EPNDR: Emotion Prediction For News Documents Based On Readers' Perspectives

Ramya R S, Madhura K, Sejal Venugopal K R, Iyengar S S, Patnaik L M

Abstract: Due to the rapid rise in Internet population, the content over the web is increasing and a large number of documents assigned by reader's emotions have been generated through new portals. Earlier works have focused only on author's perspective, this work focuses on reader's emotions generated by news articles. In this work, Emotion Prediction for News Documents based on Readers' Perspectives (EPNDR) is proposed More specifically, we form four communities based on the higest ratings that are present in the news articles. Further, a textual relevance is computed based on the word frequency for a particular document and insert all the remaining articles to the four communities. When a new document arrives, the probability of the new document being near to all the documents in a community is found. The emotion rating for the new document is predicted using nearest neighbour analysis. Experiments are conducted on the news articles and as a result, it is observed that the proposed method results in predicting reader's emotions are much better when compared with the existing method Opinion Network Community (ONC) [1].

Keywords: Community, Emotion Detection, Sentiment Analysis, Text Mining, Textual Relevance.

1. INTRODUCTION

Human emotion prediction plays an important role in the developing world. Many Internet users try to convey information through online by expressing their personal opinions than just gaining online information. Social emotions can be expressed in textual format that can be predicted using the terms present in the documents. For example: Happy, cheerful, joyful, etc. convey happiness and on the other side the terms such as depressed, down, regretful convey sad. Based on these terms, the emotion of the text document is predicted. As a result enormous amount of news documents and comments on these documents have published and is updated and shared rapidly through social media services.People on the social media fetch the information and try to enforce their feelings to it by clicking on the different emotions. Likewise, different news documents convey different emotions to readers. Compared with the typical tasks of sentimental analysis, text mining is based on subjective text that conveys the fact but does not give opinion from reader's perspective. Always news articles are written based on what writer thinks and no where the reader's emotion is considered. Human emotion prediction has potential applications on social and economic problems such as political issues and brand perception. In this work, the priority is given to reader and the actual feeling of the reader. The readers' emotion always varies from one person to other. For a news, people may have different emotions. The same news may convey different feeling to another group of people. It is a reader sentiment analysis, in which the opinion prediction is achieved by real time opinion networks. News documents are categorised under four emotions. They are, Moved, Anger, Funny and Sad. Each emotion has certain number of ratings from the user. In the training phase, the communities are formed based on these emotions using textual relevance score. Further, when a new document arrives, textual relevance score is computed and the new document is placed in a proper community. Once the documents are categorised based textual relavance score, using pre-computed semantic distance, the probability of new document being near to the other articles in the community is found. Hence, the probability function is applied to precisely find out the emotion of any news. Emotion detection acts as a reviews to the society. The reviews are important for the people reading the news. It helps in categorizing the data based on emotions. Emotion detection is widely used for short texts like twitter data and comments on the tweets. Machine learning techniques are applied to form the opinion network for twitter data and hence user generated social emotion is predicted. The data set is classified under positive and negative emotional categories. Hence, social emotion detection is value to market analysis and it helps in making political decisions [2] [3]. Emotion analysis is also widely used in product based and e-commerce related companies. Automatic classification of user reviews into different categories is achieved by emotion analysis. Hence, emotion detection plays an important role in online media and is also used as a document classifier. In the previous works [4], the emotion analysis on text corpora is achieved by Latent Dirichlet Allocation (LDA) algorithm, which is used to classify the documents based on positive and negative feelings.

Most of the works that are carried out on emotion analysis detects the emotion from writer's perspective. However, for the news articles, it is necessary to detect the reader's emotion. News articles invoke different meaning to different set of people. Hence, it is necessary to predict reader's emotion. Reader's emotions are different from writer's emotions. When the emotion prediction methodology is applied on the news articles, it is possible to classify the news into good or bad. Hence, emotion prediction from news articles helps to measure the moods of the people in the society. It also helps social scientists quality of the society which is essential for public policy making. In this work, news documents are classified under four different emotion categories using user ratings.

Motivation:

In the existing schemes [1], the performance of the model used for clustering is not much convincible. The model takes the weights as input and communities are formed based on the given input. However, the community structure varies for every restart of the algorithm. Hence, the communites formed are not stable that leads to wrong prediction of emotions in

- *Author name is currently pursuing masters degree program in electric power engineering in University, Country, PH-01123456789. E-mail: author_name@mail.com*
- *Co-Author name is currently pursuing masters degree program in electric power engineering in University, Country, PH-01123456789. E-mail: author_name@mail.com*
  (This information is optional; change it according to your need.)

later stages. In order to overcome the problem. we proposed a community clustering algorithm which is stable and has better performance when compared with existing system.

Contributions:

The principle contributions of this work are as follows:

1) Term vectors for all the news documents are trained according to the Wikipedia word corpus.

2) Semantic distance is computed between the documents using term vector.

3) Textual Relevance (TR) method is proposed that computes the weight of the document based on the emotion ratings that allows to construct the four communities namely Moved, Anger, Funny and Sad.

4) Probability of the documents in a community is computed to obtain the nearest neighbour to predict follow up news social emotion based on the probability score.

Organization:

The rest of the paper is organized as follows: Section II introduces a detailed overview of related works. Section III defines the Emotion Detection Framework. Section IV discusses the performance evaluation. Finally, Section V contains the conclusions.

