International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

SENTIMENT ANALYSIS BASED ON USER REVIEWS IN E-COMMERCE SITE

G. Rajadurai
M. Tech (Computer Science and Engineering)
Mail id: berajadurai@gmail.com
Mrs. D. Vinotha, M.S.,(US)
Asst. Professor
PRIST University, Thanjavur.

Abstract - Learning sentiment-specific word embeddings dubbed sentiment embeddings is proposed in this paper. Existing word embedding learning algorithms typically only use the contexts of words but ignore the sentiment of texts. It is problematic for sentiment analysis because the words with similar contexts but opposite sentiment polarity, such as good and bad, are mapped to neighboring word vectors. This issue is addressed by encoding sentiment information of texts (e.g. sentences and words) together with contexts of words in sentiment embeddings. By combining context and sentiment level evidences, the nearest neighbors in sentiment embedding space are semantically similar and it favors words with the same sentiment polarity. In order to learn sentiment embeddings effectively, a number of neural networks with tailoring loss functions, and collect massive texts automatically with sentiment signals like emoticons as the training data is developed. Sentiment embeddings can be naturally used as word features for a variety of sentiment analysis tasks without feature engineering. Sentiment embeddings is applied to word-level sentiment analysis, sentence level sentiment classification and building
sentiment lexicons. Experimental results show that sentiment embeddings consistently
outperform context-based embeddings on several benchmark datasets of these tasks.

Keywords: Text mining, sentiment analysis, sentiment classification, bag of words, feature based sentiment.

1. INTRODUCTION

Google defines sentiment analysis as “the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral”. In other words, sentiment analysis is a line of research that harnesses people’s opinions and attitude in relation to different topics, products, events and attributes. It is an extension of data mining in the nlp (natural language processing) domain. This concept involves studying each opinionated word or phrase in the text and labelling it as positive, negative or neutral. Today, social media has become an integral part of our lives, where people express

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

their thoughts and opinions about various things. websites such as twitter, facebook, instagram, tumblr, and myspace have become extremely popular. recently, on 24th august, 2015, facebook made a record by counting one billion people accessing the website in a single day. That is, 1 in 7 people on earth were connected at the same time. In comparison, the microblogging site, twitter has a user base of 316 million active users. The content accumulation of so many opinions, comments are useful if extracted and analyzed in the correct way.

2. LITERATURE SURVEY

Title 1: A Neural Probabilistic Language Model

A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. This is intrinsically difficult because of the curse of dimensionality: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training. Traditional but very successful approaches based on n-grams obtain generalization by concatenating very short overlapping sequences seen in the training set. We propose to fight the curse of dimensionality by learning a distributed representation for words which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences.

Title 2: Distributed Representations of Words and Phrases and their Compositionality

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling. An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases.

Title 3: GloVe: Global Vectors for Word Representation

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arithmetic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global logbilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word cooccurrence

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task.

Title 4: Word Alignment Modeling with Context Dependent Deep Neural Network

We explore a novel bilingual word alignment approach based on DNN (Deep Neural Network), which has been proven to be very effective in various machine learning tasks (Collobert et al., 2011). We describe in detail how we adapt and extend the CD-DNNHMM (Dahl et al., 2012) method introduced in speech recognition to the HMMbased word alignment model, in which bilingual word embedding is discriminatively learnt to capture lexical translation information, and surrounding words are leveraged to model context information in bilingual sentences. While being capable to model the rich bilingual correspondence, our method generates a very compact model with much fewer parameters.

Title 5: Parsing with Compositional Vector Grammars

Natural language parsing has typically been done with small sets of discrete categories such as NP and VP, but this representation does not capture the full syntactic nor semantic richness of linguistic phrases, and attempts to improve on this by lexicalizing phrases or splitting categories only partly address the problem at the cost of huge feature spaces and sparseness. Instead, we introduce a Compositional Vector Grammar (CVG), which combines PCFGs with a syntactically untied recursive neural network that learns syntactico-semantic, compositional vector representations. The CVG improves the PCFG of the Stanford Parser by 3.8% to obtain an F1 score of 90.4%. It is fast to train and implemented approximately as an efficient reranker it is about 20% faster than the current Stanford factored parser.

