International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

# AUTOMATIC EMOJI LEXICON CONSTRUCTION FOR SENTIMENT ANALYSIS USING WORD KNOWLEDGE

**P.Akshaya[1], Mrs.K.Krishnakumari[2]**

*[1]M.E Student, Department of CSE, A.V.C College of Engineering, Tamil Nadu, India*
*[2]Associate Professor, Department of CSE, A.V.C College of Engineering, Tamil Nadu, India*

---

***Abstract -*** *In the field of sentiment analysis, emoji has received only little attention. Now-a-days, the users also share or express their opinions through emoji's and without any text. Most of the existing approaches use manually labeled dataset and train their classification system. It requires more training data, which is an expensive process. So, here the system proposed is lexicon-based sentiment analysis for word-level using lexicon graph with databases like WordNet and WNRH. In this approach, relatedness score helps to identify fine grained sentiment analysis. Thus, the emoji sentiment scores are calculated using co-occurrence frequency between emoji and sentiment words. Thus, the experimental results show this analysis provides unsupervised way of finding relatedness score for both text and emoji.*

***Key Words:*** **Sentimental Analysis, graph-based approach, lexical-based, Emoji, WordNet**

## 1. INTRODUCTION

Sentiment Analysis (SA) is the process of extracting sentiment such as positive, negative and neutral from a given dataset, which may contain the user reviews about the product or services collected via the social media like twitter, Facebook, blogs, etc. SA has two sub-tasks, which are data acquisition and data pre-processing. The data acquisition is the collection of reviews or feedback forms from the websites or social media networks. The data pre-processing step includes the tokenization, stemming, stop word removal, etc. SA helps to save the large amount of processing time of unstructured data instead of manual data processing. Now-a-days, the real time sentiment analysis is done, to know about the user real emotion in the political issues, product, services and sales marketing. The sentiments can be extracted by means of different levels such as word, phrase, sentence, document and aspect based from the source.

### 1.1 Lexical Sentiment Analysis

The LSA method utilize the Natural Language Processing and Machine Learning techniques and the automated tools to extract the emotions such as positive, negative, neutral in the textual data from user feedbacks and reviews in the web platforms. In recent approaches, the automatic creation of sentiment corpora techniques used. The LSA can be classified as Corpus-based approach, which is the data-driven method and lexicon-based approach, which is the knowledge-driven method. The corpus-based approach is used to analyze the large text corpora e.g., ISEAR. It is used to identify the probability of occurrence of textual features such as lexical forms, POS tags, n-grams or phrasal patterns and it enables sentiment predictions for new texts. But it is generally data hungry and it requires a considerable amount of manual effort, to produce a relevant sense-annotated corpus. The lexicon-based approach gets the sentiment clues from the readily available sentiment lexicon. The sentiment lexicon is nothing but contains the list of words or phrases which expresses its sentiments such as positive and negative sentiments. For example, the sentiment lexicon is available for 81 languages in Kaggle website and Senti-WordNet, WNA, etc. Due to the huge manual effort in the corpus-based approach, the lexicon-based approach is widely used for research purpose lately. Generally, the sentiment extraction can be processed in different granular levels. They are by processing of a word, a phrase, a sentence, a document and aspect. The Phrase sentiments deduced from word-level sentiments whereas the sentence-level lexical sentiment extraction based on either the word-level or phrase-level LSA. The word-level LSA is the fine-grained approach, where each word is associated with sentiment categories. And it allows LSA to utilize the text at higher granularities. The words can be processed by means of either separately i.e., standalone method or by considering their textual surroundings of the context. Some other processes are by considering the domain or

topic or from author's perspective. The word or phrase-level LSA usually gives the single opinion such as positive or negative i.e., happy or excited. The document or sentence may contain opposing opinions.

### 1.2 Emoji Sentiment Analysis

The effect of Emoji in text mining and sentiment analysis plays an important role as the usage of Emoji characters on social networks increasing. *I*n sentiment analysis , the utilization of Emoji characters results in higher sentiment scores. It is observed that the usage of Emoji characters in sentiment analysis appeared to have higher impact on overall sentiments of the positive opinions in comparison to the negative opinions. The emoji are considered as NLP perspective in constract to their predecessors i.e., the emoticons. Most of the emoji sentiment lexicons are done using the supervised methods which leads to wastage of time by manual labelling and training period. Mostly, the sentiment prediction is done for three categories in general such as positive, negative, neutral.

### 1.3 Scope of the Project

While using the lexicon-based LSA based techniques most of them suffers from the ambiguity problem in it. It is possible by the lexicon graph created using the WordNet and WordNet Relatedness Hierarchy (WNRH). Because, the WNRH is a independent hierarchy of relatedness categories, which is used to navigate between relatedness categories. Here, it also used to find out the relatedness score considering only their relatedness connections between relatedness nodes. This proposed LSA can be applied in the web application for the blog sentiment analysis, client feedback analysis, user review analysis in the e-commerce websites, specific issue analysis or trendy topics analysis in day-to-day world life via social media. Also using the WSD heuristic method, computation time of word concepts decreases by means of direct relatedness word if the gloss of target word contains the "feeling" or "emotion". The most of the methods usually produce discrete sentiment labels such as joy, happy, etc., without their intensity scores. Here, the discrete sentiment label with their evaluated intensity scores of different sentiments produced. In emoji sentiment analysis, extraction of sentiment words using graph navigation in lexicon graph avoids the ambiguity in relatedness category.