2 RELATED WORK

Predicting emotions of the text data are achieved with the help of different techniques such as Optimal Network Community, Latent Dirchlet Allocation (LDA), Topic-Level Maximum Entrophy (TME) and Joint Sentiment Topic model (JST).Li et al., [1] proposed a social opinion model to predict the readers' emotions from news articles. Euclidean distance and word mover distance metric is used to find the semantic similarity between the articles. The distance obtained is fed to the algorithm to form the communities. For the new article, the emotion ratings over different emotion labels is found. The model outperforms the other models like Convolutional Neural Network (CNN), Convolutional Neural Network - SVM and Sentiment Latent Topic Model(SLTM). However, the communities formed by the algorithm is not stable, it changes on every restart of the system. Sun et al., [4] aimed at instinctive prediction of social emotion from the massive text. A Latent Dirichlet Allocation (LDA) model is developed that generates a set of latent topics from emotions and further generates affective terms from each topic. LDA improves the performance of social emotion prediction. However, the experiment can be enhanced for thelarge scale online document collections and emotion-aware recommendation of advertisements. Poria et al., [5] presented the sarcasm detection model from the sentences. The sarcasm can be differentiated as positive and negative polarity. Natural language tool kit is used for text classification inorder to find the sarcasm. The sarcasm is detected using a pre trained Convolutional Neural Network (CNN) and obtains the sentiment and emotion inorder to find the sarcasm. Hence the model identifies the sarcasm present in the text using CNN and Support Vector Machine. The experiment is carried for the short twitter data. The model can be enhanced by using large text corpora. Further, to improve the performance of the model, the past tweets of the user can be studied to categorise the sarcastic and non-sarcastic texts. Cambria et al., [6] observed the product reviews with the help of user ratings over different products and it is categorized based on the sentiments. Each review is represented as a distributed vector that is converted into a one-dimensional neural network. Recurrent Neural Network (RNN) with Rated Recurrent Unit (RRU) is used to study the neural network and further, machine learning classifier is applied for the sentiment classification. To achieve the better performance, other sequence learning models like bidirectional RNN and rated feedback RNN for sentiment analysis can be developed. Li et al., [7] [8] [9] considered the spread of subjects on the news forums as subject lifecycles. A sequence network is developed that contains vertex and edges. The vertex indicates the subject and edges represents the relationship. A prediction model is developed to find the attenuation of subject's relationship, interaction and the feedback tendency. The real news datasets are fed into the model. To calculate the coefficient of the model, swarm optimization algorithm is used. Attenuation of the subject and feedback tendency helps to achieve the better performance in the complex situations for the news forums. Interactions of the subjects sometimes leads to negative effects and degrades the performance in the complex situations. Bao et al., [10] [11] [12] proposed the importance of short text such as tweets, questions, instant messages and news headlines. Topic-level Maximum Entropy (TME) model is constructed to classify the emotions over the short texts. Effectiveness of the system is evaluated using real world short documents. However, the system do not retrieve the part-of-speech information for feature generation and also the model does not recommend the event in social media based on emotions. Rao et al., [13] [14] proposed a probabilistic model frame-work based on Latent Dirichlet Allocation, called Joint Sentiment/ Topic model (JST) to fetch the topics and sentiments together from massive text corpora. Since JST is an unsupervised technique, it is adaptable and flexible to other applications. JST represents each document as the bag of words. Therefore, the word ordering is ignored. However, the model can be enhanced by adding higher order information such as bigrams or trigrams. Rao et al., [15] developed a supervised multilabel classification by considering the label frequency and dependency, Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA) are developed. DFLDA is the improvement to the FLDA, where the relationship among the different layers is considered. Gibbs sampler algorithm is used to train the model. Prediction of labels at the word level and simultaneous modelling of the labels increases the efficiency of the supervised model. However, DFLDA is not efficient to handle large text corpora. X. Li et al., [16] [17] [18] considered the truth that bag-of-concepts model are more sophisticated than bag-of-words. Therefore, concept level sentiment analysis is proposed. Data is extracted from the multimodal short video clips and converted to vectors. Feature level fusion and decision level fusion are used to model the multimodal data. Based on the dependency relationship of input sentence, the sentiment is allowed to flow from concept to concept. Feature level modelling is faster when compared with decision level. However, the model do not consider more relevant features.

Meo et al., [19] [20] used microblogs such as Twitter to predict the emotions. Emotions are categorised according to Pletcher's method in which eight emotions are structured into

four dislike pairs such as Joy versus Sadness, Fear versus Anger, Anticipation versus Surprise, and Disgust versus Trust. The first method called lexicon-based method is used to form a dictionary of words without any duplicates and has the separate list of words for each emotion. Content based and Natural Language Processing methods are used to remove stop words from the dataset. The last method used is latent factors model that is based on low dimensional factor model.

Kang et al., [21] proposed a Bayesian inference approach to obtain a basic knowledge of emotion expressions with respect to semantic dimensions. The technique is also used to find the co – occurrence of multiple emotion labels among the different words in the data set. Gibbs Sampling inference is used to generalize emotions from words to documents. Document and Word Emotion Topic (DWET) model represents words and emotions as the two-level hierarchical structure. However, revealing the authors identity is not addressed. Hasan et al., [22] proposed automatic classification of text messages to find the emotions. Twitter messages with hash-tags are used as a dataset. Multiple categories of emotions are found using hash-tags as labels. Lexicon of emotions is used as a solution to large dimensional feature of vectors of message. The proposed model has the accuracy of 90% when compared with SVM, KNN, Decision Tree, and Naive Bayes for classifying Twitter messages. However, temporal nature of the emotions and also emotions that change over time is not predicted. Sykora et al., [23] considered the text from twitter and classified the emotions into positive and negative category. Ontology engineering approach is used to detect the eight high level emotions like anger, confusion, disgust, fear, happiness, sadness, shame and surprise. Grammatically improper and tweets with hashtags are processed using NLP pipeline method. The performance of the ontology-based approach is more effective when compared with lexicon-based approach. The system can be further enhanced for evaluating the larger data set on different systems. Czopp et al., [35] presented the solutions to differentiate the emotions with the help of data driven methods using machine learning techniques. Interactional hypotheses of appraisal theory is used for modelling and interpretable model is developed based on theoretical assumptions that in turn increases the flexibility of the data driven approaches. However, limited operationalization of emotion is investigated. Shivhare et al., [36], presented the importance of emotion detection from large text corpora. Emotion ontology and emotion detector are the two models used for emotion detection. The document and emotion ontology are fed as input to the system and one of the six emotion (love, joy, anger, sadness, fear and surprise) is produced as output. Keyword spotting technique and the use of ontology helps to increase the efficiency of the system. However, the model do not address for larger text corpora involving different categories of data. Hajar et al., [37] considered the YouTube comments as a rich resource of the publicly available text. These comments hold different styles of expression, languages and raise different issues like opinions, stories and emotions that are used for text-based emotion detection. The model uses the algorithms to extract Adjectives, Nouns, Verbs and Adverbs from text and further compute the probabilities for each word. The system has two major advantages: (i) No Labelling is required in the textual data, which is usually a time-consuming task, (ii) The system is flexible enough to allow easy update of the data, and can easily incorporate new ways of expressions and concepts. However, linguistic/semantic features are not considered.

