

# A SURVEY ON GRAPH BASED APPROACHES IN SENTIMENT ANALYSIS

Mrs. K. Krishnakumari<sup>1</sup>, P. Akshaya<sup>2</sup>

<sup>1</sup>Associate professor, Department of CSE, A.V.C. College of Engineering, Tamil Nadu, India

<sup>2</sup>M.E student, Department of CSE, A.V.C College of Engineering, Tamil Nadu, India

\*\*\*

**ABSTRACT:** Sentiment Analysis is a computation of opinions or sentiments and subjectivity of text or from the generic and specialized social media like Facebook, Twitter and LinkedIn respectively. Sentiment analysis helps in achieving various goals like observing public mood regarding political effort, market intelligence, the measurement of customer satisfaction, movie sales prediction and many more. Massive amount of information related to customer opinions or reviews is quite burdensome to be evaluated. Sentiment analysis systems allows companies to make sense of this sea of unstructured text by automating business processes, getting actionable insights, and to save hours of manual data processing, in other words, by making teams more efficient, scalable, real time analysis and consistent criteria. The sentimental analysis is done by using various algorithms like rule based, automatic systems and hybrid systems. The different approaches of graph based sentiment analysis are discussed for the improved performance, computation, and storage and accuracy prediction in the sentiment classification.

**Key Words:** Sentiment Analysis, Graph Based Approach, Sentiment Classification, keyword Extraction

## 1. INTRODUCTION

The usage of online events like chatting, conferencing, blogging, ticket booking, online transactions, e-commerce, social media communications, observations, micro-blogging, clicks streams, etc. leads us to extract, transform, load, and analyze very massive amount of structured and unstructured data, at a fast step, stated to as Big Data. The two tasks of sentimental analysis are data preprocessing and data acquisition. After, the opinions or reviews collected as document about any product, service, or organization leads to preprocessing step. The data are pre-processed using tokenization, pos tagging, feature extraction and representation, etc and then extracted, and implemented in the systems using techniques like machine learning. For example opinions given on e-commerce sites like Amazon, IMDb, etc can influence the customer's decision in buying products and subscribing services.

### 1.1 Sentimental Analysis

Sentiment analysis is the automated process of understanding an opinion about a given subject from a written language. The graph based approaches in sentimental analysis is having many techniques and approaches are there. The various types of graph based approaches are using machine learning techniques, statistical model, etc. The sentimental analysis is having the sub-tasks like data acquisition and data pre-processing. In the graph- based approaches, the nodes and vertices are selected according to the features present in the input source materials like reviews and stock prices data.

### 1.2 Need of Sentimental Analysis

To automatically mine the positive, negative and neutral classification in the source material or from user reviews. To know about the real feelings of user about the products or services. The graph based approaches in sentimental analysis can provide the higher performance when compared with lexicon-based methods and Machine learning based methods like Naive Bayes (NB), Maximum Entropy (ME) and support vector machine (SVM).

### 1.3 Types of Sentimental Analysis

There are many types of sentiment analysis. The tools of sentiment analysis that focus on polarity like positive, negative, neutral and to detect the feelings and emotions such as angry, surprise, happy, sad, etc. Also to identify the intentions of user, whether the user is interested or not towards a particular topic or online products. The most important types are listed below.

#### 1.3.1 Fine-grained Sentiment Analysis

It is used to deals with the precise level of polarity for a given opinion. Its polarity levels are of very positive, positive, neutral, negative and very negative. For example, mapped onto a 5-star rating

in a review, e.g.: Very Positive = 5 stars and Very Negative = 1 star for any online services. Some systems such as e-commerce websites like Amazon provide different flavors of polarity, by identifying the positive or negative sentiment is associated with a particular feeling. The anger, sadness, worries are considered as negative feelings and the happiness, love, or enthusiasm are considered as positive feelings.