## 2. RELATED WORKS

**Mireille Fares et al. [5]** proposed an unsupervised word-level knowledge using graph-based approaches and LSA techniques. It provides the unsupervised learning method without manual effort and the large corpus. The lexicon-based approach in LSA used to match target word in a lexical KB with seed words in a sentiment lexicon. Unlike the most of system using supervised method, this paper solves the semantic connectivity problem. It has two main modules; they are affect navigation and affect propagation, affect lookup. The affect navigation proceeds in the lexical affective graph created using lexical KB and affect KB such as WordNet and WordNet Affect Hierarchy. It uses Dijkstra algorithm for finding the shortest path for maximum weight between the nodes. As the result of affect navigation, the sentiment score in the form of sentiment vector is obtained in polynomial time. The, the WNAH and back propagation determine the affect score of all affect nodes in WNAH. The WNAH_propagation computes the sentiment scores of every affect node w.r.t every other affect node in WNAH. So, each affect category becomes fully representative of all others. Then, the sentiment lexicon generated with pre-computed values of WNAH_Propagation, which is used to efficiently search any word concept affect score through affect lookup component in logarithmic time.

**Milagros Fernandez-Gavilaneset et al. [6]** discussed about the blogs, social networking sites generate enormous amount of unstructured data of great interest to extract the sentiments of individual and organization. The sentiment expressed by the emojis and usage of symbols received the little attention. The twitter has about twenty billion emojis and new ones keep appearing in each new Unicode version, making them increasingly relevant to sentiment analysis tasks. This method proposed a novel approach beyond the NLP, to evaluate these sentiments polarity metrics. It predicts the sentiments expressed by emojis in online texts such as tweets. And, it does not require manual effort of human to annotate data and saves valuable time for other tasks. So, here automatic sentiment lexicon for emoji were constructed. It is based on unsupervised sentiment analysis, which based on the definitions given by emoji creators in Emojipedia. Additionally, the automatic creation of lexicon

variants by considering the sentiment distribution of the informal texts accompanying emoji's. These lexica are evaluated shows promising result. But, users use emoji without knowing the emoji meaning leads to less accuracy in the emoji sentiment prediction.

**Shan Huang et al. [7]** proposed the novel based graphical expression are derived from emoticons, emojis have a wide application in social networks. The emojis can assist people in expressing stronger sentiment or show subtler sentiment indirectly which are helpful to sentiment analysis. In this paper, the proposed method called emojis-based recurrent neural network for sentiment analysis in Chinese microblogs. The differentiation of ambiguous emojis and explicit emojis by using pre-trained word embedding and a new sentiment lexicon. Then, it can verify that users information can eliminate the ambiguity of ambiguous emojis to some extent and confirm the sentiment polarity of ambiguous emojis. On the basis, the obtained emoji representations by utilizing the position vector, semantic vector and sentiment vector of emojis, then put the emoji representations into Bi-directional gated recurrent unit(BiGRU) neural network model to conduct sentiment analysis. The experimental results on a Chinese microblog dataset demonstrate that compared with other baselines, the proposed model can improve the accuracy significantly in sentiment analysis.

**Darsha Chauhan et al. [1]** proposed various methods to determine sentiment score of a statement with semiotics. The purpose of these social media websites or microblogging sites available like Twitter, Tumbler, and Facebook is that its user can express their feelings without being pressurized by anyone. They can also give their point of view regarding the recent events in their surroundings as well as give suggestions to improve surroundings in text-based format. User can convey their emotions which they are not able to easily verbalize using emoticons and emojis. Each emoticon and emoji has a particular emotion / sentiment attached to it. For better understanding of people's opinion, it is important to analyze semiotics as well as sentence. Here, the importance of semiotics in sentiment analysis was discussed.

**Khalifa Chekima et al. [2]** proposed the framework to tackle some of the most common challenges posed by Malay social media text (informal text). The researchers have shown a tremendous interest in building automated Sentiment analysis applications for English language and non-English languages such as Arabic Language, French language, Deutsch language, Chinese language, Italian language, etc. Yet, very limited researches have been attributed to Malay opinionated social media text despite the big number of Malay native speakers which recorded to be approximately 215 million native speaker worldwide. in this paper, the features discussed were the handling of Bahasa Rojak also known as Mix language i.e., Malay-English language, Bahasa SMS, Emoticon and Valance shifter. As a result, RojakLex lexicon was constructed consists of 4 different lexicons combined together, namely MySentiDic: a Malay lexicon, English Lexicon: Translated version of MySentiDic, Emoticon lexicon: a combination of 9 different well known lists of commonly used online emoticons, Neologism lexicon: consists of common neologism words used in Malay social media text. The proposed system shows better improvement in the accuracy by recording 79.28% compared to baseline which recorded 51.38% only.

**Petra Kralj Novak et al. [9]** proposed the first emoji sentiment lexicon, called the Emoji Sentiment Ranking, and draw a sentiment map of the 751 most frequently used emojis The new generation of emoticons was called emojis. These emojis were used in mobile communications and social media increasingly. Emojis are Unicode graphic symbols, used as shorthand to express concepts and ideas. The sentiment of the emojis is omputed from the sentiment of the tweets in which they occur. The 83 human annotators were engaged to label over 1.6 million tweets in 13 European languages by the sentiment polarity such as negative, neutral, or positive and where about 4% of the annotated tweets contain emojis. The sentiment analysis of the emojis leads to draw several interesting conclusion such as most of the emojis are positive and well especially the most popular ones. The inter-annotator agreement on that the end of the tweets with emojis is higher and their sentiment polarity increases with the distance. It was observed that no significant differences in the emoji rankings between the 13 languages and the Emoji Sentiment Ranking. This method was proposed as Emoji Sentiment Ranking for European language independent resource for automated sentiment analysis. Finally, this paper provides a formalization of sentiment and a novel visualization in the form of a sentiment bar.

**Meng Joo Er et al. [4]** proposed the objective of extracting useful information from the opinion-rich data on Twitter, both supervised learning-based and unsupervised lexicon-based methods for sentiment analysis on Twitter corpus have been studied in recent years. the lack of labels and frequent usage of emoticons were the unique characteristics of tweets which poses challenges to most of the existing learning-based and lexicon-based

methods. In twitter, it mainly focus on domain specific tweets were about personal feelings and comments on daily life events with the maximum amount of tweets. The hybrid approach of augmented lexicon-based and learning-based method were designed to handle the distinctive characteristics of tweets and perform sentiment analysis on a user level. It also provided information of specific Twitter users typing habits and their online sentiment fluctuations. This model is capable of achieving an overall accuracy of 81.9 %, largely outperforming current baseline models on tweet sentiment analysis.