33 EMOTION PREDICTION FOR NEWS DOCUMENTS BASED ON READERS' PERSPECTIVES EPNDR FRAMEWORK

A. Problem Definition and System Model

For a given news document D, the problem is to predict the social emotions of that news. The set of news documents are denoted by D = dn. The list of emotions are denoted by EM = emi, where emi indicates four emotion labels namely Moved, Anger, Funny and Sad. T = tk is the set of terms (word) present in the particular dn document. The ratings over each of the emotion label is indicated by ra = $r_{a,em}$. The list of notions used in the experiment are as shown in Table III: The social opinion model is to represent as quadruple < ev, f t, s, t > where ev is the social event of a news document, f t is the feature text set of the event, s is the result of voting over the emotion labels and t is the time in which the event occured. The aim of the model is to predict s for the new article based on the former quadruples training set . The framework consists of the following phases: 1) Preprocessing Phase (PP) 2) Training Phase (TRP) and 3) Testing Phase (TP). Fig. 1 shows the flow of the entire framework.

1) Preprocessing Phase (PP): In this stage, the raw chinese articles are preprocessed and translated to english language. Further, the stopwords are removed from the news documents and stemming algorithm is applied on these data to perform better results. The model is defined as a quadruple < ev, f t, s, t >. Initially the aim is to extract the terms present in the documents.

TABLE I
COMPARISON OF DIFFERENT EMOTION ANALYSIS APPROACHES

Sl.no.	Authors	Year	Dataset	Emotion Labels	Approach	Lexicon
1.	Mozafari *et al.*, [24]	2019	ISEAR	7 Emotions - Joy, Disgust, Fear, Shame, Guilt, Anger, Sad	Latent Sematic Analysis, Vector Space Model	WordNet
2.	X. Li *et al.*, [1]	2018	News Articles	4 Emotions - Moved, Anger, Funny, Sad	Optimal Network and Word Mover Distance	Emoticons
3 .	Meo *et al.*, [19]	2017	Blogs	8 Emotions - Surprise, Anticipation, Trust, Anger, Disgust, Fear, Joy, Sad,	Random Forest, Logistic Regression	Emotion lexicon, Polarity Lexicon
4.	Kang *et al.*, [21]	2016	Blogs	8 Emotions - Expect, Surprise, Anxiety, Joy, Love, Sorrow, Anger, Hate	Hierarchical Bayesian Model	Latent Factors
5.	Hasan *et al.*, [22]	2014	Blog	4 Emotions - Happy, Inactive, Unhappy Active, Unhappy Inactive	Supervised learning	ANEW lexicon, Emoticons, Punctuations
6.	Calvo *et al.*, [25]	2013	Headlines, Fairy tales	5 Emotions - Joy, Sadness, Anger, Disgust, Fear	Unsupervised Learning,	Emotional Thesaurus, Bag of words
7.	Wang *et al.*, [26]	2013	ISEAR	8 Emotions- Guilt, Shame, Fear, Joy, Anger, Disgust, Sad, Surprise	LSA Algorithm	Scientific Library
8.	Sykora *et al.*, [23]	2013	Blogs	8 Emotions - Surprise, Sad, Confusion, Anger, Disgust, Fear, Joy, Shame	Ontology Approach, Custom NLP Pipeline	Emotive ontology Lexicon, Feature IntensifiersS
9.	Purver *et al.*, [27]	2012	Blogs	6 Emotion - Surprise, Anger,Joy, Sad, Disgust, Fear	SVM	Hashtags, emoticons
10.	Wang *et al.*, [28]	2012	Blogs	7 Emotion - Sad, Anger, Joy, Love, Fear Thanks, Surprise	Multi Nomial Naive Bayes	MPQA lexicon, ngrams, POS, Affect words
11.	Roberts *et al.*, [29]	2012	Blogs	7 Emotion - Anger, Love, Fear, Disgust, Joy, Sad, Surprise,	Supervised Learning	WordNet synset and hypernyms, punctuations, topic scores
12.	Balabantaray *et al.*, [30]	2012	Blog Data	6 Emotion- Anger, Sad, Fear, Disgust, Joy, Surprise	Multi class SVM	n-grams, WordNet Affect, dependency parsing
13.	Chaffar *et al.*, [31]	2011	News,Fairy Tales, Blogs	6 Emotion- Disgust,Fear,Sad,Surprise,Happiness,Anger	Supervised learning	N-grams and bag of words, WordNet Affect
14.	Balahur *et al.*,, [32]	2011	ISEAR, Documents	10 Emotion - Anticipation, Surprise, Trust, Disgust,Anger,Fear, Sad,Joy, Shame, Guilt	Common Sense Knowledge	EmotiNet
15.	Ghazi *et al.*, [33]	2010	Web-blogs, children Stories	7 Emotions - Sad, Joy, Anger, Surprise, Fear, Disgust, Neutral	Hierarchical classifier, Support Vector Machine	Bag of words, Polarity Feature Set, Prior Polarity Lexicon,
16.	Bellegarda *et al.*, [34]	2010	News Headlines	6 Emotions-Surprise, Anger, Joy, Sad, Fear, Disgust	Supervised Learning, Latent Semantic Mapping (LSM)	WordNet synset, Bag of words, WordNet Affect