### 1.3.2 Emotion detection

It aims at detection of emotions like, happiness, frustration, anger, sadness, and the like. The lexicons i.e., list of words and the emotions they convey or complex machine learning algorithms are widely used for emotion detection system. The disadvantage of resorting the lexicon is that customer expressing their feeling about a product reviews in the different ways. Their emotions vary a lot for each customer and so do the lexical items they use. For example, some of the words that would express anger and also happiness like shit or kill.

### 1.3.3 Aspect-based Sentiment Analysis

This is based on the analyzing the sentiments towards a subject, whether the customer is interested in this subject or not. If the user is interested, then the polarity of the product but also particular aspects or features of the product in which the most of the people are interested should be identified. So, its results are more detailed and accurate and it is not only based on overall sentiment classification.

### 1.3.4 Intent analysis

Intent analysis basically detects the underlying intention behind a sentence of opinion. A human being can manually predict the intended action of a customer from reviews by their emotions. When it comes to intent analysis, the intended action can be inferred from the text, but sometimes, inferring it requires some contextual knowledge. It is a level deeper than sentiment analysis and the usage of trained datasets of social media gives the idea of whether a string is a complaint, a suggestion or a query.

### 1.3.5 Multilingual sentiment analysis

Multilingual sentiment analysis is a difficult task. It leads to a lot of pre-processing and for that pre-processing many number of resources can be utilized. The resources such as sentiment lexicons are mostly available in online and some may have to be created like translated corpora, noise detection algorithm. Also the usage of the resources available or to create new requires a lot of coding experience and can take long to implement.

**Table 1: Different Approach on Sentimental Analysis**

Approach	Sentiment extraction
Document level	To obtain the sentiment of a complete document.
Sentence level	To obtain the sentiment of a single sentence.
Sub-sentence level	To obtain the sentiment of sub-expressions within sentence.
Corpus-based	To obtain the sentiment from the dictionaries related to a specific domain which is generated from a set of seed opinion through the search of related words.
Dictionary-based	To obtain the sentiment from the dictionary by usage of seed terms and this is collected and annotated manually by searching the synonym and antonym of the dictionary.
Graph-based	To obtain the sentiment from the graph structured framework of opinions with the extracted features.
Manual	Not feasible, time consuming