**Ben Eisner et al.[3]** proposed the release of emoji2vec, pre-trained embeddings for all Unicode emojis which are learned from their description in the Unicode emoji standard. The most of natural language processing applications for social media used the representation learning and pre-trained word embeddings. The several publicly-available or the pre-trained sets of word embeddings contain few or no emoji representations even as emoji usage in social media has increased. The emoji embeddings can be readily used in downstream social natural language processing applications alongside word2vec. It also demonstrates for the downstream task of sentiment analysis that emoji embeddings learned from short descriptions outperforms a skip-gram model trained on a large collection of tweets. And it avoid, the need for contexts in which emojis need to appear frequently in order to estimate a representation.

**Mayu Kimura et al.[8]** proposed automatically constructing an emoji sentiment lexicon with arbitrary sentiment categories. The emojis used frequently to express users sentiments, emotions, and feelings in text-based communication. By using manually labeled tweets, an emoji sentiment lexicon constructed to facilitate sentiment analysis of user posts with category such as positive, neutral, and negative. In those emoji sentiment lexicons, the number of emojis listed in is smaller than the currently existing emojis. To reconstruct the labeled dataset and expansion of the lexicon manually requires time and effort. The proposed method extracts sentiment words from WordNet-Affect and calculates the co-occurrence frequency between the sentiment words and each emojis. Each emoji was assigned to a multidimensional vector whose elements indicate the strength of the corresponding sentiment which was based on the ratio of the number of occurrences of each emoji among the sentiment categories,. In experimental results shows a high correlation between the conventional lexicon and our lexicon for three sentiment categories. Also, the results for a new lexicon constructed with additional sentiment categories.

## 3. PROPOSED SYSTEM

### 3.1 Lexicon Graph Creation

It is created by interconnecting the lexical KB and relatedness KB, where the WordNet is used as lexical KB and WordNet relatedness hierarchy (WNRH) is used as relatedness KB. The lexicon graph creation is only possible when atleast one of their lexical concepts matched with any one of the relatedness categories in WNRH.

#### 3.1.1 WordNet

The WordNet is a machine-readable lexical KB. It is a graphical representation of semantic relationship between the set of concepts with set of labeled links. The concepts are the nodes, which expresses the word senses called Synsets and the links are the edges, which expresses the relationship between nodes. Also, the WordNet has more than 18 semantic relationships such as meronymy, pertainymy, etc. The WordNet uses unique ID's to identify the polysemous of concepts .i.e., a word or phrase has several meanings. The WordNet concepts are categorized into four groups and they are Noun concept, Verb concept, Adjective concept, Adverb concept. The Noun concept has hypernymy (HasA) and hyponymy(IsA). E.g., the color is the hypernym and its subdivision such as purple, red, blue, etc are the hyponym.