The raw feature text ft is processed as a Bag of Words (BOW) irrespective of the grammer and the word order. BOW is a word embedding method that makes machine learning algorithm to understand the words. BOW finds the multiplicity of the words and the term frequency is calculated to find out the occurance of each word in the documents. BOW ft should be normalized based on the assumption that readers concern towards each news documents are equal. Hence, a normalized histogram is obtained at this stage with finite

dimensional vector with their non negative co-ordinates sum is equal to one.

2) Training Phase (TRP): The training phase includes 2 stages:

(i)Community Formation (ii)Textual Relevance Computation.

(i) Community Formation:

Intially, in training phase, the news articles with user ratings are considered. They are categorised based on the four emotions - Moved, Anger, Funny and Sad. Hence, four groups of articles are formed. Further, four communities are formed and one document with highest rating for a particular emotion is put in each community. Hence, four communities consisting of one document in each group is constructed. During later stages using Textual Relevance score, remaining documents are placed in each communities.

(ii) Textual Relevance Computation:

Textual Relevance for each documents are computed based on the term frequency in each document. It is mainly computed to realize the importance of a term in the document. Initially, Term Frequency (TF) for all the terms in the document is computed. Further, textual relevance is calculated using computed TF. Let Wt,C denote the weight of the term t in a community C as shown in Equation (1)

$$W_{t,C} = (1-\xi)\frac{tf_{t,C}}{|\,C.T\,|} + \xi\frac{tf_{t,D}}{|\,D\,|} \qquad (1)$$

where t $f_{t,C}$ is weight of the terms in a community C.T and t $f_{t,D}$ weights of the terms per article D. Here, D represents the input news documents and ξ is smoothing parameter. Using user ratings present in the documents, Textual Relevance score is computed for all the documents over four emotions. Hence one document has four textual relevance values. The document belongs to a particular emotion community for which it has highest Textual Relevance value. Hence, four textual relevance values for a document are compared to find the highest value and placed into a community. Likewise, Communities are formed for all the documents.

3) Testing Phase: Testing phase includes (i) Word Mover Distance Calculation and (ii) Probability Calculation.

(i) Word Mover Distance Calculation: In this work, we try to leverage the results of Word2Vec [38] which show that the model learn high quality embedding of terms by co-occurrences in the sentences. Word2Vec is a neural network that process the text and converts them into feature vectors. Input given to the model is large text corpus and the output obtained is the feature vectors. The model can be applied on different groups of text after the removal of stopwords. The main advantage of the model is that the similar word vector representation will always be together. Using mathematics, the model identifies the similarity between the words. To transfer a word into vector, the human intelligence is not needed. It is purely done using machine learning techniques. Hence, using the vector representation of the words, the semantic distance between the two news documents d1 and d2 is found by Euclidean Distance (ED) and Word Movers Distance metrics. Euclidean Distance, d(i, j) is computed as shown in Equation (2).

$$d_{(i,j)} = \| t_i - t_j \|_2 \qquad (2)$$

where d(i,j) is the distance between term i and term j and ti and tj denotes the corresponding term. The meaningful distance between two documents is computed by Word Mover Distance.

TABLE II
COMPARISON OF DIFFERENT EMOTION ANALYSIS APPROACHES

Sl.no.	Authors	Year	Algorithm	Concept	Advantages	Disadvantages
1.	Mozafari *et al.*, [24]	2019	STASIS (Emotion detection by semantic nets and corpus statistics) and Vector Space Model(VSM)	Similarity between short texts is determined using STASIS and cosine similarity for detecting emotions is achieved by (VSM)	VSM results in creation of feature vectors, hence the results obtained are efficient.	STASIS method does not work well with short texts
2.	Li *et al.*, [1]	2018	Optimal Network Community and Word Mover Distance	Communities are formed using Word Mover Distance and Emotion is detected using nearest neighbour analysis	Semantic similarity calculation makes the model to produce efficient results	Communities formed by the algorithm is not stable
3.	Cambria *et al.*, [6]	2016	Convolutional Neural Network and Recurrent Neural Network with Gated Recurrent Unit (RNN)	CNN reviews embedding with label scores and RNN with Gated Recurrent Unit to produce vectors.	-	RNN does not yield better performance
4.	Sun *et al.*, [4]	2016	Swarm optimization (PSO) algorithm	PSO forms Subject network. It represent the casual relationship based on which a prediction model constructed.	Effective performance in the complex situations in the news forums	Negative effects results in degrading the performance.
5.	Bao *et al.*, [13]	2015	Latent Dirichlet Allocation (LDA)	Affective terms are generated by LDA, creates the emotions from emotion distribution. A latent topic generates a document at the end.	Meaningful latent topics for each emotion is obtained from LDA	LDA is not efficient enough for large scale dataset.
6.	Sykora *et al.*, [23]	2013	Ontology and NLP Pipeline	Ontology engineering approach is used to detect levels emotions. Grammatically improper and tweets with hashtags are processed using NLP pipeline method.	Ontology approach has better performance when compared with lexicon approach.	The system do not evaluate for larger dataset on different systems.
7.	Rao *et al.*, [15]	2014	Joint Sentiment/Topic Model (JST)	JST helps to identify the topic and sentiment of a document simultaneously.	JST is adaptable and flexible for other applications	Since JST represent the text as bag of words, word ordering is ignored.
8.	Li *et al.*, [16]	2015	Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA) and Gibbs Sampler algorithm	FLDA is developed using the label frequency and in order to keep track of dependency among the different labels DFLDA is developed.	Label frequency and label dependency make the model efficient	DFLDA is prone to parameter explosion.