3. PROBLEM STATEMENT:

In sentiment analysis, the main fields of research are sentiment classification, opinion summarization and feature based sentiment classification. Sentiment classification is used to classify the whole document according to the opinion around certain objects. In opinion summarization, the features of the product are mined on which the customers have expressed their opinions. Feature based sentiment classification considers the opinion on features of certain objects. For instance, user-generated reviews of products not only assess the overall product but also reveal some sentiments on product specific aspects, such as performance, price etc

4. Existing System

Word representation aims to represent aspects of word meaning. A straight-forward way is to encode a word w as a one-hot vector, whose length is vocabulary size with 1 in the w_i^{th} position and zeros everywhere else. However, such one-hot word representation only encodes the indices of words in a vocabulary, without capturing rich

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

relational structure of the lexicon. One common approach to discover the similarities between words is to learn a clustering of words .Each word is associated with a discrete class, and words in the same class are similar in some respects. This leads to a one-hot representation over a smaller vocabulary size. Instead of characterizing the similarity with a discrete variable based on clustering results which corresponds to a soft or hard partition of the set of words, many researchers target at learning a continuous and real-valued vector for each word, also known as word embeddings.

Drawbacks of Existing System

- Existing embedding learning algorithms are mostly based on the distributional hypothesis, which states that words in similar contexts have similar meanings

Proposed System

In this paper, learning sentiment-specific word embeddings dubbed sentiment embeddings for sentiment analysis is proposed. The effectiveness of word contexts and exploit sentiment of texts for learning more powerful continuous word representations are retained. By capturing both context and sentiment level evidences, the nearest neighbors in the embedding space are not only semantically similar but also favor to have the same sentiment polarity, so that it is able to separate good and bad to opposite ends of the spectrum. In order to learn sentiment embeddings effectively, a number of neural networks to capture sentiment of texts (e.g. sentences and words) as well as contexts of words with dedicated loss functions is developed.

Advantages of Proposed System

- Sentiment embeddings are useful for discovering similarities between sentiment words
- On sentence level sentiment classification, sentiment embeddings are helpful in capturing discriminative features for predicting the sentiment of sentences
- On lexical level task like building sentiment lexicon, sentiment embeddings are shown to be useful for measuring the similarities between words

5. IMPLEMENTATION

Training Data

Sentence level sentiment information are collected automatically from Mobile Apps. This is based on the consideration that larger training data usually leads to more powerful word representation, and it is not practical to manually label sentiment polarity for huge number of sentences. In order to collect sentiment information of words, the word clusters from Urban Dictionary to expand a small size of manually labeled sentiment seeds. Specifically, label the top frequent 500 words from the vocabulary of sentiment embedding as positive, negative or neutral.

Word Representation

In this module a notation is defined for the meaning of variables used. In particular, w_i means

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

a word whose index is i in a sentence, h_i is context words of w_i in one sentence, e_i is the embedding vector of w_i.

Modeling Context of Words

In this module a prediction model and a ranking model is described to encode contexts of words for learning word embeddings. These context-based models will be naturally incorporated with sentiment-specific models for learning sentiment embeddings.

Ranking Model

The scoring function is achieved with a feed forward neural network. Its input is the concatenation of the current word w_i and context words h_i , and the output is a linear layer with only one node which stands for the compatibility between w_i and h_i. During training, an artificial noise w^n is randomly selected over the vocabulary under a uniform distribution.

Modeling Sentiment Polarity of Sentences

This module presents the approaches to encode sentiment polarity of sentences in sentiment embeddings in this part.

Ranking Model

Describe an alternative of prediction model, which is a ranking model that outputs two real valued sentiment scores for a word sequence with fixed window size. The basic idea of ranking model is that if the gold sentiment polarity of a word sequence is positive, the predicted positive score should be higher than the negative score. Similarly, if the gold sentiment polarity of a word sequence is negative, its positive score should be smaller than the negative score.

Modeling Sentiment of Sentences and Contexts of Words

This module introduce two hybrid models, which naturally capture both sentiment of sentences and contexts of words for learning a more powerful sentiment embeddings based on aforementioned models.

Hybrid Prediction Model

These modules combine the context-based prediction model and the sentiment-based prediction model to get a hybrid prediction model in this part.

Hybrid Ranking Model

Similar with the hybrid prediction model, context ranking model and sentiment ranking model are merged to get a hybrid ranking model in this part.