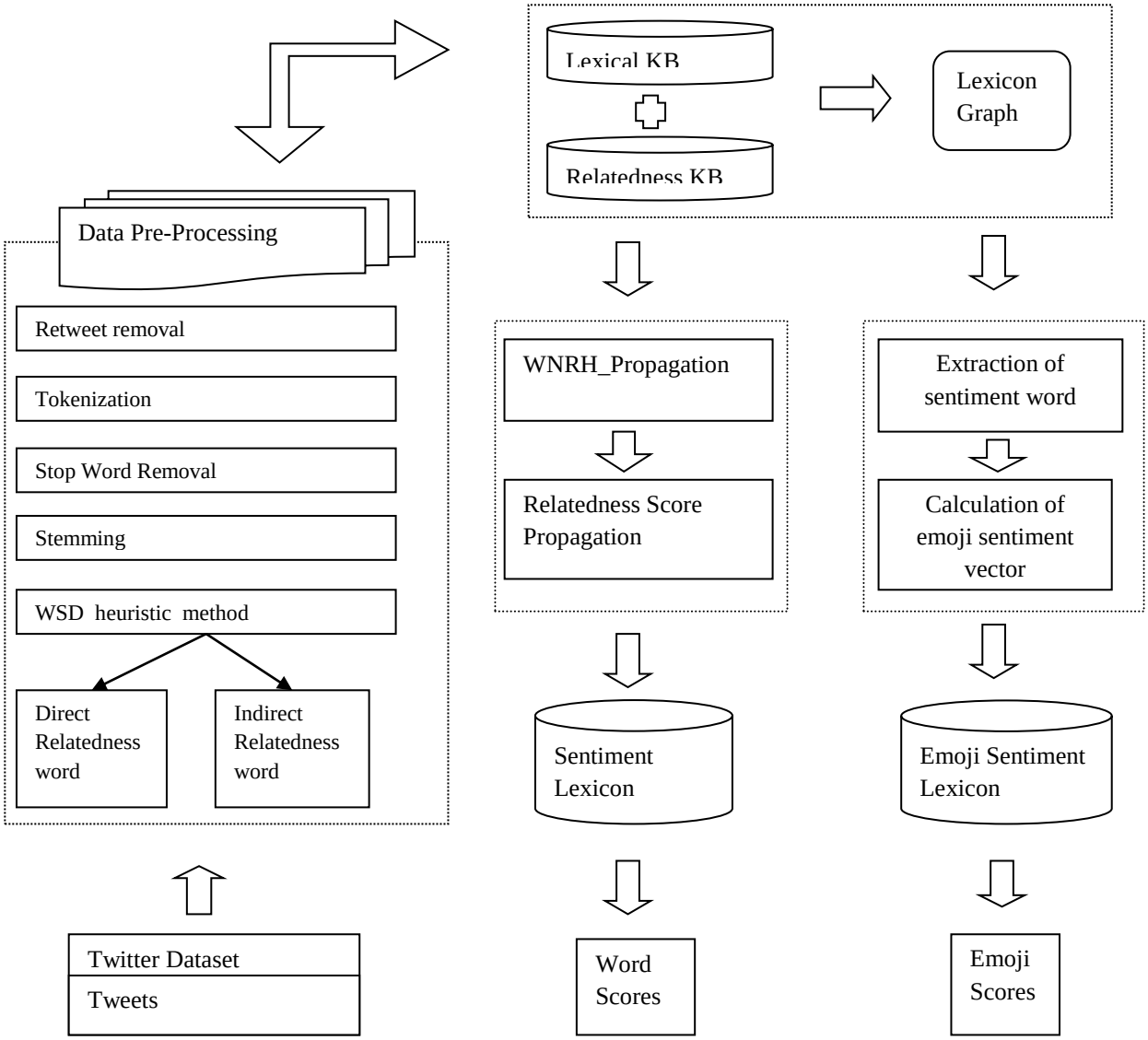



**Fig-1:** System Architecture

By means of linguistics, a hyponym is a word or phrase whose semantic field is included within of another word and a hyponym is a type of relationship with its hypernym. Another example is for the IsA is that violet is a hyponym of purple and the purple is a hyponym of color, therefore violet is a hyponym of color. The Verb concept has hypernymy, entailment, troponymy and verb group. The entailment is representation of relationship between the two statement i.e., when one statement is true then another statement must be true. The troponymy is the presence of a manner relationship between two lexemes. The Adjective concept is classified using the following relationship between nodes such as attribute, similar to, related to. The Adverb concept such as pertainymy which is a lexical relationship and that allows deriving the adjective of noun.

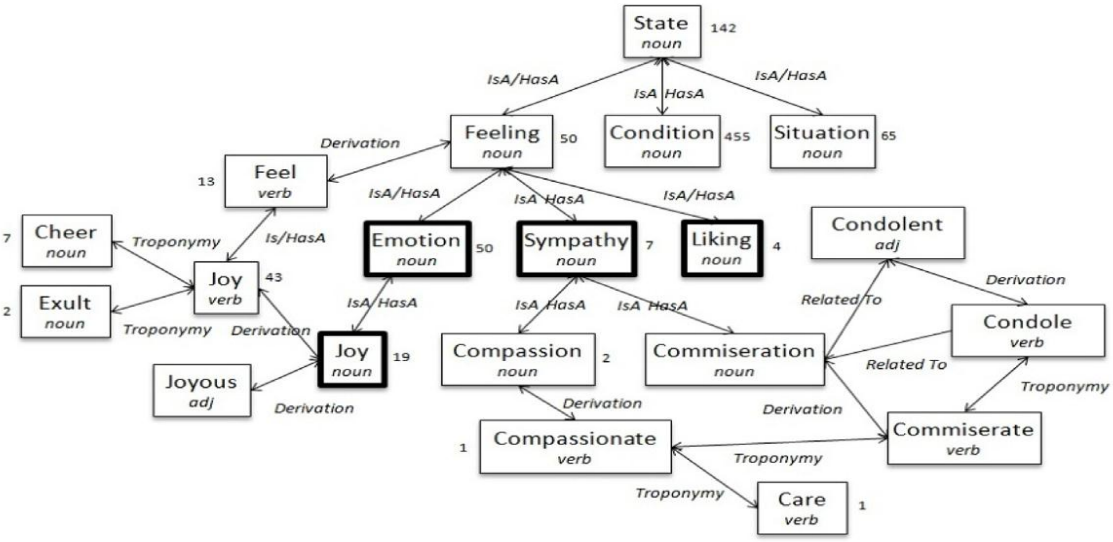



**Fig-2:** WordNet –Lexical KB

### 3.1.2 Wordnet Relatedness Hierarchy

The WNRH have subset of Synsets and which is represent the relatedness concepts correlated with relatedness words. It is Knowledge Base which consists of 294 different categories such as positive emotion, joy, love, apathy, euphoria, etc. The WNRH is organized in hierarchical structure of hypernymy/hyponymy (IsA/HasA). The graph creation using the WNRH with WordNet highly improves the efficiency of sentiment classification, by detecting the emotion with relatedness categories such as happy, sad, anger. The categories in WNRH are mapped to corresponding word concepts in WordNet. The certain composite relatedness category like self-pride, general-dislike and neutral-languor are exceptional.

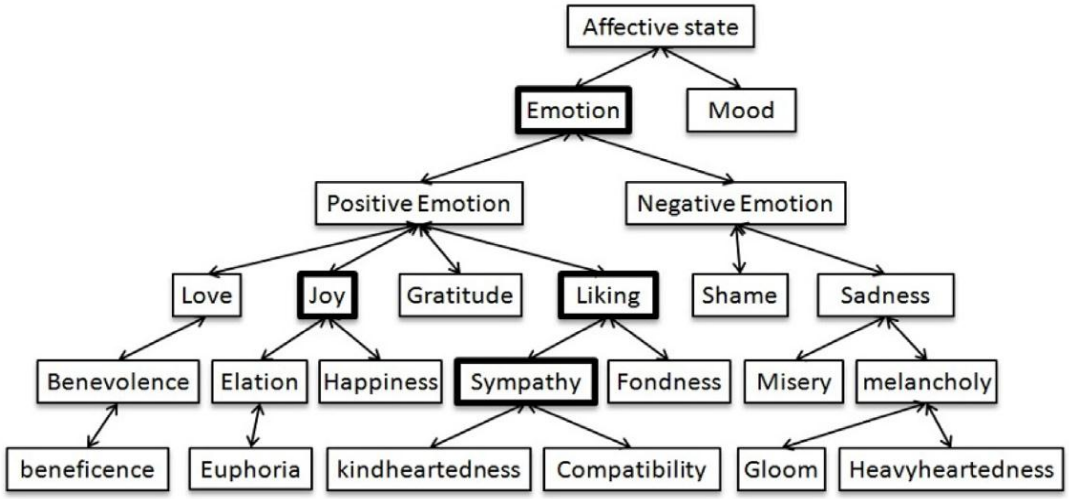



**Fig- 3:** WordNet Relatedness Hierarchy-Relatedness KB

### 3.2 Data Pre-Processing

The data pre-processing step include the Retweet Removal, Tokenization, Stop Word Removal, Stemming, Word Sense Disambiguation (WSD) for the user given or selected input. Here, the twitter dataset is used. The retweet removal will remove retweets, #, UserID's, URL, unnecessary symbols such as @,",!. The tokenization process involves the identification of separate tokens for user selected or given input. The Stop Word Removal is used for removing the prepositions, semantically meaningless words such as "the, a, of, to". The next processes are the Stemming which is used to convert the derived words to their stem. E.g., enjoyed – enjoy. The WSD is used to identify and select the proper word concept in the WordNet.