TABLE III
THE LIST OF NOTIONS

Notations	Discription
D	All the news set
$T = t_k$	Term set of documents
$EM = em_i$	Set of emotions
$W_{t,C}$	Weigth of word t in Community C
$d_{(i,j)}$	Distance between document d_i and d_j
$r_{(d,em_k)}$	The set of emotion ratings

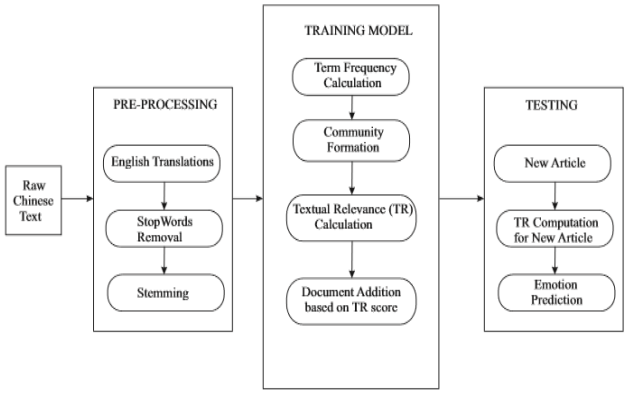

Fig. 1. Emotion Prediction for News Documents based on Readers' Perspectives Architecture EPNDR

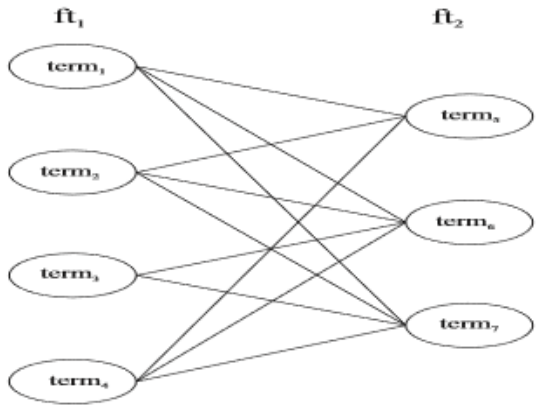

Fig. 2. Flow Diagram

Intially, let us assume that the term i in f t_1 can transformed partially or completely into any word in ft_2 as shown in Fig. 2. Flow matrix exists ie. $M \in \Re^{n*n}$, the element $M_{ij} \geq 0$ denotes the term i in ft1 is switched to term j in ft_2. There are two conditions for transforming ft_1 completely to ft_2. They are:

a) The sum of outgoing flow from term i equals ft_{1i}, $\sum_j M_{ij} = ft_{1i}$;
b) The sum of incoming flow from to term j equals ft_{2j}, $\sum_j M_{ij} = ft_{2j}$;

For example, let us consider three sentences from three different documents.

d_1 = Obama speaks to the media in millions.
d_2 = The president greets the press in Chicago.
d_3 = The band gave concert in Japan.

The stopwords are removed from the above sentences. The task of word mover distance is to move the terms present in the document d2 to the document d1 and d3 at the cost of the calculated distance d(i,j). If the cost of moving is less, then it is concluded that those documents are semantically near. In the above example, the cost of moving the terms from the document d2 to d1 is significantly low when compared with moving the document from d2 to d3 as shown in Fig. 3. Hence, The sentence d1 and d2 consists of different words however they are semantically related to each other. Probability calculation phase involves computation of distance between the documents. In order to compute the distance, the terms of the documents are converted to vectors using Word2Vec model. Further, the distance between each term in the document to all the other terms in the other document is computed using Euclidean Distance (ED). Further, optimal transport distance is computed in

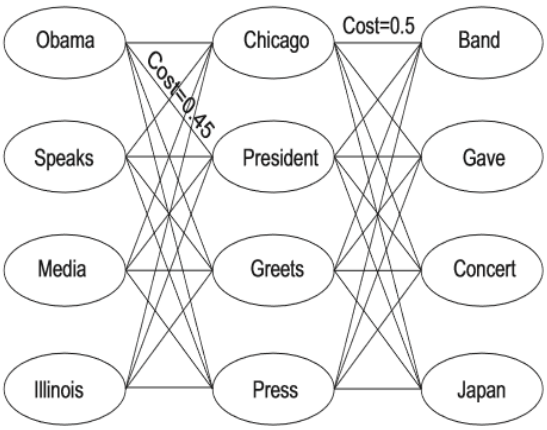

Fig. 3. Example for Word Mover Distance Diagram

order to find out the semantic similarity between the document. The computed distance is used as weights for each document.

