International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

Sentiment Analysis of Book Reviews Using Semantic Network

Yogeshwaran T[1], Neena Jasmine S[2], Kishore Kumar. B[3], Saraswathi S[4]

[1,2,3,4]Information Technology, Pondicherry Engineering College, Pondicherry, India
[1]yogeshwarantj@gmail.com

Abstract: In new emerging technologies the processing of large textual data is challenging. Text mining can be defined as the process where exploring and analysing of enormous amounts of unstructured text data assisted by software that can identify patterns, concepts, keywords, topics and other attributes in handling the data. From the data, the main challenge is to identify the meaning of the text and correlate it to the text representation model for the sentiment classification. It is complex to capture the sentiments in the document level sentences to correctly classify them into positive, negative and neutral. In this work, we suggest a novel semantic network method using WuPlamer word similarity method of WordNet to compare the Part-Of-Speech (POS) tagged words of new documents to the manually obtained POS tagged words from the documents. This results in generating a similarity score, which is then used to classify the documents into three categories. We experimented to evaluate the different existing algorithms. The results show that the proposed approach allows for the precision of the classification when considering analysis of the document level.

Keywords: Sentiment Classification, Semantic Network, document level sentences, WuPlamer, POS tagger, WordNet.

1. INTRODUCTION

Recently the world online market functions based on the reviews given by the customers and determines the sales of products. The reviews which are posted by the reviewers generally have different polarity ranging from positive to negative. These reviews collectively leave an opinion for the product which is sold. These opinions also leave the manufacturer to shape their way to improve and inculcate innovations in future products so there is a need to analyse the review and obtain the feedback.

The review can also be available at different levels, it can be a one-line review or it can be a sentence or a detailed explanation about the product. The online reviews are mostly expressed in the form of text which is a major challenge. Hence, these reviews are needed to be classified to reveal the sentiments which are expressed. A semantic network is a knowledge base that is used in a network to describe the semantic relationships between the concepts. This is often used as a type of representation of the knowledge. It is a graph which can be either directed or undirected, comprising of vertices and edges which signify concepts and semantic relations between concepts, mapping or connecting semantic fields respectively. Semantic networks are gaining insights towards the unstructured form of text and are used to classify them.

The previous works in the semantic network where the semantic words are generated manually from the reviews and are used to classify the reviews are done only for the small size of the sentences but in the newly proposed method, we consider the document level reviews for the classification and the semantic words in the document is automatically generated using the POS tagger by considering all forms of verb, adjectives and conjugations. The WordNet is used to combine the English words into the set of synonym so the words which are tagged from the input are compared and the scores are generated by the WuPalmer formula are used to categorize the review as positive, negative and neutral. After the classification, the newly identified unrepeated tagged words are appended to the existing respective list. This increases the probability of identifying the polarity of the sentences accuracy.

2. LITERATURE SURVEY

Opinion mining or Sentiment analysis is the use of natural language processing, computational linguistics and text analytics to recognize and extract subjective knowledge from source materials. Opinion Mining is largely applied to social media and reviews for applications, ranging from marketing to customer service. Opinion Mining defines a speaker's mind set or a writer's overall contextual polarity with regard to any subject or paper.

Ricardo B. Scheicher et al. in their paper-**Sentiment classification improvement using semantically enriched information** propose a technique to improve the sentiment classification performance by using

semantically enriched information derived from domain expressions. An experiment was conducted in which various categorization algorithms were applied to three datasets composed of reviews on various products and services. The results showed that this technique improves the classification precision when dealing with reviews of a narrow domain. This experiment shows that the technique is suitable to reviews related to entities of the same nature [1].

In the paper-**The main factors affecting cultural exchange between Korea and China: A semantic network analysis based on the cultural governance perspective** by Sang Do Park et al. text data were collected based on the cultural exchange between the two countries and text mining was conducted. Semantic network analysis and ego network analysis were performed by selecting 80 key words from the data collected from Korea and China's digital portals. The frequency-inverse document frequency (TF-IDF) technique was introduced to supplement the drawbacks of assessing the value of terms based on their frequency. In this case, the higher the frequency of a word in a particular document and the lower the number of documents that contain the word, the higher the value of the TF-IDF[2].