The WSD processes the sentences as sequences of stemmed words. Mostly, in WSD processing the LESK algorithm is used. Here, the WSD heuristic method is used based on two observations for efficient running time. The first observation is that the LESK algorithm uses the linear time with respect to number of possible meanings for a given word with their size. And the second observation is that the WordNet contains the words "feeling" or "emotion" as the most sentiment carrying concepts in their gloss definitions. E.g; Elation- "A feeling of joy and pride" and Apathy- "An absence of emotion or enthusiasm".

The proposed WSD heuristic method will search for proper concept in WordNet by means of either direct relatedness word method or indirect relatedness word method. The direct relatedness word method is used to search for the gloss definition of word concept with terms containing either "feeling" or "emotion" and also to identify the polysemous concepts of word concept. The indirect relatedness word method is used to find proper word concept either using the LESK algorithm or first most common concept among multiple concepts identified according to user preference. The usage of LESK algorithm is reliable, which compares the target word's context with the contexts of its different possible concepts in the lexical kb. It chooses the concept whose context is most similar to the target word as its proper disambiguated meaning. And, it requires linear time with respect to number. of. possible meanings for a given word and their context sizes. And the selected word concept in the WordNet is given as input to the extraction of sentiment words.

### 3.3 Extraction of Sentiment Word

For the extraction of sentiment word, the selected word concept in the WordNet should be direct relatedness word or indirect relatedness word. While the input is given as user disambiguated word concepts through WSD leads to the list of sentiment words with their category extracted using lexicon graph by an below function:

- The source concept node, target relatedness node, new incoming node are represented by the mathematical expressions such as $c_i$, $c_j$, $c_r$ respectively. Before starting the graph navigation in lexicon graph, initialize the weight of$c_i$= 0.The weight of a source concept node $c_i$ w.r.t a target relatedness node $a_j$, where Relatedness category $a_j$ is not expressed in $c_i$ represented as 0 and 1 as if $a_j$ is totally expressed in $c_i$. it is mathematically represented as $\mathbf{w(c_i,a_j) \in (0,1)}$
- The weight of source concept node $c_i$ w.r.t a set of target relatedness categories $A=\{a_1,a_2,\ldots\ldots a_J\}$,which consists of a vector of relatedness weights $V_i=<w(c_i,a_1),\ldots,w(c_i,a_J)>$.The J is the dimension, where dimension j corresponds to a target relatedness category $a_j$ belongs to A and its vector coordinate $W(c_i,a_j)$ represents the relatedness weight of $a_j$ w.r.t $c_i$. The weight of an edge outgoing from node $c_i$ and incoming into node $c_r$. it is mathematically expressed as $\mathbf{w(c_i,c_r) \in (0,1)}$, where the edge does not carry any sentiment expressiveness from $c_i$ to $c_j$ represented as 0. Otherwise, the edge carries all the sentiment expressiveness from $c_i$ to $c_j$ as 1.
- The computation of edge weight is based on either edge label or out-degree of $c_i$. The edge label method is used, when the semantic relationship between two nodes are such as hypernymy, related to, etc. The out-degree of $c_i$ method is used, when the computation of semantic relationship being processed.

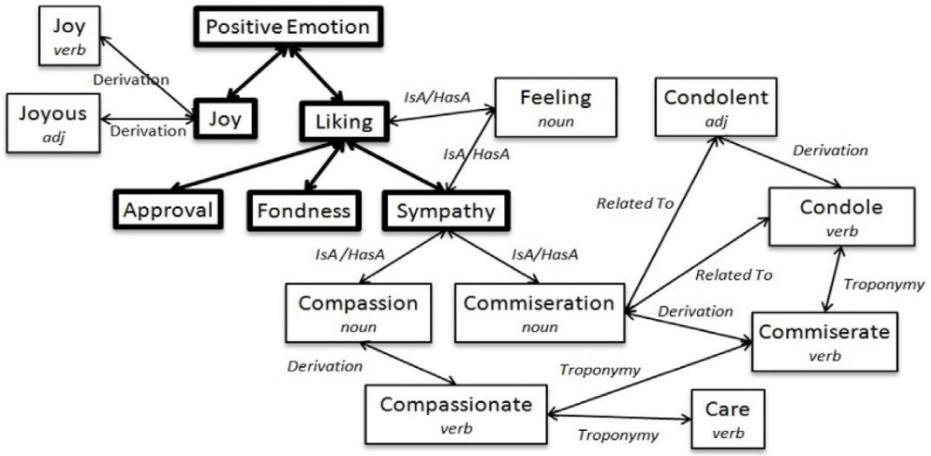



**Fig-4:** Lexicon Graph

- **Reliable and Partially Reliable:**

  The semantic relationships between the nodes are considered in two ways such as Reliable and Partially Reliable. The emotion carries between the nodes such as Hyponymy, similar to, related to, attribute, usage, derivation and pertainymy are considered as Reliable and the emotion carries between the nodes such as hypernymy, entailment, verb group, troponymy are considered as partially reliable. The $c_i$ is the weight of edge outgoing from $c_i$ , $c_r$ as incoming into node, rel as edge's label and the set of sentiment reliable relationships($R_{reliable}$). It is mathematical function is

$$W(c_i, c_r) = \begin{cases} \dfrac{1}{out - degree_{rel}(c_i)} & if\ rel \notin R_{reliable} \\ 1 & otherwise \end{cases}$$

