

A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification and Emotion Recognition

Komal D. Murkute, B. R. Bhamare

Department of Information Technology, Sinhgad College of Engineering, Pune, Maharashtra, India

Department of Information Technology, Sinhgad College of Engineering, Pune, Maharashtra, India

ABSTRACT: Opinion mining (also known as sentiment analysis) refers to the use of natural language processing, text analysis and computational linguistics to discover and dig out subjective information in source resources. However, typically a sentence was split as a word sequence in previous sentiment classification algorithms, which is difficult for efficiently handling the unpredictable sentiment polarity between a phrase and the words. The sentence level sentiment classification generates segmentations of a sentence and predicts sentence-level polarity simultaneously based on the segmentation results. This framework includes a candidate generation model which is used for generating segmentation of a sentence, a segmentation ranking model to calculate the ranking score for each segment, and a classification model for forecasting the sentiment polarity of a segmentation using WordNet. It gives meaningful segmentation to predict the positive, negative or neutral polarity. This system also performs mood analysis or emotion recognition of a sentence (happy, sad and afraid).

KEYWORDS: Joint Segmentation and Classification, Natural Language Processing, Sentence Level Classification, Mood Analysis.

I. INTRODUCTION

Sentiment analysis is the mechanized mining of states of mind, feelings and suppositions from content, discourse and database sources through Natural Language Processing (NLP). Now days on web the number of reviews, suggestions, feedbacks are increasing in enormously manner. Because every person wants to share his views and experience about the product like review on product, review on movie, person tweets etc. Reviews play vital role in helping and suggesting other person in their decision making. But on the other hand it becomes difficult to read all reviews and make decision as per. Thus, mining this data, identifying the user opinions this is done by performing detailed sentiment analysis on the data. The fields of opinion mining and sentiment analysis are distinct but deeply related. Opinion mining focuses on polarity detection [positive, negative or neutral] whereas sentiment analysis involves emotion recognition. Because detecting the polarity of text is often a step in sentiment analysis, the two fields are usually combined under the same umbrella.

The sentiment classification approaches can be categorized into: (i) machine learning (ii) lexicon based and (iii) hybrid approach. For guessing the polarity of sentiments based on training and testing data sets, the machine learning approach is mainly used [4]. A pipelined framework with two stages was mainly used in the previous studies. In the first stage, the segmentation result of a sentence with bag-of-words or a separate text analyser like standard syntactic chunker was produced. In the second stage, they take the segmentation results as input and classification algorithm was mainly used to build sentiment classifier. But this pipelined framework was having some drawbacks and suffers from the difficulty of error transmission, since the errors generated from sentence segmentation cannot be corrected. A typical kind of error is caused by the unpredictable sentiment polarity between a phrase and the words it contains, like {"not bad," "bad"}. The bag-of-words and syntactic chunkers are not good enough to handle this polarity conflict phenomenon. [1]

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The joint framework for Sentence Level Sentiment Classification is proposed for generation and prediction of useful segmentation simultaneously. It mainly deal with the error circulation problem which was present in pipelined framework. It provides improved performance as compare to pipelined framework. The purpose is to define the sentiment polarity (positive, negative, neutral) of a sentence based on its textual content and emotion recognition of a sentence as happy, sad or afraid.

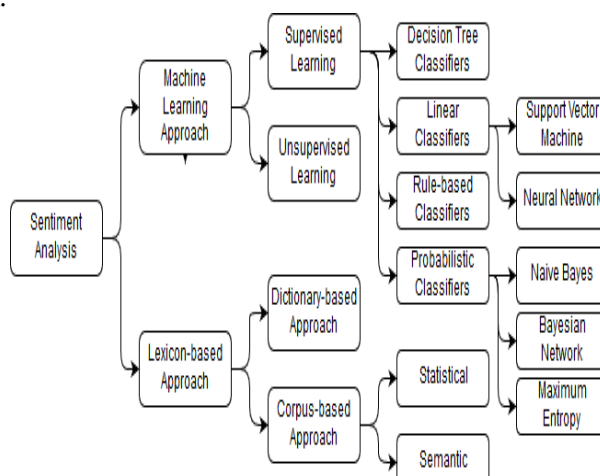


Figure 1 Classification of Sentiment Analysis

II. RELATED WORK

Tang et al [3] has proposed sentiment-specific word embedding's called as sentiment embedding's. Feeling of writings was overlooked in past word inserting calculation and just the setting of words were utilized so it gets to be distinctly troublesome for opinion investigation on the grounds that the words having inverse extremity however with comparable settings, are mapped to adjoining word vectors. In this system, sentences and words of sentiment information of texts are encoded together with contexts of words in sentiment embedding's and existing problem is solved. The adjacent neighbours in sentiment embedding space are semantically related by combining context and sentiment level evidences and it favours words with same sentiment polarity.