(ii) Probability Calculation: Probability is computed for testing document to predict the nearest neighbour of the input document. We consider two documents d1 and d2 in a community. The document d1 selects d2 as approaching nearest neighbour with probability Pr(d1,d2). If both the documents are closer then the probability value is larger else it is smaller. The Equation (3) can be written as,

$$Pr(di,j) = \frac{exp(-dist_{i,j}^2/2\sigma^2)}{\sum_{k \neq i} exp(-dist_{i,k}^2/2\sigma^2)} \qquad (3)$$

where Pr(di,j) is the probability of document i being near to j. disti,j is the distance between two document i and j. σ defines the normal distribution variance and k is the number of documents in a community

EXPERIMENT EVALUATION

4.1 Data collection

The Yanghui Rao's corpus is a collection of news documents from the society channel of Sina. In total 4000 news documents are considered among which 2400 documents are used for training the model and 1600 documents are used for testing purpose. Each document comprises of publication date, news title, news id, total number of ratings and ratings over each of the four emotion lables. Since the documents are in chinese language, the data is converted to English before modelling. In pre-processing stage, Word2Vec model is constructed by wikipedia corpus.

4.2 Experiment Setup

The proposed Textual Relevance (TR) method produces the stable communities for the documents. Hence, Optimal Network Community (ONC) [1] formation method is used as a baseline for comparison. In the baseline method, communities are formed based on the distance computed by optimal transport distance. The communities that are formed using distance score are not stable. Communities were varying everytime when the algorithm is restarted. In baseline method, the threshold to prune the network is set manually which results in improper community formation. Hence, in this work,

the Textual Relevance (TR) method forms the communities based on the ratings of the user. All the documents are considered for community formation without any pruning. Once the training of data is completed, the result for any new document is determined over four emotions Moved, Anger, Funny and Sad for both the methods. The results are compared using two metrics:
Accuracy@1(ACC) and P earson−Correlation(P C).

.

4.3 Performance calculation

In this section, emotion detection results are compared and discussed for both ONC and Textual Relevance methods. The experiment is carried out on a 8GB RAM, Intel(R) Core(TM) i5-7000U CPU @2.50GHz 2.70GHz processor system. The higher configuration system may yield better results. The dataset discussed in data col-lection is used as an evaluation for both the methods. Emotion Ratings over Moved, Anger, Funny and Sad

Algorithm 1: Emotion Prediction for News Document based on Readers' Perspectives s (EPNDR)

Input: Document Set D = d_n with ratings over four emotion labels EM = em_i

Output: Prediction of em_i for the new document

(1) Phase I: Training

(2) Pre-Processing all the documents in the Document set D.
(3) Convert the Terms T = t_k to vectors for each document D, using Word2Vec Model.
(4) For each em_i of Document D, compute the highest values of user rating $r_{a,\,em.}$

(5) For each emotion em_i, consider one document d_i with highest value $r_{a,em}$ and place it in particular emotion community C.
(6) For remaining d_{n-1} documents, repeat the below steps to compute Textual Relevance (TR).

(7) For each d_{n-1} documents, Compute the Term Frequency T = t f_{tD}.

(8) For each community C, Compute the Term Frequency for particular community, T = t f_{tC}
(9) Compute the Textual Relevance using Equation. 1

(10) Repeat the following steps for Community formation for remaining d_{n-1} documents.

(11) For each Textual Relevance value, compute the highest value.

(12) Add the document D to a particular community.

(13) Phase II: Testing

(14) For a new input document D = d_{new},
(15) Preprocessing
(16) Convert the terms T = t_k to vectors.
(17) Compute the similarity between the D = d_{new} document and D = d_{n-1} using Equation 2.

(18) For a new document D = d_{new}, and Distance D = dnew,i
(18) Find the probability of new document being near to other

n − 1 documents in all the community.

(20) For all probability values for a new document, select P r(di, j) which is equal to 1.

---

emotions is considered as performance metric. 1600 are used for testing. The ACC and P C method are used for evaluation in both the experiments.

According to ACC the predicted emotion compared with the actual emotion rating list is calculated using Equation

(4). If the predicted emotion is identical to the top-rated emotion in the actual list, then ACC is equal to 1. If two emotion labels have same rating, then any one of them can be considered.

$$Acc_d@1 = \begin{cases} 1 & \text{if } e_{pr} = EM_{top} \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

$Acc_{@}1$ is determined with the number of correctly predicted documents versus total number of documents as in Equation (4).

$$Acc@1 = \frac{\sum_{d \in D} Acc_d@1}{|D|} \qquad (5)$$

Here, d is a particular document in the set of documents D. In order to find the ACC, the ratings are not compared. Just the emotion label is compared. For example, while testing, if the predicted emotion is" Moved" with rating 40 and actual top-rated emotion for the same document is" Moved" with rating 60. Here the result ACC is considered as 1 inspite of rating mismatch.

Emotional distributions are not considered in ACC. The reaction of the people of a news is not clear. It comprises of concentrated emotions. Hence, to measure the correlation among the news documents P earson−Correlation (PC) metric is calculated as shown in Equation (6) and

(7). Here, the predicted votes for the testing document is compared with the actual vote. The value of the PC varies from -1 to +1. where, +1 indicates the best correlation.

where PC is the P earson − Correlation that measures the correlation between predicted distribution over the actual distribution. To compare the efficiency of TR and ONC method, reducing network strategy is used. Accordingly, the experiment is repeated by varying the size of training data. The variation of tranining data is represented in the graph (both the graphs) for ACC and P C. The horizontal axis shows the percent of training data and verticle axis shows the ACC and P C values. The percentage of training data plays a very important role in

emotion prediction. The model produces the best result only when the training data is more. Later the accuracy starts reducing. Hence the model should be properly trained in order achieve the best results. From the Fig. 4 it is clear that the model gives accurate results for ACC upto 40% of training data, later the accuracy of model starts reducing. Similarly, the results for P C is shown in Fig. 5, the experiment gives the best results upto 35% of training data. Later the value of P C starts reducing. So, it is necessary to consider proper percentile of training data for the experiments.Fig. 6 shows the number of correctly predicted documents in both ONC and TR methods. The number of documents predicted for each emotion for TR method is more when compared with ONC. Testing documents are placed into proper community with the help of Term Frequency. The percentage of correctly predicted documents for four communities namely “Funny”, “anger”, “sad” and “moved” are 73%, 85%, 78% and 90% respectively. The communities formed by TR score method is efficient and reliable when compared with ONC method. In ONC, the communities formed are not constant and varies with every restart of the algorithm. A threshold value is manually set to prune the network in ONC method and it results reduced prediction in ONC method. In TR method the communities are formed by user ratings and they are constant throughout the execution. Hence, the proposed TR method outperforms the ONC method.