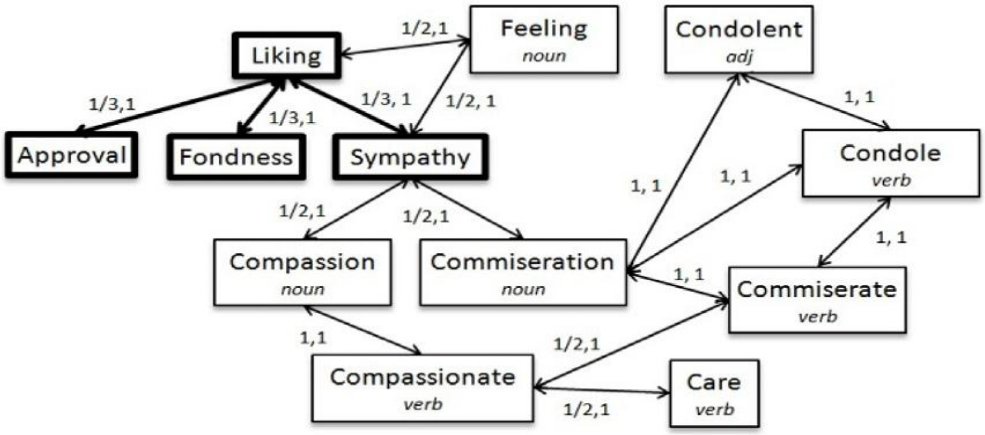



**Fig-5:** Computed Scores of Lexicon Graph

Its maximum score will be 1 (if its edge label corresponds to a sentiment reliable relationship).Otherwise, its determined by out-degree of incoming node $c_i$. For example, a tweet before data pre-processing looks like : ": 7.98577E+17 Broke and sad \U0001f643 https://t.co/vivqvFniQe" . Here, the unicode\U0001F643 is the upside-down face. While in pre-processing of WSD Heuristic method, break & sad are the indirect relatedness word and direct relatedness word respectively. It leads to the extraction of sentiment word by the lexicon graph navigation using synsets ID's.

### 3.4 Calculation Of Emoji Sentiment Vector

In the twitter dataset, the Unicode conversion is made in it. Then, the each emoji are collected separately called as target emoji. The sentiment scores of a target emoji calculated using the ratio of appearance of emojis for each sentiment. Then , calculate the co-occurrence frequency between target emoji and sentiment word. The sentiment score ES($e_i$,s) of emoji $e_i$ for sentiment using the following equation. It also avoids the ambiguity problem between two relatedness category by using this graph navigation technique. A larger value of ES($e_i$,s) indicates a close relationship between the emoji $e_i$ and the sentiment. Each emoji is reprsented by a |S|-dimensional vector whose elements are{ES($e_i$,s)}s €S. the mathematical representation is $\mathbf{ES(e_i,s)=\sum_{j=0}^{W} a(w_j,s)n_{ij}/\sum^{W}_{j=0}n_{ij}}$ .

The representation of the number of their co-occurrence frequency in the dataset ($n_{ij}$), sentiment(s), sentiment word($W_j$), each emoji($e_i$) where i=1,2,....E i.e., E is the total no of emojis where j=1,2,....Wand W is total no of sentiment words.

### 3.5 WNRH_Propagation

The relatedness score computation happens between relatedness nodes considering only their relatedness connections, instead of lexical or semantic connections. It helps to solve the semantic connectivity or inconsistency problem. The sentiment score is computed by the every relatedness node $a_j$ w.r.t every other relatedness node in WNRH. So, that each relatedness node becomes fully representative of all others. Every relatedness node $a_j$ in WNAH will be associated with a sentiment vector $V_j$ consisting of 294 dimensions, where every dimension represents every other relatedness node in WNAH with its corresponding relatedness score w.r.t to $a_j$.It computes the path from $c_i$ to the closest relatedness node $a_j$.That relatedness node $a_j$ would provide sentiment scores for all other WNAH relatedness category through sentiment vector $V_j$. it is represented as $\mathbf{V_j=W(c_i,a_j)*V_i}$ .