Another paper-**Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit** by Minjoo Yoo et al. collected postings for bipolar and depressive disorders from subreddit communities and utilized them for semantic network analysis with four steps. From the first stage all posts were collected from Reddit subcommunities for bipolar and depressive disorders. Secondly, tokenization, word stemming and text cleaning improved the texts in the posts. The third stage is where word matrixes were used to create the pre-processed words. Semantic networks were organized according to word matrix. The texts relating to the important topics were finally explored [8].

The paper-**Dynamic semantic network analysis for identifying the concept and scope of social sustainability** written by Keeheon Lee and Hosang Jung establishes a data-driven method for recognizing the scope as well as the concept of social sustainability. The latest literature on social sustainability has been applied over time to dynamic semantic network research. By identifying shift points, it divides the diffusion of social sustainability in academia into some times. This research also periodically constructs semantic networks to explore the scope and definition in the perspective of network analysis. The scope of the concept depending on the properties of a semantic network, while the concept relies on the semantic network's central nodes. The scope is the range of topic that the concept covers. The range in the semantic network is mirrored in the network's structural properties [6].

Ismael Ali and Austin Melton in their paper-**Graph-Based Semantic Learning, Representation and Growth from Text: A Systematic Review** give a systematic review on the graph-based processes of Semantic Learning, Representing and Growth (SLRG) from the text. This enumerates a new branch inspired by the cognitive-semantics of graph-based SLRG modeling. Graph-based text representation is a various method to the Vector-Space-Model (VSM) or Bag-Of-Words (BOW), in the form of text graphs. Text-graphs learn, represent and develop semantics from the text, giving rise to semantics graphs that encodes non-noise terms as nodes and offered semantic relations as edges among the nodes. In semantic SLRG processes, the computational cognitive semantic modeling is represented in detail with its applications. The four primary cognitive models of semantic memory of Steyvers, CI, ICAN and SAT are discussed [3].

The paper-**An Improved Study of Multilevel Semantic Network Visualization for Analyzing Sentiment Word of Movie Review Data** by Hyoji Ha et al, define an approach for refining and visualizing an enormous amount of collective intelligence information with a multilevel sentiment network for intuitively and semantically understanding the relevant information. After the extraction and analysis of the sentiment words from the movie review data, a film network based on the similarities between the words was developed. The network shaped like this will appear as a representation of the network of multilevel sentiments. Results showed that this approach provided an enhanced cognitive experience for the user [4].

Bolanle A. Ojokoh and Olumide Kayode in their paper-**A Feature–Opinion Extraction Approach To Opinion Mining** propose a method that extracts feature and opinion pairs from online reviews, determines the polarity and strength of these opinions to summarize and determine the recommendation status of the customers' reviews. Experiments were carried out with online reviews of four different digital cameras. During the analysis of the product features extraction the features found in the data set were tagged manually and compared with those evaluated by the system. The SentiWordNet sentiment score for each opinionated word was used [10].

Another paper-**Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis** by Zhiping Hou et al. recognizes themes and compares differences in online travel reviews. A semantic association analysis was used to extract thematic words from reviews collected

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

from three major online travel agencies (OTAs) in China and construct a semantic association network. The key concept of semantic association analysis is to define semantics as a node by the co-occurrence of two terms in a phrase with high-frequency terms, often considering the frequency of high-frequency phrase co-occurrence as a relation between nodes. Analysis of the semantic association, statistical analysis of thematic terms, and visualization was carried out on the data collected [5].

