Accuracy for ANEW Dataset

The figure shown below in 4 is the Comparative analysis of Accuracy on BAWL-R dataset. The proposed methodology implemented here provides high accuracy as compared to the existing technique. BAWL-R dataset mainly contains German words. It contains a list of over 2900 German words which are taken from CELEX Database and represents words or sentence polarity of Positive, Negative or Neutral valencies.

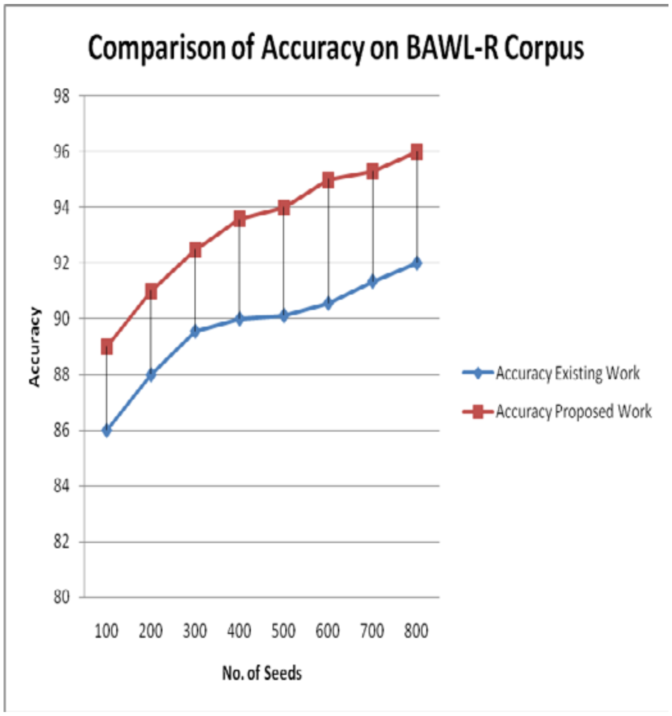


Figure 5 Analysis of Accuracy for BAWL-R Dataset

V. Conclusions

Here defined and revise a work on literature text mining trouble pass on to as proportional text mining. It has to do with find out any concealed frequent premises transversely a set of equivalent collected works on text as well as summarizing the relationship and differences of these collections beside each of the premises available. The previous technique come within reach of to sustain users in the search process do not deliberate on the level of semantic matching required for searching concepts data on Web. Here presents a concise preface to the different text representation proposals and classifiers utilized in the field of text mining. The obtainable techniques are measure up to and dissimilarity found on a variety of constraints to be precise criteria used for classification. Since the above argument it is implicit that no single representation method and classifier can be suggested as a common representation for any application.

The new technique implemented here for the sentence polarity provides high accuracy as compared to the other existing techniques of creating semantic model for the text discovery. The Experimental result analysis shows the performance of the proposed methodology. The techniques is tested on three dataset and comparative analysis of the existing and proposed is done.

The proposed methodology is implemented for various datasets such as SemEval 2007 and BAWL-R dataset and it was found that the proposed methodology provides efficient results as compared to the existing technique implemented for the text analysis.

References

[1]. Witten, I.H. and Frank, E.,"Data mining: Practical machine learning tools and techniques with Java implementations", Morgan Kaufmann, San Francisco, CA, 2000.

[2]. Mani, I.,"Automatic summarization. John Benjamins", Amsterdam, 2000.

[3]. Sebastiani, F. ,"Machine learning in automated text categorization", *ACM Computing Surveys, Vol. 34*, No. 1, pp. 1–47, 2002.

[4]. Witten, I.H. and Bainbridge, D. ,"How to build a digital library", Morgan Kaufmann, San Francisco, CA, 2003.

[5]. E. Gabrilovich and S. Markovitch. ,"Computing semantic relatedness using Wikipedia-based explicit semantic analysis", *In Proc. of the Inl. Joint Conference On Artificial Intelligence (IJCAI)*, 2007.

[6]. A. Budanitsky and G. Hirst. ,"Evaluating WordNet-based measures of semantic distance", *Computational Linguistics*, 32:13–47, 2006.

[7]. P. Resnik. ,"Using information content to evaluate semantic similarity in a taxonomy", *In Proc. Of International Joint Conference for Artificial Intelligence*, pages 448–453, 1995.

[8]. S. Banerjee and T. Pedersen,"An adapted Lesk algorithm for word sense disambiguation using WordNet", *In Proc. Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145, 2002.

[9]. Hassan and R.Mihalcea ,"Cross-lingual semantic relatedness using encyclopedic knowledge", *In Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, 2009.

[10]. Thomas Hofmann,"Unsupervised learning by Probabilistic Latent Semantic Analysis", *Machine Learning*, 42(1-2):177-196, 2001.

[11]. Jonathon Hare, Sina Samangooei, Paul Lewis, and Mark Nixon,"Semantic spaces revisited: Investigating the performance of auto-annotation and semantic retrieval using semantic spaces", *In Proceedings of CIVR*, pages 359-368, Niagara Falls, Canada, 2008.

[12]. Jeon Wook Kang, Hyun-Kyu Kang ,"A Term Cluster Query Expansion Model based on Classification Information", *Published in IEEE (2010) and 2010 International Conference on Artificial Intelligence and Computational Intelligence*, 2010.

[13]. Francisco Joo Pinto, Carme Fernndez Prez-Sanjulin,,"Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model", *Published in SISTEDES*, 2008.