Tang et al [2] has implemented a new model called User Product Neural Network (UPNN) to capture user-and product-level information for sentiment classification of documents (e.g. reviews). Only semantics of texts was used in the existing neural network models, while users who express the sentiment and products which are evaluated was ignored. In this paper this problem is addressed by integrating user-and product-level information in neural network for document-level sentiment classification. A variable-sized document, the review written by the user and the product which is evaluated is taken as an input in UPNN and sentiment polarity label of a document is provided as an output.

Duyu Tang et al [6] developed a large-scale sentiment lexicon from twitter with a representation learning approach. Mostly propagation methods were used to estimate the score of each phrase in the existing sentiment lexicon learning algorithm yet this data is not reasonable for building large-scale sentiment lexicon. So, a dense, low-dimensional, and real-valued vector called as a sentiment-specific phrase embedding (SSPE) is learned, by encoding the syntactic contexts and sentiment related information into the representation of continuous phrases. As a result, the nearest neighbours having similar semantic usage also having the same polarity in the embedding space of SSPE are chosen.

Mikolov et al [8] has introduced an algorithm called Paragraph Vector that is used for studying fixed-length representations of features from variable-length pieces of texts, such as sentences, documents and paragraphs. Various machine learning algorithms need as an input a feature vector of fixed-length. When it comes to texts, bag-of-words is used as most common fixed-length features. There are two drawbacks related to these features as they fail to deal with the ordering of the words and they also ignore semantics of the words. In this work, each document is denoted by a

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

dense vector which is skilled to guess words in the document. The algorithm is mainly used to overcome the flaws of bag-of-words model.

Tang et al [5] proposed an approach for learning sentiment specific word embedding (SSWE) for sentiment analysis. Sentiment information is encoded in to the persistent portrayal of words, so that great and awful can be separated to inverse closures of the range. For learning constant word portrayals ordinarily the syntactic setting of words was demonstrated in many past calculations however the sentiment of text was neglected. This is troublesome for notion investigation since words having comparative syntactic setting yet inverse extremity are mapped by them, similar to great and terrible, to neighbouring word vectors. With the help of sentiment specific word embedding (SSWE), this problem is addressed by encrypting sentiment related data in the representation of non-stop words.

Saima Aman et al [9] introduced an emotion annotation task of identifying emotion category, emotion intensity and the words/phrases that indicate emotion in text. In this work annotation scheme is introduced and results of an annotation agreement study on a corpus of blog posts is presented. Also computational techniques is explored for emotion classification. In this experiment, knowledge-based approach is used for automatically classifying emotional and non-emotional sentences. Here data used is emotion-rich data which come from blogs.

III. PROPOSED SYSTEM

Sentence-level sentiment classification is a basic and broadly studied area in sentiment analysis. The main purpose is to define the polarity (positive, negative, neutral) of a sentence based on its textual content. A joint framework for sentence-level sentiment classification is proposed, which all the while produces segmentation results of sentence and by utilizing this outcome sentence-level polarity is predicted and also emotion (happy, sad or afraid) related to sentence is recognised.. Three models are developed (1) a candidate generation model to generate the segmentation of a sentence, (2) a segmentation ranking model to calculate ranking score for each segment of a sentence, and (3) a classification model to predict the sentiment polarity of each segmentation also Emotion recognition contributed to find out mood related to sentence. There are two main process of Joint Framework namely Prediction Process and Training Process. Prediction Process is mainly used to predict sentence polarity. Training Process is used to train the ranking model from sentences with only sentiment polarity labels and for obtaining better classification performance [1].

