IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 01-07

Sentiment Analysis of Google Reviews of a College

Dr.S.Gomathi @ Rohini, Punitha. R

Associate Professor/BCA, Sri Ramakrishna College of Arts and Science (autonomous), Affiliated to Bharathiar University, Coimbatore.

M. Phil. (Computer Science) Scholar, Sri Ramakrishna College of Arts and Science (autonomous), Affiliated to Bharathiar University, Coimbatore.

Abstract: *Sentiment analysis is one of the key challenges for mining online and offline user generated content. This paper focuses the reviews of college which are an important form of opinionated contents. The objective of this work is to classify every sentence's semantic orientation (e.g. positive, negative and neutral) of the reviews. This paper presents a weakly-supervised embedding model to learn the weak labels for college reviews sentiment analysis. This paper introduce the max entropy method to investigate the problem of incorporating sentiment prior knowledge to learn weak label meaningful word embeddings for sentiment analysis using R tool.*

I. INTRODUCTION

Today extensive datasets are accessible on-line, holding text data or numerical. It has been the major focus for many practitioners and researchers to apply reasonable approaches and techniques and extract useful information from those datasets. Wide range of techniques have been used and tested to retrieve information. In addition to text mining and data mining, lately interest for non-topical text analysis has increased drastically and sentiment analysis is part of them.

Sentiment analysis is a process of analyzing the given text in order to find out the emotions in it. Sentiment analysis is about "Text analysis, Information Extraction and Natural Language Processing are kind of tasks which aim towards getting the writer's feelings expressed in negative or positive comments by analyzing sentences or documents". In simple words, opinion mining is a process of detecting the sentiment of the writer concerning a particular topic. It is a blend of techniques and strategies about distinguishing and detecting subjective information from a text such as opinions and attitudes [3]. Usually, it has been about opinion polarity to find out whether someone has negative, positive or neutral opinion about college.

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards college. Since early 2000, sentiment analysis has grown to be one of the most active research areas in Natural Language Processing (NLP). It is also widely studied in data mining, Web mining, text mining, and information retrieval. In fact, it has spread from computer science to management sciences and social sciences such as marketing, finance, political science, communications, health science, and even history, due to its importance to business and society as a whole. This proliferation is due to the fact that opinions are central to almost all human activities. To beliefs and perceptions of reality, and the choices we make to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, whenever need to make a decision we often seek out the opinions of others. This is not only true for individuals but also true for organizations.

College reviews dataset shows good results by finding positive opinion of the people. **"Best college in Coimbatore, So, I'm happy"** is the opinion whose irrelevant characters are removed and shows the results as **Best college in Coimbatore, So I'm happy.** Split the reviews to find out the polarity whether it is positive, negative and neutral. Further it found he polarity and accuracy by using different methods to prove better prediction.

Existing research has produced numerous techniques for various tasks of sentiment analysis, which include both supervised and unsupervised methods. In the supervised setting, early papers used all types of supervised machine learning methods (such as Support Vector Machines (SVM), Maximum Entropy, Naïve Bayes, etc.) and feature combinations. Unsupervised methods include various methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. Several survey books and papers have been published, which cover those early methods and applications extensively. Since about a decade ago, deep learning has emerged as a powerful machine learning technique and produced state-of-the-art results in many application domains, ranging from computer vision and speech recognition to NLP. Applying deep learning to sentiment analysis has also become very popular recently. But in practice, the long-range dependencies are still problematic to handle. So this paper presents max entropy for college reviews sentiment analysis.

II. RELATED WORKS

***Collobert et.al.,* (2011)** [3] proposed an unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. This versatility is achieved by trying to avoid task-specific engineering and therefore disregarding a lot of prior knowledge. Instead of exploiting man-made input features carefully optimized for each task, our system learns internal representations on the basis of vast amounts of mostly unlabeled training data.