## 2. RELATED WORKS

**Andrew B. Goldberg, Xiaojin Zhu [1]** presented a graph-based algorithm to address the sentiment analysis task of rating inference. It demonstrates that considering unlabeled reviews in the learning process can improve rating inference performance. The graph is created on both labeled and unlabeled data to encode certain assumptions for this task and it also solves an optimization problem to obtain a smooth rating function over the whole graph. **Kavita Ganesan, Cheng Xiang Zhai, Jiawei Han [2]** presented a novel graph-based summarization framework of using opinions. This framework generates concise abstractive summaries of highly redundant opinions. The generated abstractive summaries of highly redundant opinions are readable, reasonably well-formed and also they are informative enough to convey the major opinions. **Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang [3]** proposed hashtag-level sentiment classification instead of presenting the sentiment polarity of each tweet relevant to the topic. It will automatically generate the overall sentiment polarity for a given hashtag in a certain time period, which markedly differs from the conventional sentence-level and document-level sentiment analysis. The task done by using sentiment polarity of tweets containing the hashtag, co-occurrence relationship and the literal meaning of hashtags. **Fotis Aisopos, George Papadakis, Theodora Varvarigou [4]** suggested that the extraction of information about a particular or various topics using the Social Media which facilitates the extraction of useful conclusions about the average public opinion. The graph which created is with whose node correspond to the character n-grams of a document and its weighted edges denote the average distance between them. **Natalia Ponomareva, Mike Thelwall [5]** presented a graph-based approaches for cross-domain sentiment classification. The two existing methods in this paper analyzed such as an optimization problem and a ranking algorithm. It shows, that the best parameters are analyzed and it reveals that there are no optimal values valid for all domain pairs. **Arturo Montejo - Raez, M. Teresa Martin Valdivia, Eugenio Martinez - Camara, L.AlfonsoUrena-Lopez [6]** presented a novel based approach to Sentiment Polarity Classification using Twitter posts. By using the wordNet graph, vectors of weighted nodes are extracted and those weights are used in SentiWordNet which is used to compute a final estimation of the polarity. **Linhong Zhu, Aram GalstyanJames Cheng, Kristina Lerman [7]** proposed an unsupervised tri-clustering framework, which analyses both user-level and tweet-level sentiments through co-clustering of a tripartite graph. Using an efficient online algorithm, the updating of sequentially arriving new data can be done on the clustering of tweets, users and features. This online framework provides better quality of both dynamic user-level and tweet-level sentiment analysis. And, also improves the computational and storage efficiency. **Esteban Castillo, Ofelia Cervantes, Darnes Vilarino, David Baez, Alfredo Sanchez [8]** presented an co-occurrence graph to find the relationship among the terms in a document. This approach uses a centrality measures to extract the most representative words that express the sentiment. Using a supervised learning algorithm, the extracted words used as feature in graph based representation. **Zhikui Chen, Fei Lul, Xu Yuan, Fangming Zhong [9]** proposed a sentiment classification paper for online service recommendation using the reviews. This approach uses the both single model information and complementary information of online product reviews and leads to satisfied accuracy of sentiment classification. The cross-modal hypergraph coupled with the Latent Dirichlet Allocation (LDA) topic is designed to mitigate the ambiguity of some specific words, which may express opposite polarity in different contexts. **Monali Bordo, Dr. S. K. Biswasloi [10]** suggested an effective sentiment analysis technique for deep understanding of the underlying meaning and sentiment of a given set of text which is highly desirable in many application domains. This model gives the better performance than the three popular machine learning techniques used in this domain namely SVM, NB and MaxEnt for the dataset used. **Xu Yuan, Mingyan Gsun, Zhikui Chen, Jing Gao, Pengli [11]** suggested that a word embedding clustering based deep hypergraph model called ECDHG. The ECDHG introduces external knowledge by employing the pre-training word embeddings to express reviews word based sentiment classification of online reviews. The other existing works ignore the semantic correlation among different reviews and ineffective for sentiment classification. To solve this problem, the ECDHG is proposed for sentiment analysis of online reviews. The improved hierarchical fast clustering algorithm is used to detect the semantic units under the supervision of semantic cliques discovered. **Mario Casillo, Fabio Clarizia, Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale [12]** presented a graphical model called mixed graph of terms. The complex structure mGT is used for classification of a new document. So, it will automatically mine the positive and negative sentiments in the document. It provides better results when compared to the other methods like Naive bayes, SVM and in the literature using standard and real datasets such as Facebook. **Jue Liu, Zhuocheng Lu, Wei Du [13]** suggested to predict the stock price movement using the correlation between the stocks. Because, correlated stocks could cause co-movement. The traditional method of measuring of correlation factor are based only on the similarity between corresponding stock price data, while ignoring the business relationships between companies, such as shareholding, cooperation and supply-customer relationships. This problem is solved by using the enterprise knowledge graph to calculate the correlation. **K. Krishnakumari, E. Sivasankar, Sam Radhakrishnan [14]** proposed a solution to the domain adaptation problem of sentimental classification using neural networks. The convolutional neural networks which are used to learn the knowledge of a particular domain using Doc2Vec feature representation and it derives into a suitable CNN architecture with hyperparameters which favour domain adaptation between the different domains. It shows

that the CNN with Doc2Vec model provides a strong capability of learning large data representation semantically with other state-of-the-art methods for the domain adaptation sentiment classification.