SYSTEM ARCHITECTURE

System Architecture deals with overall flow of the entire system. A system architecture is a theoretical model that outlines the behaviour, structure, and more outlooks of a system. An architecture is a formal clarification and demonstration of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

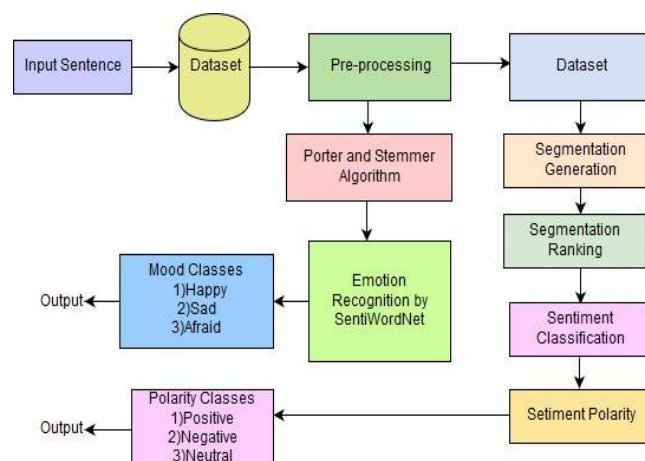


Figure 2 System Architecture

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

System architecture mainly consists of Candidate generation, Segmentation Ranking, Sentiment classification and Emotion Recognition models. Also it consists of training and Prediction process. Each model, process and algorithms used in system is described below in detail.

CANDIDATE GENERATION MODEL

In this model, segmentation candidates are generated for each sentence. It mainly deals with Beam search approach and Phrase dictionary. Phrase dictionary is obtained or induced from massive corpora. Phrases are identified based on the occurrence frequency of unigrams and bigrams as given in following equation. These phrases are then stored in phrase dictionary.

$$score(w_i, w_j) = \frac{freq(w_i, w_j)}{freq(w_i, w_j)}$$

After obtaining phrase dictionary, Beam Search Approach is used for obtaining segmentation candidates. While applying Beam Search approach, every phrase is taken from phrase dictionary or cross checked with phrase dictionary. These segmentation candidates are then passed to Segmentation Ranking model [1].

SEGMENTATION RANKING MODEL

In this model ranking score is calculated for segmentation candidates. Only segmentation candidates having not as a prefix are accepted in this model. (E.g. this is not good or this is not bad, from these segments only not good or not bad is given as input to this model). In this model positive and negative dictionary is used. This dictionary contains positive and negative words and is stored in database. Second word from Segmentation candidates (E.g. good or bad) are checked in positive and negative dictionary and ranking score is assigned to this candidates. If word is present in positive dictionary the score 1 is assigned and if word is present in negative dictionary then score 0 is assigned and candidates with 0 or 1 score are stored in database.

$$\phi_{ij} = \sum s f e_{ijk}$$

Above equation is used to obtain the k -th segmentation feature. These features are nothing but unigrams and bigrams. Also it contains positivity and negativity of these n-grams [1].

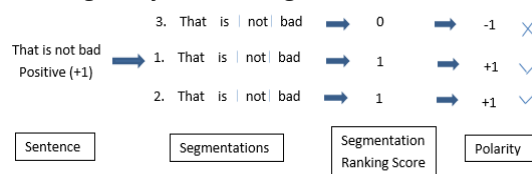


Figure 3 Segmentation Ranking Model

SENTIMENT CLASSIFICATION MODEL

Sentiment Classification model is used to obtain sentiment polarity. It mainly deals with testing process. User will enter input and then segmentation will be done first. After getting segmentation candidates, first these candidates are checked in database if are present or not. If present then final score will be taken from database and polarity is decided based on that. If ranked candidates are not present in database, the Sentiwordnet is used to obtain score and polarity is decided. Positive and negative score for candidates are taken from Sentiwordnet and which score is greater is assigned or selected as final polarity (If positive score > negative score) polarity positive and vice versa. If sentence contains both positive and negative polarity then it is assigned neutral class.

EMOTION RECOGNITION OR MOOD ANALYSIS

In mood analysis three moods are analysed mainly:

- 1) Happy
- 2) Sad
- 3) Afraid

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

In this process of analysing mood, input sentence is given and it is pre-processed. Then it is passed to sentiwordnet to get positive or negative score. According to obtained score, greater score (positive score > negative score or negative score > positive score) word is passed to Naive Bayes. Then in Bayes it is checked whether word is present in happy, sad or afraid class and mood class is returned as output. Three classes are used in Naive Bayes mainly i.e. Happy, Sad, Afraid. If positive score from Sentiwordnet is greater and Bayes result is happy then emotion of sentence is happy. If negative score is greater and Bayes result is sad then emotion of sentence is sad. If negative score is greater and Bayes result is afraid then mood is afraid.