Algorithm Used

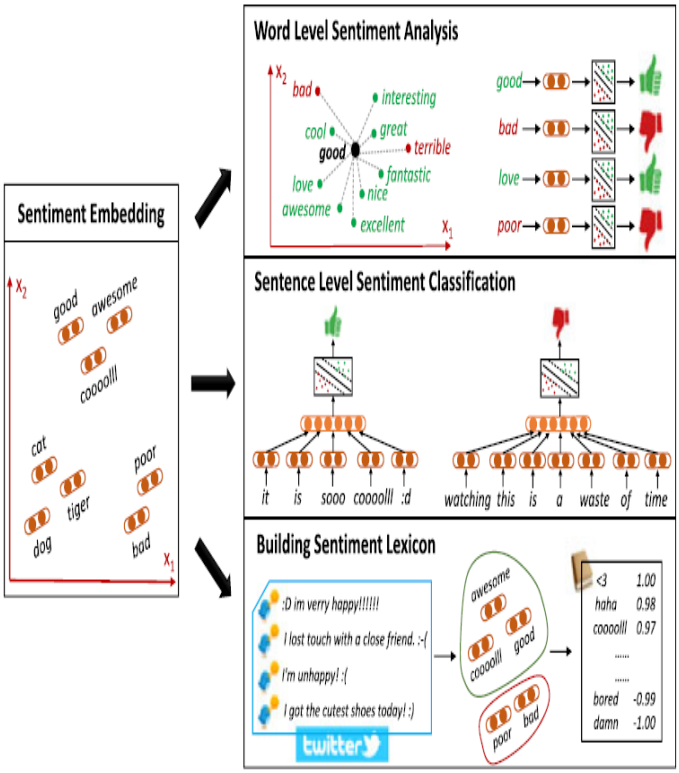
Introduce two hybrid models, which naturally capture both sentiment of sentences and contexts of words for learning a more powerful sentiment embeddings.

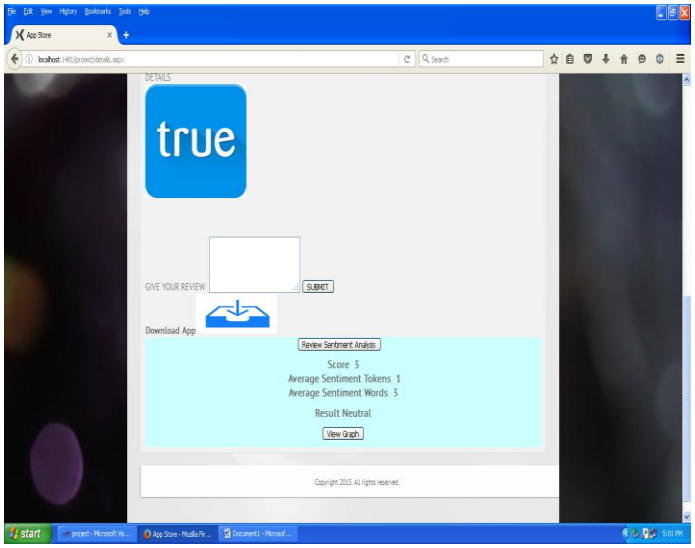
SE-HyPred –Hybrid Prediction Model

SE-HyRank – Hybrid Ranking Model

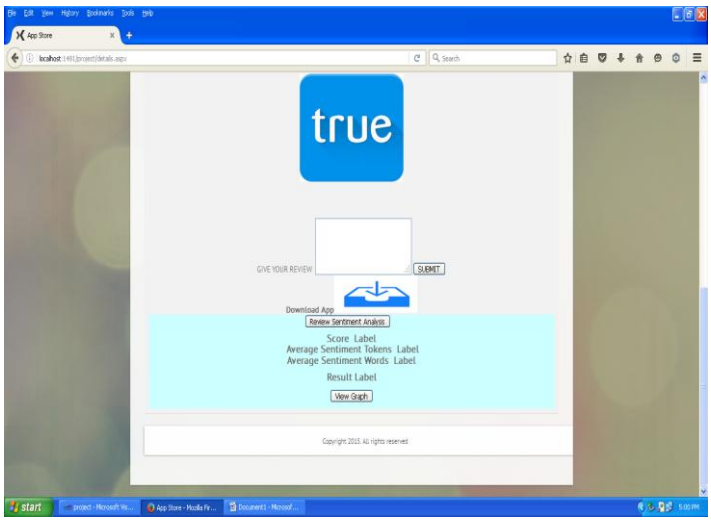
International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

7. SYSTEM MODEL





Screenshots:



8. CONCLUSION AND FUTURE WORK

This paper depicts the available methods for carrying out sentiment analysis of reviews and have showcased the methods which the survey has shown to be the most efficient. Instead of using purely supervised or unsupervised learning algorithms advanced by heuristic approaches like ANN to increase accuracy, a semi-supervised approach is proposed in which emphasis is given to meaningful opinion words identified using WordNet. Sentiment analysis will be carried out at sentence level using NLTK with Naïve Bayes probabilistic model. Representation of results will be done graphically and statistically.

9. REFERENCES:

- D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 28, 2017 ISSN: 2348-6600
Reference ID: IJCS-293 PAGE NO: 1934-1940

Proc. 8th Int. Workshop Semantic Eval., 2014, pp. 208–212.

- C. D. Manning and H. Sch€utze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.
- D. Jurafsky and H. James, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.
- Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learning Res., vol. 3, pp. 1137–1155, 2003.
- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proc. Conf. Neural Inf. Process. Syst., 2013, pp. 3111–3119.
- J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.