***L. Qu, R. Gemulla, and G. Weikum* (2012)** [4] discussed the weakly supervised Multi-Experts Model (MEM) for analyzing the semantic orientation of opinions expressed in natural language reviews. In contrast to most prior work, MEM predicts both opinion polarity and opinion strength at the level of individual sentences; such fine-grained analysis helps to understand better why users like or dislike the entity under review. A key challenge in this setting is that it is hard to obtain sentence-level training data for both polarity and strength. For this reason, MEM is weakly supervised: It starts with potentially noisy indicators obtained from coarse-grained training data (i.e., document-level ratings), a small set of diverse base predictors, and, if available, small amounts of fine-grained training data.

***R. Socher, et al.,* (2013)** [5] introduced the Recursive Neural Tensor Network. When trained on the new tree bank, this model outperforms all previous methods on several metrics. It pushes the state of the art in single sentence positive/negative classification from 80% up to 85.4%. The accuracy of predicting fine-grained sentiment labels for all phrases reaches 80.7%, an improvement of 9.7% over bag of features baselines.

***H. Lakkaraju, R. Socher, and C. Manning* (2014)** [6] focused on the problem of aspect-specific sentiment analysis. The goal here is to not only extract aspects of a product or service, but also to identify specific sentiments being expressed about them. They presented a novel approach based on a hierarchical deep learning framework which overcomes the aforementioned drawbacks. To experimented with various models of semantic compositionality within this framework.

***J. Wieting, M. Bansal, K. Gimpel, and K. Livescu* (2015)** [7] considered the problem of learning general-purpose, paraphrastic sentence embeddings based on supervision from the Paraphrase Database. Compare six compositional architectures, evaluating them on annotated textual similarity datasets drawn both from the same distribution as the training data and from a wide range of other domains. They find that the most complex architectures, such as max entropy perform best on the in-domain data. However, in out-of-domain scenarios, simple architectures such as word averaging vastly outperform LSTMs.

***D. Tang, B. Qin, X. Feng, and T. Liu* (2016)** [8] developed two target dependent Long Short - Term Memory (LSTM) models, where target information is automatically taken into account. The authors presented a neural network models to deal with target-dependent sentiment classification. The approach is an extension on LSTM by incorporating target information. Such target-dependent LSTM approach models the relatedness of a target word with its context words, and selects the relevant parts of contexts to infer the sentiment polarity towards the target. The model could be trained in an end-to-end way with standard back propagation, where the loss function is cross-entropy error of supervised sentiment classification.

***D. Tang, B. Qin, and T. Liu,* (2016)** [9] introduced a deep memory network for aspect level sentiment classification. Unlike feature-based SVM and sequential neural models such as LSTM, this approach explicitly captures the importance of each context word when inferring the sentiment polarity of an aspect. Such importance degree and text representation are calculated with multiple computational layers, each of which is a neural attention model over an external memory.

***K. Greff, et.al.,* (2017)** [10] presented the first large-scale analysis of eight LSTM variants on three representative tasks: speech recognition, handwriting recognition, and polyphonic music modeling. The hyper parameters of all LSTM variants for each task were optimized separately using random search, and their importance was assessed using the powerful ANOVA framework. In total, to summarize the results of 5400 experimental runs (≈15 years of CPU time), which makes our study the largest of its kind on LSTM networks.

***Wei Zhao, et.al,* (2018)** [11] proposed a novel deep learning framework for product review sentiment classification which employs prevalently available ratings as weak supervision signals. The framework consists of two steps: (1) learning a high level representation (an embedding space) which captures the general sentiment distribution of sentences through rating information; and (2) adding a classification layer on top of the embedding layer and use labeled sentences for supervised fine-tuning. To explored two kinds of low level network structure for modeling review sentences, namely, convolutional feature extractors and long short-term memory. To evaluated the proposed framework, to construct a dataset containing 1.1M weakly labeled review sentences and 11,754 labeled review sentences from Amazon.

III. RESEARCH METHODOLOGY

This paper proposed a weakly-supervised deep embedding process to learn the weak labels for College Reviews Sentiment Analysis. The proposed method follows a well established word embedding technique for annotation makes level of underlying linguistic representation of text. The proposed architecture diagram is described in figure 1.