PREDICTION PROCESS

In the prediction process, the features of top-ranked segmentation candidates is used to predict the polarity of sentence. The prediction process is formalized in following Fig.4. Given a sentence whose polarity to be predicted. For each sentence, segmentation candidate is generated based on CG. Then based on SEG ranking score is assigned to each segment. Afterwards, top-ranked K segmentation results are selected and polarity is predicted for that sentence using sentiwordnet score and training dataset. In following figure (CG-Candidate Generation Model, SC-Sentiment Classification Model and SEG-Segmentation Ranking Model).[1]

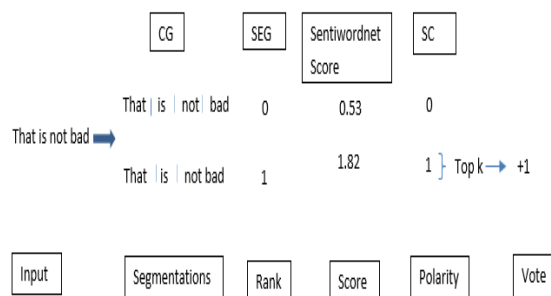


Figure 4 The Prediction Process

Algorithm 1: The prediction algorithm of the joint framework for sentence –level sentiment classification.

Input:

Test data: $T' = [s_i], 1 \leq i \leq |T'|$

Segmentation feature extractor: $sfe(.)$

Candidate generation model: CG

Segmentation ranking model: SEG

Output:

Test data with predicted polarity $T' = [s_i, pol_i], 1 \leq i \leq |T'|$

- for $i \leftarrow 1 \rightarrow |T'|$
- Generate segmentation candidate Ω_i for each sentence s_i in T' based on CG
- Calculate the segmentation scores for Ω_i based on SEG and $sfe(\Omega_i)$
- Select the top-ranked K segmentation candidate $\bar{\Omega}_i$ from Ω_i from training dataset
- For $j \leftarrow 1 \rightarrow |\bar{\Omega}_i|$ do
- Predict the sentiment polarity pol_{ij} for $\bar{\Omega}_{ij}$ Based on Sentiwordnet score and training dataset.
- End for
- $\widehat{pol}_i \leftarrow \text{majority}(pol_{ij}), 1 \leq j \leq K$
- End for
- To calculate mood generate segments of sentence.
- Pre-process data and pass to next step.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

12. Get mood score from sentiwordnet.i.e. positive score and negative score and match it according to that class (happy, sad, and afraid).
13. Display mood to user.

TRAINING PROCESS

In the training process, the ranking model is optimized for obtaining a better classification performance. The training process is formalized in following Fig.5. For each sentence s_i segmentation candidates are generated. Unigrams and bigrams are obtained from dataset and segmentation score is calculated based on SEG and sfe . The top-ranked candidates are chosen and stored in database and ranking score is obtained.. In following figure (**CG**-Candidate Generation Model, **SC**-Sentiment Classification Model and **SEG**-Segmentation Ranking Model).[1]

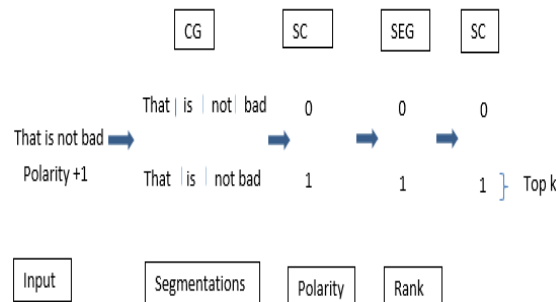


Figure 5 The Training Process

Algorithm 2 The training algorithm of the joint framework for sentence-level sentiment classification

Input:

Training data: $T = [s_i], 1 \leq i \leq |T|$

Segmentation feature extractor: $sfe(.)$

Candidate generation model: CG

Output:

Segmentation ranking model: SEG

1. Generate segmentation candidate Ω_i for each sentence s_i in T based on CG
2. Find unigram and bigram of datasets.
3. For $j = -1, \dots, |\Omega_{ij}|$ do
4. Calculate the segmentation score for Ω_{ij} based on $SEG^{(r)}$ and $sfe(\Omega_{ij})$
5. End for
6. Select the top ranked K segmentation candidate $\bar{\Omega}$ from Ω_i
7. End for
8. Get ranked score $1 \leq i \leq |T|$
9. End for
10. $SEG \leftarrow SEG^{(R)}$