3. PROPOSED WORK

In this proposal, the document level reviews are classified based on the similarity of words by the semantic network at first. Few sentences which are manually identified as similar sentences are collected as three different types from the datasets of book reviews and these documents are passed to pre-processing and POS tagger. Then the tagged words are maintained as a list which will be used to compare with the reviews which are given as input in the semantic network. In the semantic network the adverb, adjective, verb and negation words are taken into account by the same POS tagger and are connected as a graph by the semantic network in WordNet, then used to identify the similarity with the help of WuPlamer formula. Based on the similarity, the best score is obtained after comparing with other scores. The sentences are classified as positive, negative and neutral and stored.

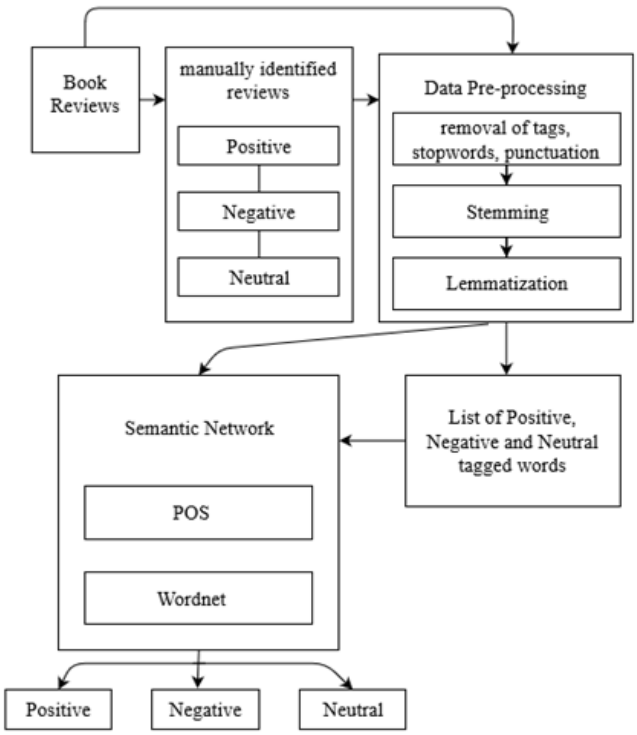

Figure 1. Proposed System Diagram

The proposed system can be divided into three components

A. Data Pre-processing
B. Obtaining Training set of words by POS tagger
C. Classifying Sentiments through the Semantic Network

A. Data Pre-processing

Data pre-processing improves identifying the specific details of data which is needed for analysis and increase the speed of processing.

In this module, the pre-processing of manually identified five documents each belonging to the positive, negative and neutral sentiments are given as input. Then these reviews are pre-processed to avoid the tags, stop words and punctuation. Stemming and lemmatization are done to identify the exact root meaning of the words. This helps to capture the exact sentimental words which are reflected in the review.

B. Obtaining Training set of words by POS tagger

A POS Tagger is used to read the text and allocate parts of the speech to each word or token, like noun, verb, adjective, etc.

The pre-processed documents of positive, negative and neutral are fed into the POS tagger to obtain tagged words such as verbs, adjectives, adverbs and negated words. These tagged words from each category are collected as a different list by avoiding repeated words.

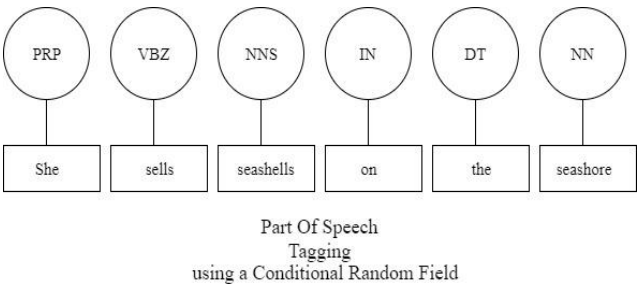

Figure 2: POS tagger

C. Classifying Sentiments through the Semantic Network

In this module, the list of POS tagged words are fed into the semantic network. The new reviews are given into the semantic network where once again the POS tagger is used to identify the words as adverbs, adjectives, verbs and negation then the tagged words are fed into WordNet where each of the words are connected and by synsets, their various synonyms are also considered and compared

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

with the list of tagged words and return a score for each word. The score for each word is calculated by WuPlamer formula.