Data Extraction

Data extraction is to retrieve the data from google reviews of a leading autonomous arts & science college in a city. The collected data are preprocessed and the words are converted into vectors, this data set contains many sentences labeled with 0 or 1 depending on its polarity. Tokenization refers to the procedure of splitting a set of text into meaningful words (stems), phrases. Documents hold stop words. Stop words are as well as field specific and its function words like prepositions, articles, conjunction and pronouns.

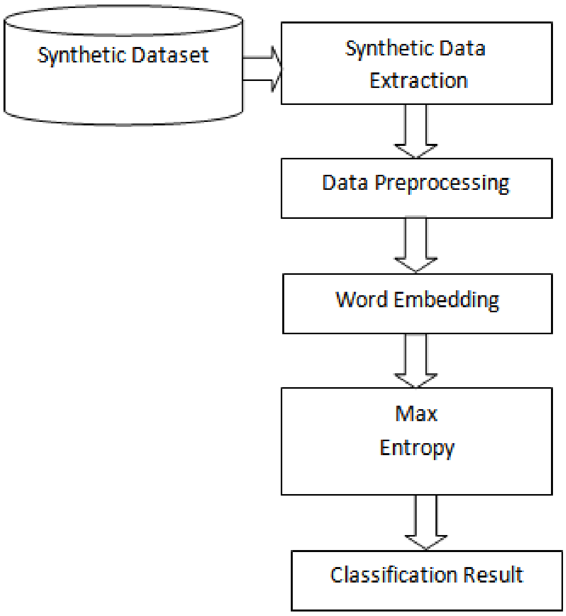

Figure 1: Architecture Diagram

IV. DATA PREPROCESSING

Data preprocessing is a process of removing the unrelated and duplicate features from a dataset in order to develop the performance of unsupervised machine learning algorithm in terms of accuracy and time to build the model. Data cleaning or preprocessing is the most important part of this paper. Preprocessing of data is the process of planning and cleaning the irrelevant reviews for clustering. Reducing the irrelevant noise in the reviews should help to progress the performance of the clustering and speed up the clustering process. This process performed the following processes on reviews during cleaning and normalization.

Word Embedding

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. Word embedding is a popular method for NLP that aims to learn low-dimensional vector representations of words from documents. Word embedding is the collective name for a set of language modeling and feature learning techniques in NLP where words are mapped to vectors. The proposed max entropy for word embedding is one of the popular techniques.

V. EXPERIMENT

Experiment Setup and Datasets. We conducted experiments on the google reviews sentiment classification. Google reviews contains 254 reviews and those reviews are splitted to find the sentiments for a particular words. Optimal performance for all models was achieved at word vector sizes. For all models, we used the cross validate for the word vector size. We also directly evaluated the effectives of the max entropy by measuring the word similarity in the embedding space for sentiment lexicons from google reviews. Max entropy would usually achieve its best performance on 4 cross validation using R tool.

Table:1 Statistics of Google Reviews Sentiment Classification Dataset

Positive	Negative	Neutral	Total
245	3	6	254

Set 1: Classifying Positive Sentences

The first set contains positive sentences, in this set negation changes the overall sentiment of a sentence. In the results based on sentiment conditional probability, the importance of a positive sentenc**e** is higher than that of a negative sentence in all positions. This indicates that a positive sentence is more powerful to distinguish the sentiment polarity of a document. Hence we compute accuracy in terms of correct sentiment for positive sentences**. "The placements are good. Faculties are really helpful. The college campus is good"**, data preprocessing has the good role to removing the irrelevant data and shows the results as **The placements are good Faculties are helpful The college Campus is good.** Above sentence are split in vectors to find out the sentiments and its polarity.