IV. APPLICATIONS

Sentiment analysis is used mainly in various types of fields such as political, advertising and sociological. In marketing field, companies use it to advance their strategies, to understand customer's feelings towards products or brand how people react to their campaigns or product launches and why consumers don't buy some products. It is used in political field, for tracking the political view, also to detect reliability and irregularity between statements and actions at the government level; it can be used to predict election results. Sentiment analysis is also used to observe and inspect social phenomenon, for identifying of possibly hazardous situations and defining the general mood of the blogosphere. The

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

sentiment analysis then represents a main part for any subject (investors, policy makers etc.) to accomplish various types of activities such as: guess financial performance, define election outcomes etc. [4]

V.RESULT ANALYSIS

Following graph in Fig.6. Represent the no of ranked candidate for testing entered data. That is sentence entered to test. That is if sentence is hiii, today she is not happy. In previous methods it gives 6 segments to classify. In proposed joint framework it ranks the candidate according to feature and then pass it to classification. In this method, it will pass 5 segments for classification.



Figure 6 Graph representing ranked candidates.

Graph in Fig.7. Represent unigram and bigram score from whole training dataset. According to size of dataset value of unigram and bigram will change. Unigram score is find on basis of occurrence of single word (unigram) in whole document. Bigram score is find on basis occurrence of two words in whole dataset.

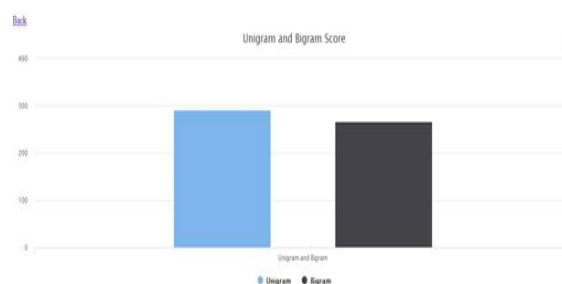


Figure 7 Unigram and Bigram Score

RESULT TABLE 1

	Pipelined framework	Joint Framework
First sentence	4	3
Second sentence	4	4
Third sentence	8	7
Fourth sentence	6	6
Third sentence	6	5

Above Result table 1 shows no. of candidates passed to classification in pipelined framework and joint framework.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

RESULT TABLE 2

	No. Of candidate
First sentence	5
Second sentence	7
Third sentence	8
Fourth sentence	7
Third sentence	3

Above Result table 2 shows Ranked candidates for sentences.

VI.CONCLUSION

A joint framework is developed based on sentence segmentation for sentence-level sentiment classification. The proposed method in this work produces segmentations results and sentence-level polarity is predicted based on the segmentation results. This joint framework deals with the trained data and test data. Final polarity is predicted using WordNet. Three parts of the framework are defined as: a candidate generation model with a constrained beam-search approach and n-gram model, a segmentation ranking model with positive and negative dictionary providing ranking score for segmentation results, and a sentiment classification model with Sentiwordnet for predicting sentiment polarity(positive, negative, neutral) and also an Emotion Recognition model for mood detection(happy, sad and afraid).

REFERENCES

- [1] D. Tang, F. Wei, B. Qin, L. Dong, T. Liu, and M. Zhou, "A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015.
- [2] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. ACL*, 2015.
- [3] D.Tang,"Sentiment Embedding's with Applications to Sentiment Analysis," IEEE Transactions on Knowledge and Data Engineering, 2015.
- [4] AlessiaD'Andrea, Fernando Ferri, PatriziaGrifoni, TizianaGuzzo,"Approaches, Tools and Applications for SentimentAnalysis Implementation", International Journal of Computer Applications, September 2015.
- [5] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proc. ACL*, 2014, pp. 1555–1565
- [6] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building large-scale twitter-specific sentiment lexicon: A representation learning approach," in *Proc. COLING*, 2014, pp. 172–182.
- [7] D. Tang, F. Wei, B. Qin, L. Dong, T. Liu, and M. Zhou, "A joint segmentation and classification framework for sentiment analysis," in *Proc. EMNLP*, 2014, pp. 477–487, Assoc. for Comput. Linguist.
- [8] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. ICML*, 2014.
- [9] SaimaAman, Stan Szpakowicz, "Identifying Expressions of Emotion in Text", 2007.
- [10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. EMNLP*, 2002, pp. 79–86.