The WuPlamer is used to calculate the relatedness by taking into account the depths of two synsets in the WordNet taxonomies, including the depth of Least Common Subsumer (LCS).

$$\mathbf{Wu-Palmer} = 2 * \frac{\mathbf{depth}\,(\mathbf{lcs}(\mathbf{s1},\mathbf{s2}))}{(\mathbf{depth}\,(\mathbf{s1}) + \mathbf{depth}\,(\mathbf{s2}))} \quad ...(1)$$

The score generated is 0<=score<=1 which can be represented in decimal. In the equation, the depth represents the similarity of each word compared with the words in the list. The score can be zero when there is no way of similar matching. This equation (1) calculates the similarity based on how similar the word senses are and where the synsets occur relative to each other in the hypmen tree.

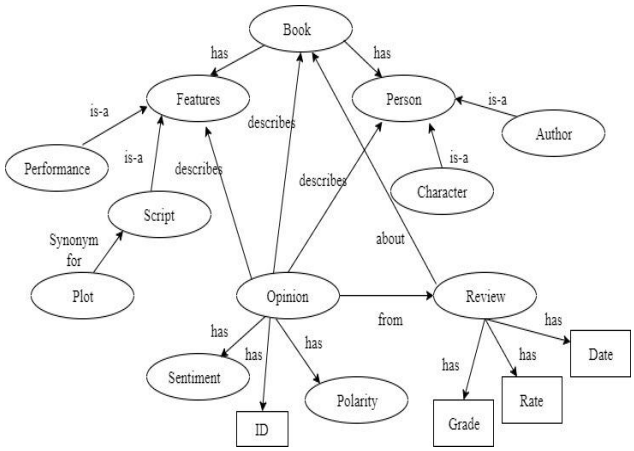

Figure 3. Hypmen Tree

Based on the generated three scores by comparing to three different lists, the best score is chosen and the input sentence is classified as positive, negative and neutral.

After classification, each word from the classified review is appended to the category of the list which is identified as output. Hence this improves the accuracy as the new words are included in the given list.

Table 1. Proposed Algorithm

Procedure steps
(1) Start
(2) From NLTK import stopwords, state_union, tokenize and WordNet
(3) stop_words = set(stopwords.words('english'))
(4) positivetag = [], negativetag=[], neutraltag=[] #list of sentiment words
(5) f = open (filepath) #open file containing reviews
(6) for x in f:
d=[] #empty list
remove stop words
tokenize words to d
for i in range(0,len(d):
tag the words : all forms verbs, adverbs, adjective
k.append(tagwords)

Table 2. WordNet to identify similarity

Procedure steps
allsyns1 = set(ss for word in sentences for ss in wordnet.synsets(word))
allsyns2 = set(ss for word in k for ss in wordnet.synsets(word))
allsyns3 = set(ss for word in sentences1 for ss in wordnet.synsets(word))
for s1,s2 in product(allsyns1, allsyns2):
best=wordnet.wup_similarity(s1, s2) # similarities identified
if(best==None):
best=0
else:
best=best
best=best*100
rounds=int(best)
now the best score are compared with each other
if(x1>x2 && x1>x3):
print type1 #positive sentence
taggedwords.append(positivetag) #the new tagged words are appended to list.
else if (x2>x3 && x2>x1):
print type2 #negative sentence
taggedwords.append(negativetag)
else
print type3 #netural sentence
taggedwords.append(neturaltag)
(7) Clear the lists and again run the for loop for next sentences
(8) End

Experimental Setup

This section presents the experiments conducted on a book review dataset as a case study to assess the performance of the proposed semantic method on the classification of text based on sentiment for improving

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 5 Issue: 3 March 2020

the performance of sentiment classification tasks. The experiments were conducted using a computer system with a processor (Intel® Core™ i5-4570 CPU @3.20 GHz 3.20 GHz), memory RAM 4.0 GB, hard disk 1TB and 64-bit operating system. Also, the Natural Language Toolkit package was installed in python ide spyder.