### 3. PAPER CLASSIFICATION AND CATEGORIZATION

Year	Name of Authors	Proposed Method	Approach used	Data Set	Critic
2006	Andrew, Goldberg, Xiaojin Zhu	Graph-based Semi-supervised learning	1.Regression 2.Metric labeling	IMDb Movie review document	When only limited labeled data is available, this method provides better predictive accuracy compared to the other methods which ignore the unlabeled examples during training.
2010	Kavita Ganesan, Cheng Xiang Zhai, Jiawei Han	Opinosis graph	Summarization Algorithm-A2,A3	Reviews of hotels, cars.	It cannot group sentences at a deep semantic level due to graph emphasizes too much on the surface order of words.
2011	Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang	Hashtag Graph Model	1.Approximate Collective Classification Algorithm 2.Enhanced Boosting Classification	Tweets with Hastag	The self-explainable hashtags are label-fixed and not updated for polarity but used for sentiment influence to neighbor hashtags.
2011	Fotis Aisopos, George Papadakis, Theodora Varvarigou	N-gram Graph	N-graph similarities: 1.containment 2.Size 3.value	Tweets	It exhibits high classification accuracy, due to a limited number of features are used which depends on the corresponding number of classes.
2012	Natalia Ponomara, Mike Thelwall	Cross-domain sentiment classification	1.OPTIM algorithm 2.RANK algorithm	Product reviews	The graph domain representations provide a competitive solution to the domain adaptation problem
2014	Arturo Montejoraez, Eugenio Martinez-Camara, M.Teresa Martin-Valdivia, L.Alfonso Urena-Lopez	Sentiment Polarity Classification	1.Random walk algorithm 2.SentiWordNet	Tweets	It proposed a non-supervised solution which is domain-independent.

2014	Linhong Zhu, Aram Galstyan James Cheng, Kristina Lerman	Unsupervised tri-clustering framework	Online algorithm	Twitter data	It provides efficient running time and storage with gloog clustering of user and tweets in dynamic setting.
2015	Esteban Castillo, Ofelia Cervantes, Darnes Vilarino, David Baez, Alfredo Sanchez	Co-Occurrence graph	Supervised learning algorithm	LiveJournal, SMS	It provides the excellent runtime with centrality measure to discover word related to each sentiment instead of traditional feature like n-gram and vocabulary.
2017	Zhikui Chen, Fei Lu1, Xu Yuan and Fangming Zhong	cross-modal hypergraph model	1.LDA 2.cross-modal hypergraph	Domain datasets (books, DVD)	The parameter adjustment may influence sentiment classification accuracy.
2017	Monali Bordo, Dr. S. K. Biswasloi	ESAGBA model	1.Degree centrality measure 2.polarity assignment technique 3.POS tagging.	Mobile handset reviews	The proposed graph leads to the significant extraction of the semantic relationship between two different nodes.
2017	Xu Yuan, Mingyang Sun, Zhikui Chen, Jing Gao, Pengli.	Embedding Clustering based Deep Hypergraph Model (ECDHG)	1.Pre-training word embeddings 2.Hierarchical fast clustering algorithm	Product reviews	The results validate that our model outperforms the compared methods in the classification accuracy.
2019	Mario Casillo, Fabio Clarizia, Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale	Latent Dirichlet Allocation (LDA) approach.	Word-based graphical model	Twitter and Facebook, Movie Reviews	The built mGT's will be more effective when there are greater number of words.
2019	Jue Liu , Zhuocheng Lu , Wei Du	Enterprise knowledge graph	TransR model	Stock dataset	The accuracy of 8.1% is improved using Enterprise Knowledge Graph when compared to traditional method.
2019	K. Krishnakumari , E. Sivasankar , Sam Radhakrishnan	Doc2Vec feature representation	1.Feature extraction using Doc2Vec. 2.Domain adaptation	Multi-domain sentiment dataset (Book, DVD, Kitchen, Electronics)	It provides a contextual understanding of terms between different domains with large-scale corpus.