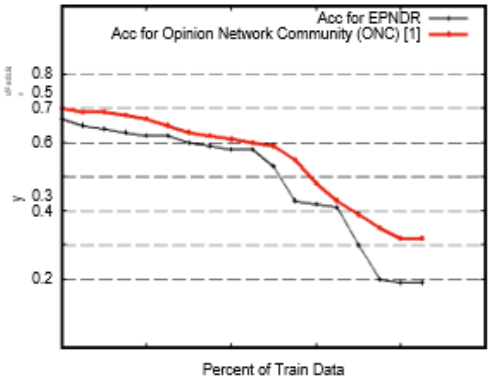

Fig. 4. Comparison of Accuracy results.

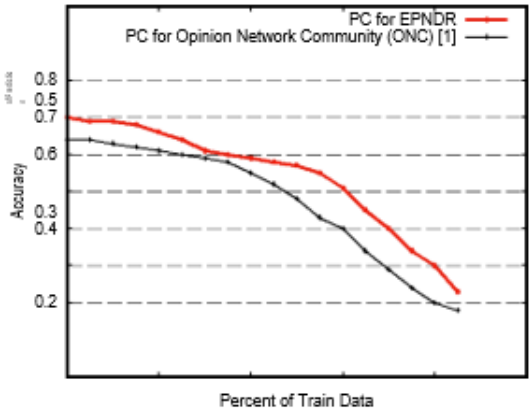

Fig. 5. Comparison of PC results.

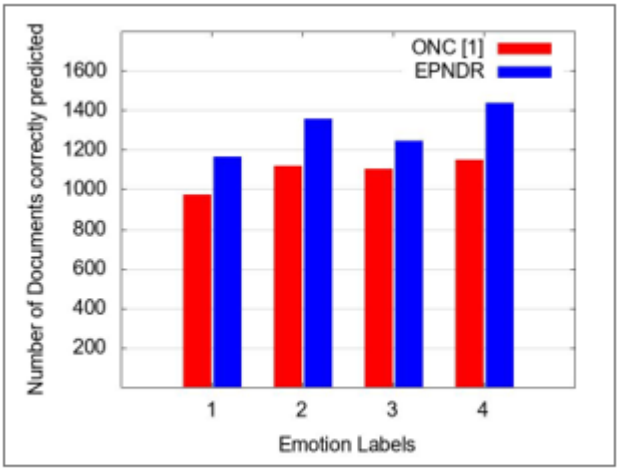

Fig. 6. Number of Documents Correctly Predicted.

CONCLUSION

Social emotion prediction is of value to market analysis and to political decision. Due to the rapid development of Web, large numbers of documents assigned by readers’ emotions have been generated through new portals. Earlier studies focused on only writers perspective. In this paper, we analyze the social opinions for the news documents considering authors’ perspective. In this work, Social Emotion Prediction based on Readers’ Perspectives for News Documents (EPNDR) for measuring similarity among news document is proposed. The news document with various emotions are extracted. TR is computed for each document using the term frequency. Based on the TR score the document is placed into Moved, Anger, Funny and Sad communities. Further, a probability metric is applied on each community to predict exact nearest emotion for the input testing of new document. The communities formed by TR method is stable and reliable. The performance of TR model is evaluated using ACC and P C with 1600 news documents. The proposed method outperforms the ONC model
[1] with stable communities and accurate emotion detection results. In future, we would like to investigate deep learning techiques to find the similarity between the documents by reducing the computation time.

REFERENCES

[1] X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang, “Predicting Social Emotions from Readers’ Perspective,” IEEE Transactions on Affective Computing, vol. 10, no. 2, pp. 255–264, 2017.
[2] V. H. Bhat, P. G. Rao, R. Abhilash, P. D. Shenoy, K. R. Venugopal, and
L. Patnaik, “A Novel Data Generation Approach for Digital Forensic Application in Data Mining,” In Proceedings of Second International Conference on Machine Learning and Computing (ICMLC), pp. 86–90, 2010.
[3] K. R. Venugopal and R. Buyya, “Mastering C++,” Tata McGraw-Hill Education, 2013.
[4] Z. Sun, Q. Peng, J. Lv, and J. Zhang, “A Prediction Model of Post Sub-jects Based on Information Lifecycle in Forum,” Information Sciences, vol. 33, no. 7, pp. 59–71, 2016.