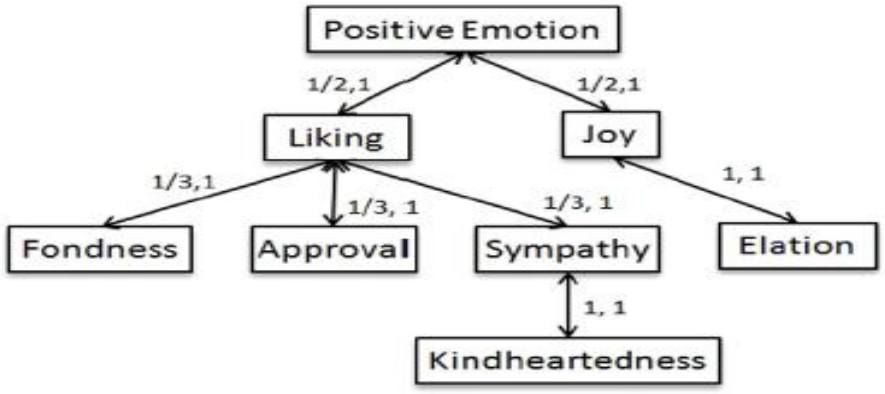



**Fig-6:** Extract of the WNAH hierarchy from

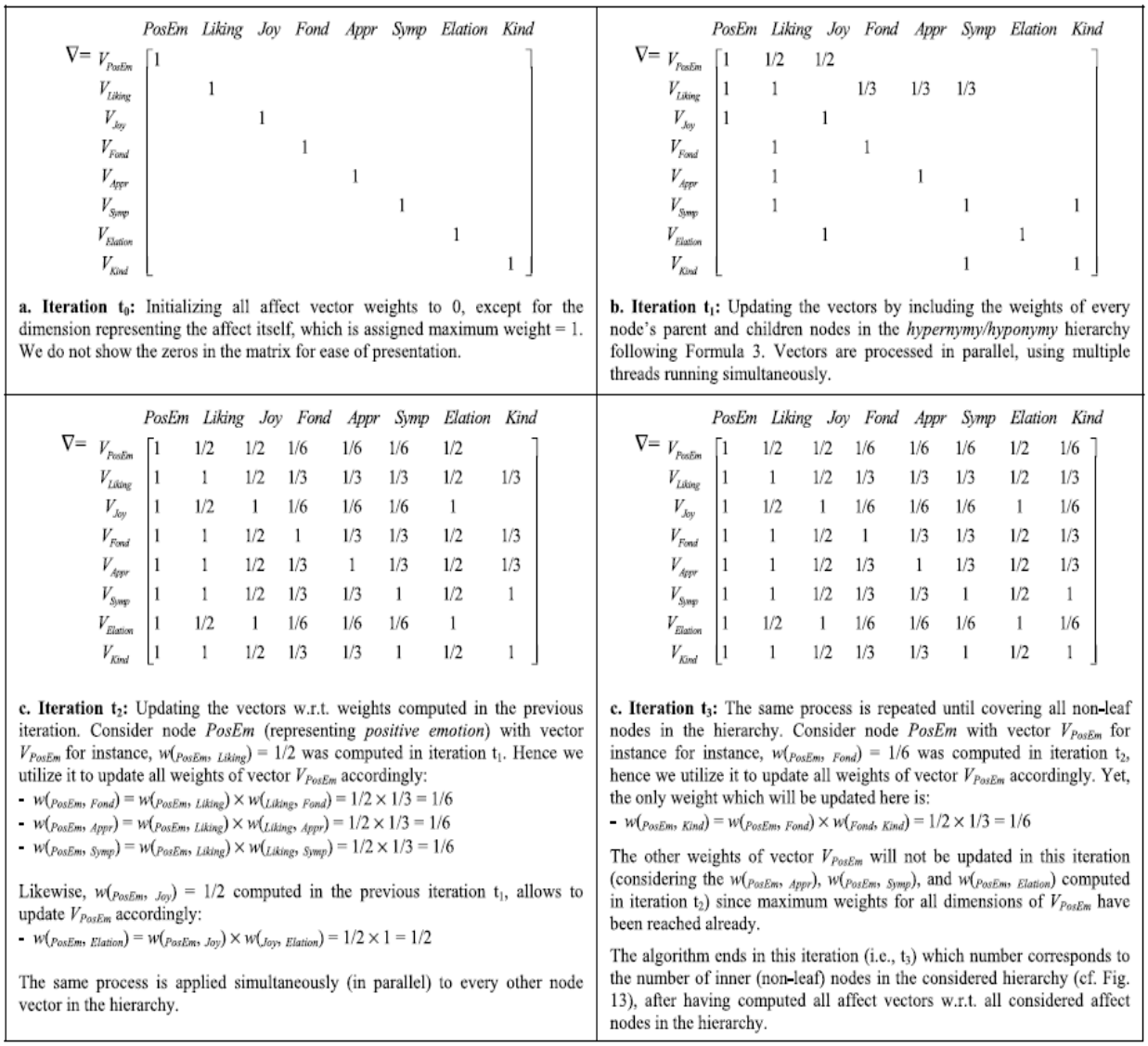
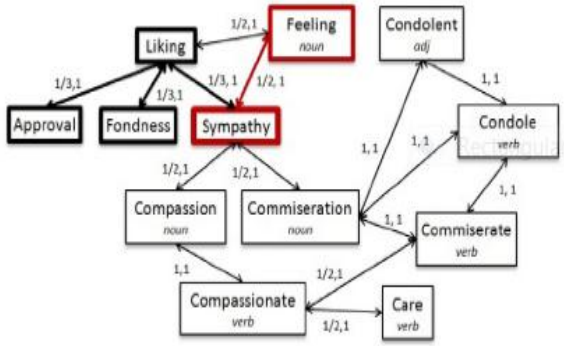



**Fig-7 : WNRH_Propagation**

By reaching any one relatedness node $a_j$ by navigating the lexicon graph using source word concept $c_i$ would povide $c_i$ with an affective vector $V_i$ containing relatedness scores of all 294 WNAH relatedness category.

### 3.6 RELATEDNESS SCORE PROPAGATION

It propagates the lexicon graph, using the pre-computed WNRH _propagation scores from user chosen relatedness nodes or category to all connected concept nodes. So, that all lexical concepts connected with any relatedness node directly or indirectly through an edge or path respectively provide the relatedness score, which leads to sentiment lexicon formation. The resulting set of relatedness scored concepts forms a sentiment lexicon. It can be efficiently used to search or lookup for any word concept relatedness score. It is done by make use of legacy indexing technique through B+ tree. It helps efficiently to access any word concept relatedness score. The several thread execution is done for parallel processing.