Dataset Description

A dataset of book reviews was used for the experiment, and this dataset was taken from amazon and used for sentiment analysis. The dataset contains 500 reviews of document level size ranging 8 GB of the total size. The reviews are pre-processed to do stemming to find the root word. Those reviews are processed via Parts Of Speech (POS) Tagger. The table (3) below shows the count of reviews which is taken for analyses.

Table 3. Dataset Description

S. No	Polarity of sentences	Count
1.	Positive	235
2.	Neutral	15
3.	Negative	60

4. RESULT ANALYSIS

The sentiment of document level reviews is classified as positive, negative and neutral and stored in different files. After each iteration, from the classified review, the keywords which are taken for comparison to the lists are appended to the available list and this provides to analyze the reviews exactly as the number of review increases. The graph below shows that the increases in the number of reviews will make the graph achieve a maturing stage where the words which are appended can be repeated and are not added.

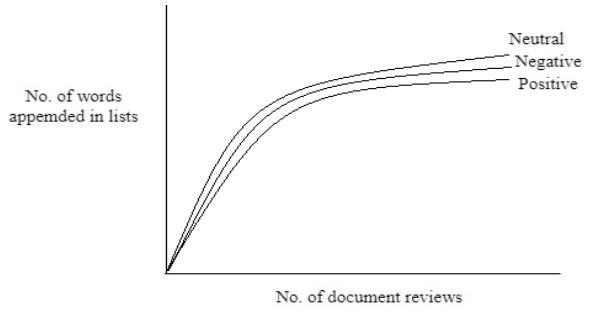

Figure 4. Graph Representing Number of Reviews

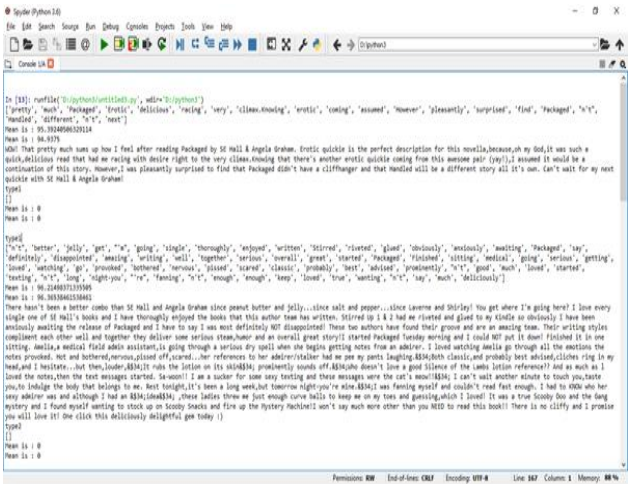

Figure 5. The semantic network identifies the types of documents based on similarity