4. REVIEW METHODOLOGY

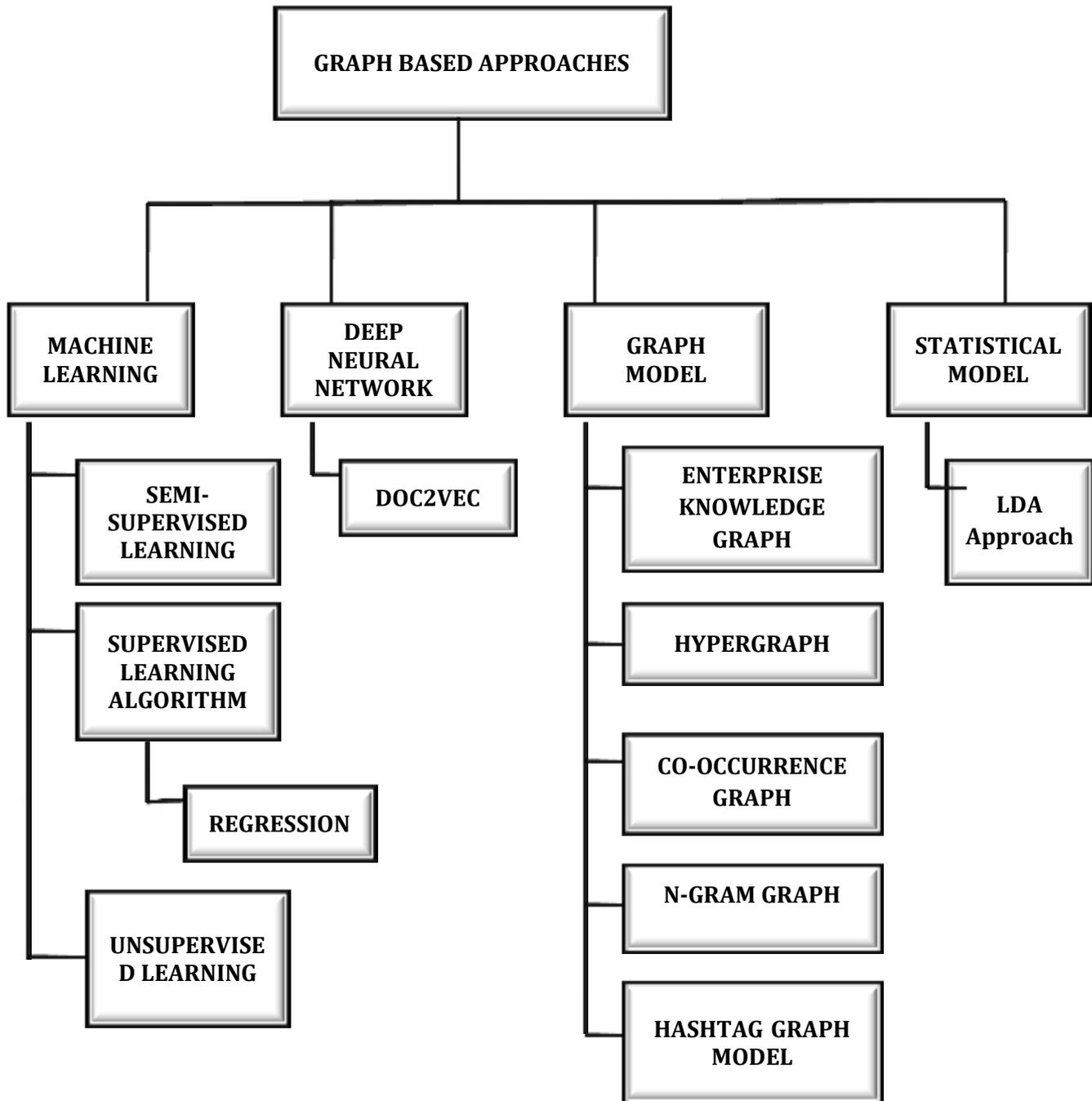


Fig 1: Graph Based Approaches

4.1 Machine Learning Techniques

4.1.1 Supervised Learning Techniques

Supervised machine learning is used to develop a predictive model based on both input and output data or variables. Then, use an algorithm to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when takes a known set of input data and known responses to the data. It leads to training a model to generate reasonable predictions for the response to new data. Supervised learning requires that the data used to train the algorithm should be already labeled with correct answers. The techniques under Supervised

Machine Learning include regression and classification models. Some of them are linear regression and Logistic regression, Decision Trees, Support vector machines, Multi-class classification

#### **4.1.1.1 Regression**

A regression problem is when the output variable is a real or continuous value, such as changes in temperature. The simplest is the linear regression which tries to fit data with the best hyper-plane which goes through the points. Some other regression algorithms include non-linear model, stepwise regression, boosted and bagged decision tree.