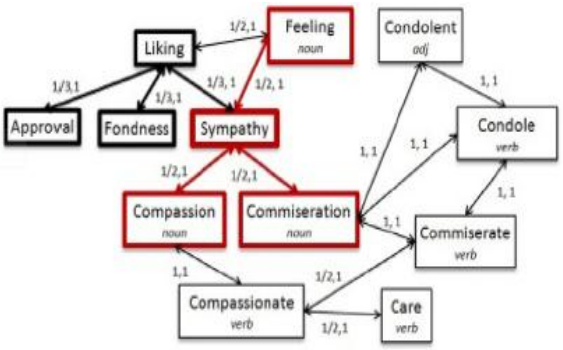



i. **Thread #1**: Considering **source** node ***Sympathy***:

- Initialize weight of source node: $w(Sympathy) = 1$
- Affective vector of sympathy w.r.t. all considered affect categories has been (pre-computed) provided: $V_{Sympathy} = <1 \quad 1/3 \quad 1 \quad 1/3>$ following dimensions *Liking*, *Fondness*, *Sympathy*, and *Approval* respectively
- Iteration #1: Fill neighbors in *Frontier* and identify maximum weight: $w_{Symp}(Feeling) = 1$
  - Compute affective vectors:

$$\nabla_{Symp} = \begin{array}{c} \\ V_{Feeling} \\ V_{Compassion} \\ V_{Commiseration} \\ V_{Compassionate} \\ V_{Condolent} \\ V_{Condole} \\ V_{Commiserate} \\ V_{Care} \end{array} \begin{array}{c} \begin{array}{cccc} Liking & Fond & Symp & Appr \end{array} \\ \begin{bmatrix} 1 & 1/3 & 1 & 1/3 \end{bmatrix} \end{array}$$

⇨ Remove node *Feeling* from *Frontier* and include in *Explored* set




b. Iteration #2: Fill neighbors in *Frontier* and identify maximum weight: $w_{Symp}(Compassion) = w_{Symp}(Commiseration) = 1/2$

- Compute affective vectors:

$$\nabla_{Symp} = \begin{array}{c} \\ V_{Feeling} \\ V_{Compassion} \\ V_{Commiseration} \\ V_{Compassionate} \\ V_{Condolent} \\ V_{Condole} \\ V_{Commiserate} \\ V_{Care} \end{array} \begin{array}{c} \begin{array}{cccc} Liking & Fond & Symp & Appr \end{array} \\ \begin{bmatrix} 1 & 1/3 & 1 & 1/3 \\ 1/2 & 1/6 & 1/2 & 1/6 \\ 1/2 & 1/6 & 1/2 & 1/6 \end{bmatrix} \end{array}$$

⇨ Remove both nodes from *Frontier* and include in *Explored* set

**Fig-8:** Thread execution in relatedness score propagation

$$\nabla_A = \begin{array}{c} \\ V_{Liking} \\ V_{Fond} \\ V_{Symp} \\ V_{Appr} \end{array} \begin{array}{c} \begin{array}{cccc} Liking & Fond & Symp & Appr \end{array} \\ \begin{bmatrix} 1 & 1/3 & 1/3 & 1/3 \\ 1 & 1 & 1/3 & 1/3 \\ 1 & 1/3 & 1 & 1/3 \\ 1 & 1/3 & 1/3 & 1 \end{bmatrix} \end{array}$$

**Fig-9:** Input Relateness Vectors

## 4. EXPERIMENT REQUIREMENT

It can be implemented in java with the help of MySQL database, Apache Tomcat Server and eclipse platform. The twitter dataset should be given as text file. The WordNet should be imported in eclipse with necessary .jar files by using WordNetSQLBuilder . Here, five-dimensional sentiment score vector calculated for each of the emojis. The principal component analysis is applied to the vectors to show the distribution of resulting sentiment vector. And its observed that the emojis have similar sentiment mapped close to each other.

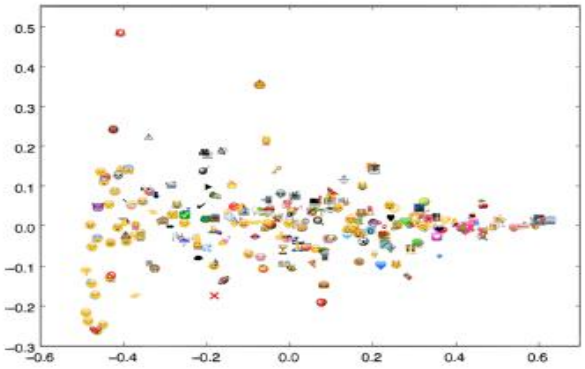



**Fig-10 : Sample visualization of five-dimensional emoji sentiment vector in a two- dimensional space**

## 5. CONCLUSION

The proposed method can be utilized for calculation of relatedness score of both text and emoji using the word-knowledge in the graph based approach of sentiment analysis. The proposed unsupervised method saves large time by automatic construction of sentiment lexicon and emoji lexicon. Also, the sentiment lexicon and emoji lexicon can be updated efficiently with new emoji or new sentiment category. This word-level LSA functionality can be updated efficiently by adding valence shifters to it.

## REFERENCES

[1] Chauhan, D., Sutaria, K., & Doshi, R. (2018, February). Impact of semiotics on multidimensional sentiment analysis on twitter: A survey. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC) (pp. 671-674). IEEE.

[2] Chekima, K., & Alfred, R. (2017, November). Sentiment Analysis of Malay Social Media Text. In International Conference on Computational Science and Technology (pp. 205-219). Springer, Singapore.

[3] Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. arXiv preprint arXiv:1609.08359.

[4] Er, M. J., Liu, F., Wang, N., Zhang, Y., & Pratama, M. (2016, July). User-level twitter sentiment analysis with a hybrid approach. In International symposium on neural networks (pp. 426-433). Springer, Cham.

[5] Fares, M., Moufarrej, A., Jreij, E., Tekli, J., & Grosky, W. (2019). Unsupervised word-level affect analysis and propagation in a lexical knowledge graph. Knowledge-Based Systems, 165, 432-459.

[6] Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castaño, F. J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Systems with Applications, 103, 74-91.

[7] Huang, S., Zhao, Q., Xu, X. Z., Zhang, B., & Wang, D. (2019, November). Emojis-Based Recurrent Neural Network For Chinese Microblogs Sentiment Analysis. In 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) (pp. 59-64). IEEE.

[8] Kimura, M., & Katsurai, M. (2017, July). Automatic construction of an emoji sentiment lexicon. In Proceedings of the 2017 ieee/acm international conference on advances in social networks analysis and mining 2017 (pp. 1033-1036).

[9] Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. PloS one, 10(12), e0144296.