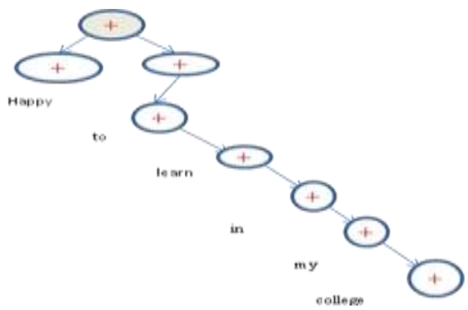

Figure 2: Classifying positive sentences

"The placements are good. Faculties are really helpful. The college campus is good."
"Happy to learn in my college."
"Good college for Students."
"Staff members were very talented"

Table 2: Sentiment Classification of Positive Sentences

Reviews	**Anger**	**Anticipation**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**	**Trust**	**Negative**	**Positive**
The placements are good Faculties are really helpful The college campus is good	0	1	0	0	2	0	1	2	0	2
Happy to learn in my college	0	1		0	0	1	0	0	0	2
Good College for Students	0	0		0	0	0	0	1	0	1
Staff members were very good talented	0	0		0	0	0	0	1	0	1

Set 2: Classifying Negative and Neutral Sentences

The second contains negative sentences, negative sentences and neutral sentences were easily predicted by ax entropy by splitting the sentences into vector to find the negative sentiments and neutral sentiments .Hence we compute accuracy in terms of correct sentiment for negative and neutral sentences. **"No support from staffs or department to student< > :>"** data preprocessing has the good role to removing the irrelevant data and shows the results as **No support from staffs or department to student**. Above sentence are split in vectors to find out the sentiments and its polarity.

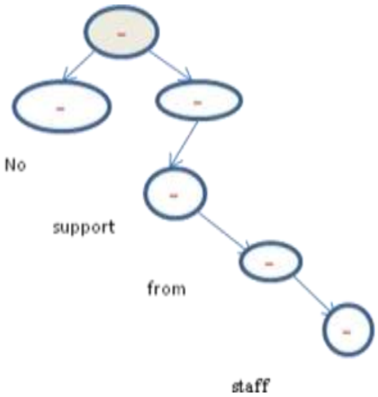

Figure: 3 Classifying negative sentences

"No support from staffs or department to student< >: >"
"Spacy Auditorium with all amenities can concentrate on toilets."
"All facilities are there excellent college."
"Top college in Coimbatore,"
"Infrastructure facilities management and membership."
"This building is going to be destroyed in this year ."

Table 3: Sentiment Classification of Positive Sentences

Reviews	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive
No support from staffs or department to student	1	0	0	0	0	0	0	0	1	0
Spacy Auditorium with all amenities Can concentrate on toilets	0	0	0	0	0	1	0	1	1	0
All facilities are there excellent college	0	0	0	0	0	0	0	0	0	0
Top college in Coimbatore	0	0	0	0	0	0	0	0	0	0
Infrastructure facilities management and membership	0	0	0	0	0	0	0	0	0	0
This building is going to be destroyed in this year	1	0	0	1	0	0	0	0	1	0

Sentiment Polarity

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation affective state, or the intended emotional communication.

Table 4: Sentiment Polarity

Sentences	Polarity
The placements are good. Faculties are really helpful. The college campus is good	Most positive
Happy to learn in my college	Positive
Good College for Students	Positive
Staff members were very talented	Positive
No support from staffs or department to student	Negative
Spacy auditorium with all amenities Can concentrate on toilets.	Negative
This building is going to be destroyed in this year	Most Negative
All facilities are there excellent college	Neutral
Infrastructure facilities management and membership	Neutral

Model Analysis: Most Positive and Most Negative

We queried the model for its predictions on what the most positive and negative are measured as the highest activation of the most positive and most negative classes. Table 4 shows some sentences which the Maximum Entropy selected their strongest sentiment. Most positive holds highest values when compared to other values and the most negative holds the least value when compared to other values.

Table 5:Predicting Sentiment Values

Sentences	Values
The placements are good. Faculties are really helpful. The college campus is good	4.0
Happy to learn in my college	3.0
Good College for Students	2.1
Staff members were very talented	3.7
No support from staffs or department to student	0.1
Spacy auditorium with all amenities Can concentrate on toilets.	1
This building is going to be destroyed in this year	-0.1
All facilities are there excellent college	Neutral
Infrastructure facilities management and membership	Neutral

Comparative study

We compared our method with the following
sentiment classification algorithms:

1. Naive Bayes: We use the 254 google reviews selected by positive, negative and neutral emoticons, it holds less training data's and build sentiment classifier (Wang and Manning). Need less training data but not for probabilistic prediction.