#### **4.1.2 Semi-supervised Learning**

Semi-supervised learning is a machine learning which falls between unsupervised learning and supervised. This approach is applied to modals which make use of unlabeled data for training with a small amount of labeled data.

#### **4.1.3 Unsupervised Learning**

Unsupervised learning is a type of machine learning algorithm used to find the hidden patterns in the data and also used to draw inferences from datasets consisting of input data without labeled responses. The technique which is used in this unsupervised learning for clustering such as k-means, k-medoids, Hidden Markov model, Gaussian mixture model, and the neural networks.

### **4.2 Statistical Model**

#### **4.2.1 LDA Approach**

Latent Dirichlet allocation (LDA) is a method used in statistics and machine learning which is used to find a linear combination of features that characterizes two or more class of objects. LDA is also generative statistical model in the Natural Language Processing. It allows sets of observations to be explained by unobserved groups and leads to that some parts of the data are similar. The LDA is widely applied in the fields of machine learning and artificial intelligence. In LDA as a topic, each document is viewed as a mixture of various topics and where each document is considered to have a set of topics. In LDA as a model, using plate notation which is represented as probabilistic graphical model. Then, the dependencies among the variables can be captured concisely.

### **4.3 Deep Neural networks**

It is a specialized form of machine learning. The most of the deep learning methods use neural network architectures, which is why deep learning models are often referred to as deep neural networks. The term deep refers to the number of hidden layers in the neural network but the traditional neural networks only contain 2-3 hidden layers, while deep networks may contain up to 150. The deep learning model with neural network are trained by using large sets of labeled data and which learn features directly from the data without the need for manual feature extraction.

#### **4.3.1 Doc2Vec**

The design for doc2vec is based on word2vec. The doc2vec also uses unsupervised learning approach and which the Word2vec method and the skip-gram neural network are the basis for graph embedding methods. The doc2vec models are used for training, a set of documents is required. The doc2vec is used to create a numeric representation of a document. The word vector is generated for each word and document vector is generated for each document. And the model trains weight for a softmax hidden layer. A new document may be presented and all weights are fixed to calculate the document vector in inference stage.

### **4.4 Graph Model**

#### **4.4.1 Enterprise Graph Model**

The Enterprise Knowledge Graph contains all of an organization data such as structured, unstructured, internal or external and size is unlimited. For example, it is presented as trillions of interlinked facts made available in any combination, on-demand to approved users. It is a representation of an organization knowledge domain or an artifact which can understand by both human and machine and shows how they are related with each other. It improves the organization to take advantage of in-memory computing at cloud-scale to bring immediate access and analysis to everyone and these tools support intuitive, interactive, coherent and transparent query generation for all users. It also provide

context behind AI statistical algorithm to achieve the sought-after cognitive applications and also helps to discover hidden facts of AI and relationship through inferences in integrated content that would be unable to catch an large scale.

#### 4.4.2 Hypergraph

In mathematics, a hypergraph is a generalization of a graph in which an edge can join any number of vertices. Formally, a hypergraph (H) is a pair  $\{H=(X,E)\}$  where X is a set of elements called nodes or vertices, and E is a set of non-empty subsets of X called hyperedges or edges. Hypergraphs have been extensively used in machine learning tasks as the data model and classifier regularization. The applications include recommender systems, image retrieval and for large scale hypergraphs, a distributed framework built using Apache Spark is also available.