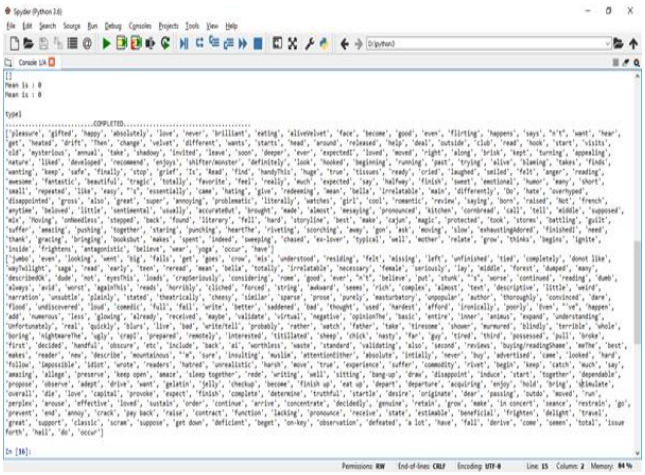

Figure 6. The set of corpus (verbs, adverbs and adjectives) collected as lists

5. CONCLUSION

In this project, we suggested a novel semantic network algorithm for classification of textual book reviews and this method is suitable for the document level reviews which are bigger in size and reduces the time for classification of documents. There are a few approaches which suggest the adoption of semantic network approach with the deep learning methods for classification tasks. So in future works, this proposed semantic network method can be used along with deep learning method as a hybrid classification approach. The proposed semantic network method can be applied with any classifier which directly depends upon the size of the dataset. This method can also be used to a large range of datasets in different domains for classification.

REFERENCES

[1] Ricardo B. Scheicher et al, "Sentiment classification improvement using semantically enriched information", in ACM Symposium on Document Engineering 2019 (DocEng '19), September 23–26, 2019, Berlin, Germany. ACM, New York, NY, USA, https://doi.org/10.1145/3342558.3345410

[2] Sang Do Park, Jong Youl Lee and Bo Wang, "The main factors affecting cultural exchange between Korea and China: A semantic network analysis based on the cultural governance perspective", International Journal of Intercultural Relations 71 (2019) 72–83, 24 May 2019 0147-1767/ © 2019 Elsevier, https://doi.org/10.1016/j.ijintrel.2019.04.005

[3] Ismael Ali and Austin Melton, "Graph-Based Semantic Learning, Representation and Growth from Text: A Systematic Review", 2019 IEEE 13th International Conference on Semantic Computing (ICSC), 978-1-5386-6783-5/19/$31.00 ©2019 IEEE DOI 10.1109/ICSC.2019.00027

[4] Hyoji Ha et al, "An Improved Study of Multilevel Semantic Network Visualization for Analyzing Sentiment Word of Movie Review Data", Appl. Sci. 2019, 9, 2419; doi:10.3390/app9122419

[5] Zhiping Hou et al, "Opinion mining from online travel reviews: A comparative analysis of Chinese major OTAs using semantic association analysis", 0261-5177/ © 2019 Elsevier, https://doi.org/10.1016/j.tourman.2019.03.009

[6] Keeheon Lee and Hosang Jung, "Dynamic semantic network analysis for identifying the concept and scope of social sustainability", 0959-6526/© 2019 Elsevier, https://doi.org/10.1016/j.jclepro.2019.05.390

[7] Livia Celardo and Martin G. Everett, "Network text analysis: A two-way classification approach", 0268-4012/ © 2019 Elsevier, https://doi.org/10.1016/j.ijinfomgt.2019.09.

[8] Minjoo Yoo, Sangwon Lee and Taehyun Ha, "Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit", 0306-4573/ © 2018 Elsevier, https://doi.org/10.1016/j.ipm.2018.10.001.

[9] Ying Xiong, Moonhee Cho and Brandon Boatwright, "Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement", 0363-8111/ © 2018 Elsevier, https://doi.org/10.1016/j.pubrev.2018.10.014

[10] Bolanle A. Ojokoh and Olumide Kayode, "A Feature–Opinion Extraction Approach To Opinion Mining", Journal of Web Engineering, Vol. 11, No. 1 (2012) 051-063 © Rinton Press

BIOGRAPHIES

Yogeshwaran. T, He is a student in Pondicherry Engineering College, pursuing B.Tech in the department of Information Technology.

Neena Jasmine. S, She is a student in Pondicherry Engineering College, pursuing B.Tech in the department of Information Technology.

Kishore Kumar. B, He is a student in Pondicherry Engineering College, pursuing B.Tech in the department of Information Technology.

Dr Saraswathi S, is a Professor in Pondicherry Engineering College of the Information Technology department, India. She received her PhD in Computer Science and Engineering from Anna University, Chennai, India in 2008. She completed her Master of Technology in Pondicherry University, India and received her Bachelor of Technology in Pondicherry Engineering College.