2. Support Vector Machine (SVM): The SVM are widely used baseline methods to build sentiment classifiers (pang et al.,) Lib linear is used to train the SVM classifier. SVM is better at computation speed and memory but worse at performance compared to deep learning

3. Random Forest: Random Forest (Socher et al.,) has been proven effective in many sentiment tasks and runs efficiently in larger databases but over fitting can easily occurs.

4. Maximum Entropy: The model uses search-based optimization to find weights for the features maximize the likelihood of the training data. It shows a good results when comparing to other methods such as RF (Random Forest), SVM,Tree, Naïve bayes.

Performance Comparison

The results are shown in Table 2. We also report performance for reviews. The key observations are as follows. Support vector machine poorly performs when compared to Max entropy, since factual statements would not contain opinion words. When no opinion word is detected, we can only make random predictions. We compared commonly used methods that use bag of words features with Naïve Bayes, SVM's, Random Forest and Tree.We also analyse performance on positive , negative and neutral.

Table 6: Accuracy for Sentiment Classification at the Sentence Level

Method	Accuracy
Naïve bayes	80%
Maxent	96%
Tree	71.6%
SVM	86%
RF	88%

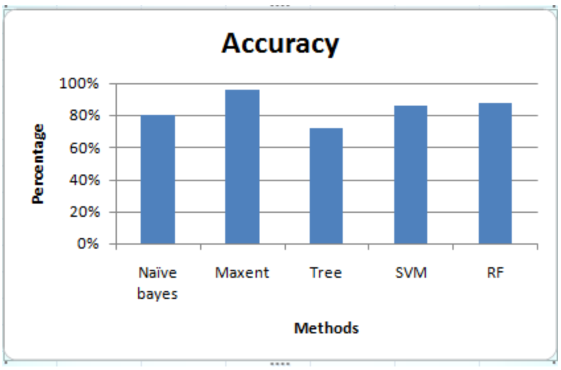

Figure-4 Accuracy of the Google Reviews

VI. CONCLUSION

Sentiment Analysis has lead to development of better products and good business management, reviews. In this work we proposed Max entropy for weakly-supervised Deep Embedding for college review sentence sentiment classification. The method max entropy performs well for finding the results as positive. The accuracy of Max entropy produces 96% for college reviews. Max entropy is more capable of modeling the long-term dependencies in sentences, but it is high efficient than Random forest, SVM, Naïve Bayes, Tree. We also find that our method can improve general text classification .For future work will focus on improving our embeddings and sentiment classification by effectively handling undertrained words as well as by exploring new models that generalize even better by using the large number of possible reviews.

REFERENCES

[1]. M. Mäntylä, D. Graziotin and M. Kuutila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers," Oulu, Finland.
[2]. B. Liu, "Sentiment Analysis and Subjectivity," University of Illinois at Chicago, Chicago, 2010.
[3]. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.
[4]. L. Qu, R. Gemulla, and G. Weikum, "A weakly supervised model for sentence-level semantic orientation analysis with multiple experts," in Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn., pp. 149–159, 2012.
[5]. R. Socher, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proc. Conf. Empirical Methods Natural Language Process, vol. 1631, Art. no. 1642, 2013.
[6]. H. Lakkaraju, R. Socher, and C. Manning, "Aspect specific sentiment analysis using hierarchical deep learning," in Proc. Neural Inf. Process. Syst. Workshop Deep Learn. Representation Learn., 2014.
[7]. J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," arXiv:1511.08198, 2015.
[8]. D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for targetdependent sentiment classification," in Proc. 26th Int. Conf. Comput. Linguistics, pp. 3298–3307, 2016.
[9]. D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in Proc. Conf. Empirical Methods Natural Language Process., pp. 214–224, 2016.
[10]. K. Greff, R. K. Srivastava, J. Koutnık, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2222–2232, 2017.
[11]. Wei Zhao, et.al., "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis", IEEE Transactions on Knowledge and Data Engineering, Vol. 30, No. 1, 185, 2018.