#### 4.4.3 Hashtag Graph Model

The hashtag graph  $HG = \{H, E\}$ , in which the edge set E consists of links between hashtags and each edge  $e_{ij}$  represents an undirected link between hashtags  $h_i$  and  $h_j$ , which co-occur in at least one tweet.

#### 4.4.4 N-Gram Graph

The N-gram graph model is mostly used in the statistical natural language processing. An n-gram is a contiguous sequence of n items from a given sample of text or speech in the fields of computational linguistics and probability. Those items can be phonemes, syllables, letters, words or base pairs according to the application and for example which is also used in speech recognition. They are collected from a text or speech corpus. Using Latin numerical prefixes, an n-gram of size 1 is referred to as a unigram, size 2 is a bigram and so on.

#### 4.4.5 Co-occurrence Model

The co-occurrence graph model is semantically related word as the vertices and the edges as representing the relation between two words. It provides the graphic visualization of relationship between any significant co-occurrence words such as people, organizations or any concepts. For example, the co-occurrence words such as biological organisms like bacteria or other entities represented within written material. The generation and visualization of co-occurrence networks helps with the practical advent of electronically stored text amenable to text mining.

### 5. SUMMARY AND CONCLUSION

Due to increasing usage of internet and online activities, the users are sharing the information among them through social media, blogging, chatting, etc. The user posts or tweets becoming trend setter and reviews in the e-commerce websites are influencing the other users to buy the product or services. The sentimental analysis of a products, services or user interests or needs are effectively analyzed using the domain-specific and cross-domain data. The Graph Based Approaches on sentimental analysis provides the efficiency and improved performance in the accuracy of sentiment classification. It also improves the running time of sentiment analysis in both static and dynamic. In dynamic method, the storage and computations methods are also improving greatly.

#### REFERENCE:

- [1] Goldberg, Andrew B., and Xiaojin Zhu. "Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization." Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. Association for Computational Linguistics, 2006.
- [2] Ganesan, Kavita, ChengXiang Zhai, and Jiawei Han. "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions." Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). 2010.
- [3] Wang, Xiaolong, et al. "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.
- [4] Aisopos, Fotis, George Papadakis, and Theodora Varvarigou. "Sentiment analysis of social media content using N-Gram graphs." Proceedings of the 3rd ACM SIGMM international workshop on Social media. ACM, 2011.

- [5] Ponomareva, Natalia, and Mike Thelwall. "Do neighbours help?: an exploration of graph-based algorithms for cross-domain sentiment classification." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.
- [6] Montejo-Ráez, Arturo, et al. "Ranked wordnet graph for sentiment polarity classification in twitter." Computer Speech & Language 28.1 (2014): 93-107.
- [7] Zhu, Linhong, et al. "Tripartite graph clustering for dynamic sentiment analysis on social media." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.
- [8] Castillo, Esteban, et al. "UDLAP: sentiment analysis using a graph-based representation." Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015.
- [9] Chen, Zhikui, et al. "TCMHG: Topic-based cross-modal hypergraph learning for online service recommendations." IEEE Access 6 (2018): 24856-24865.
- [10] Bordoloi, Monali, and S. K. Biswas. "E-commerce sentiment analysis using graph based approach." 2017 International Conference on Inventive Computing and Informatics (ICICI). IEEE, 2017.
- [11] Yuan, Xu, et al. "Semantic clustering-based deep hypergraph model for online reviews semantic classification in cyber-physical-social systems." IEEE Access 6 (2018): 17942-17951.
- [12] Casillo, Mario, et al. "A Latent Dirichlet Allocation Approach using Mixed Graph of Terms for Sentiment Analysis." Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.
- [13] Liu, Jue, Zhuocheng Lu, and Wei Du. "Combining Enterprise Knowledge Graph and News Sentiment Analysis for Stock Price Prediction." Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.
- [14] Krishnakumari, K., Sivasankar, E. & Radhakrishnan, S. Soft Comput (2019).