[5] S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Net-work Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis," In Proceedings of the V International Conference on Empirical Methods in Natural Language Processing, pp. 2539–2544, 2015.
[6] E. Cambria, N. Howard, Y. Xia, and T.-S. Chua, "Computational Intel-ligence for Big Social Data Analysis," IEEE Computational Intelligence Magazine, vol. 11, no. 3, pp. 8–9, 2016.
[7] S. Desai, V. Chandrasheker, V. Mathapati, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "User Feedback Session with Clicked and Unclicked Documents for Related Search Recommendation," IADIS-International Journal on Computer Science and Information Systems, vol. 11, no. 1, pp. 81–98, 2016.
[8] V. H. Bhat, P. G. Rao, R. Abhilash, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, "A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application," International Journal of Engineering and Technology, vol. 2, no. 3, pp. 313–319, 2010.
[9] A. Ramachandra, S. Abhilash, R. K B, and K. R. Venugopal, "Feature Level Fusion based Bimodal Biometric using Transformation Domine Techniques," IOSR Journal of Computer Engineering (IOSRJCE), vol. 3, no. 3, pp. 39–46, 2012.
[10] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu, "Mining Social Emotions from Affective Text," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1658–1670, 2011.
[11] R. S. Ramya, T. G. Singh, D. Sejal, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "DRDLC: Discovering Relevant Documents using Latent Dirichlet Allocation and Cosine Similarity," In Proceedings of VII International Conference on Network, Communication and Computing, pp. 87–91, 2018.
[12] R. S. Ramya, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "Feature Extraction and Duplicate Detection for Text Mining: A Survey," Global Journal of Computer Science and Technology, vol. 16, no. 5, pp. 1–21, 2017.
[13] Y. S Rao, "Contextual Sentiment Topic Model for Adaptive Social Emotion Classification," IEEE Intelligent Systems, vol. 31, no. 1, pp. 41–47, 2015.
[14] P. D. Shenoy, K. Srinivasa, and L. M. K R Venugopal, Patnaik, "Dynamic Association Rule Mining using Genetic Algorithms," Intelligent Data Analysis, vol. 9, no. 5, pp. 439–453, 2005.
[15] Y. Y Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment Topic Models for Social Emotion Mining," Information Sciences, vol. 266, no. 7, pp. 90–100, 2014.
[16] X. X Li, J. Ouyang, and X. Zhou, "Supervised Topic Models for Multi-Label Classification," Neurocomputing, vol. 149, no. 9, pp. 811–819, 2015.
[17] D. Sejal, T. Ganeshsingh, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "ACSIR: Anova Cosine Similarity Image Recommendation in Vertical Search," International Journal of Multimedia Information Retrieval, vol. 6, no. 2, pp. 143–154, 2017.
[18] K. R. Venugopal, K. Srinivasa, and L. M. Patnaik, "Soft Computing for Data Mining Applications," Springer, 2009.
[19] R. Meo and E. Sulis, "Processing Affect in Social Media: A Comparison of Methods to Distinguish Emotions in Tweets," ACM Transactions on Internet Technology (TOIT), vol. 17, no. 1, p. 7, 2017.
[20] V. H. Bhat, V. R. Malkani, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, "Classification of Email using Beaks: Behavior and Keyword Stemming," In Proceedings of IEEE Region 10 Conference TENCON, pp. 1139–1143, 2011.
[21] X. Kang and F. Ren, "Understanding Blog Author's Emotions with Hierarchical Bayesian Models," In Proceedings of the 13th International Conference on Networking, Sensing, and Control (ICNSC), pp. 1–6, 2016.
[22] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting Emotions in Twitter Messages," In Proceedings of the 9th International Conference on Networking, Sensing, and Control (ICNSC), pp. 10–16, 2014.
[23] M. D. Sykora, T. Jackson, A. O'Brien, and S. Elayan, "Emotive Ontol-ogy: Extracting Fine-Grained Emotions from Terse, Informal Messages," In Proceedings of the 9th International Conference on Networking, Sensing, and Control (ICNSC).
[24] F. Mozafari and H. Tahayori, "Emotion Detection by Using Similarity Techniques," In Proceedings of the seventh Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp. 1–5, 2019.
[25] R. A. Calvo and S. Mac Kim, "Emotions in Text: Dimensional and Categorical Models," Computational Intelligence, vol. 29, no. 3, pp. 527–543, 2013.
[26] X. Wang and Q. Zheng, "Text Emotion Classification Research Based on Improved Latent Semantic Analysis Algorithm," In Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering, 2013.
[27] M. Purver and S. Battersby, "Experimenting with Distant Supervision for Emotion Classification," In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 482–491, 2012.
[28] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter Big Data for Automatic Emotion Identification," In Proceedings of the International Conference on Privacy, Security, Risk, Trust and on Social Computing, pp. 587–592, 2012.
[29] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, "Empatweet: Annotating and Detecting Emotions on Twitter." Lrec, vol. 12, pp. 3806–3813, 2012.
[30] R. C. Balabantaray, M. Mohammad, and N. Sharma, "Multi-Class Twitter Emotion Classification: A New Approach," International Journal of Applied Information Systems, vol. 4, no. 1, pp. 48–53, 2012.
[31] S. Chaffar and D. Inkpen, "Using a Heterogeneous Dataset for Emotion Analysis in Text," Canadian Conference on Artificial Intelligence, pp. 62–67, 2011.
[32] A. Balahur, J. M. Hermida, A. Montoyo, and R. Munoz,˜ "Emotinet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories," International Conference on Application of Natural Language to Information Systems, pp. 27–39, 2011.
[33] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Hierarchical Versus Flat Classification of Emotions in Text," Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 140–146, 2010.

[34] D. Ghazi, D. Inkpen, and S. Szpakowicz, “Hierarchical Versus Flat Classification of Emotions in Text,” Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 140–146, 2010.
[35] J. R. Bellegarda, “Emotion Analysis Using Latent Affective Folding and Embedding,” In Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 1–9, 2010
[36] A. M. Czopp, A. C. Kay, and S. Cheryan, “Positive Stereotypes are Pervasive and Powerful,” Perspectives on Psychological Science, vol. 10, no. 4, pp. 451–463, 2015.
[37] S. N. Shivhare and S. K. Saritha, “Emotion Detection from Text Documents,” International Journal of Data Mining & Knowledge Man-agement Process, vol. 4, no. 6, p. 51, 2014.
[38] M. Hajar et al., “Using Youtube Comments for Text-Based Emotion Recognition,” Procedia Computer Science, vol. 83, pp. 292–299, 2016.
[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.