Sentiment Classification based on Latent Dirichlet Allocation

Raja Mohana S.P
Assistant Professor,
Department of IT,
PSG College of Technology

Umamaheswari K, Ph.D
Professor,
Department of IT,
PSG College of Technology

Karthiga R
PG Scholar,
Department of IT,
PSG College of Technology

ABSTRACT
Opinion miningrefers to the use of natural language processing, text analysis and computational linguistics to identify and extract the subjective information. Opinion Mining has become an indispensible part of online reviews which is in the present scenario.In the field of information retrieval, a various kinds of probabilistic topic modeling techniques have been used to analyze contents present in a document. A topic model is a generative technique for document. All topic models share the idea that documents are having mixture of topics, and the topic is a probability distribution over words. Recently topic modeling techniques have been used to identify the meaningful review aspects, but existing topic models like Latent Dirichlet Markov Allocation (LDMA), hierarchical aspect sentiment model (HASM) do not identify aspect specific opinion words and also not suitable for shared features. In the proposed system, movie review dataset is collected from the IMDB database and is preprocessed. TF-IDF is calculated for the preprocessed data and result is given to LDA model which is then used to discover both the aspects and aspect specific opinion words. After that CHI value has been determined, SVM classifier is used to classify the topics preferable to each and every document.

Keywords
Latent DirichletAllocation, Support Vector Machine, TF-IDF, Chi value.

1.INTRODUCTION
With the explosion of Web 2.0, various types of social media such as blogs, discussion forums and peer-to-peer networks present a wealth of information that can be very helpful in assessing the general public's sentiment and opinions towards products and services. Recent surveys have revealed that opinion-rich resources like online reviews are having greater economic impact on both consumers and companies compared to the traditional media [1]. Opinion mining deals with the analysis of opinion, sentiment, and subjectivity in text.It is a personal view or thought of the user about a particular issue or product. The web is providing an opportunity for the users to express their views, opinions or comments about all kinds of topics. The web opinion acts as an interface between the Internet users and web. It allows Internet users to communicate and express their opinions through blogs, social networks and forums.

Topic modeling techniques provide a simple way to analyze large volumes of unsupervised data. A topic consists of a group of words that frequently occur together.Topic models can connect words that have same meanings and difference between uses of words with multiple meanings. Topic modeling programs assume that any piece of text is composed by selecting words from possible baskets of words, where baskets represent topics. If that is true, then it is possible to decompose a text into the probable baskets from whence the words first came.

2. LITERATURE SURVEY
Ying Fu, Meng Yan, Xiaohong Zhang [1], presented a novel automatic change message classification method characterized by semi-supervised topic semantic analysis to automatically classify change messages. The original LDA only provided the probability of topics within the document and the probability of words within that topic. In semi-supervised LDA based method, the classification algorithm relies on the similarity comparison between the unclassified documents and the signifier documents. The advantage of using this algorithm is that it improves the accuracy and performance of the system. The drawback is that incompleteness, quality dependant, no precise stemming.

AyoubBagheri,MohamadSaraee [2], proposed the LDMA model by combining both the Latent Dirichlet Allocation and Hidden Markov Model to emphasize on extracting multi-word topics from text data. It is done by making use of both the LDA and Hidden Markov model. LDMA is a four-level hierarchical Bayesian model where topics are associated with documents, words are associated with topics and topics in the model can be presented with single- or multi-word terms by to relax the "bag of words" assumption from LDA to yield to a better model in terms of extracting latent topics from text. The advantage of LDMA is that it has the power to decide whether to generate a unigram, a bigram, a trigram, etc. The drawbacks associated with this system is that it does not identify aspect specific opinions.

Wayne Xin Zhao, Jing Jang introducedaJointly Modeling Aspects and Opinion with a MaxEnt LDA Hybrid [3]. The aim of Jointly Modeling Aspects and Opinion with a MaxEnt LDA Hybrid system is to model the jointly discover both aspects and aspect-specific opinion words with a relatively small amount of training data set. This is done by assuming there are aspects in a given collection of reviews from the same domain, and each review document contains a mixture of aspects and each sentence is assigned to a single aspect, which is often true based on observation. The author showed that by incorporating a supervised, discriminative maximum entropy model into an unsupervised, generative topic model could leverage syntactic features to help separate aspect and opinion words. Also it evaluated the model on two large review datasets from the restaurant and the hotel domains.

3. PROPOSED METHODOLOGY

3.1.Preprocessing

All the irrelevant information in the review files are removed which are not necessary for analysis. Data preprocessing include cleansing of data by removing extra useless information. It includes tokenization, stop words removal and stemming.

3.1.1. Tokenization

Text document has a collection of sentences.Split up the sentences into terms or tokens by removing white spaces, commas and other symbols.

3.1.2. Stop words removal

Stop words are words which are filtered out prior to, or after, processing of natural language data. For example, words likea, an, the, and,or, before,but, while and soon which do not contribute to classify the reviews.These words are removed from the dataset so as to avoid using them as features.

3.1.3. Stemming

Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form.

3.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved.In LDA, each document can be visualized as a mixture of various topics.

Fig. 1 represents the plate notation of LDA model. M denotes the number of documents, N the number of words in a document.

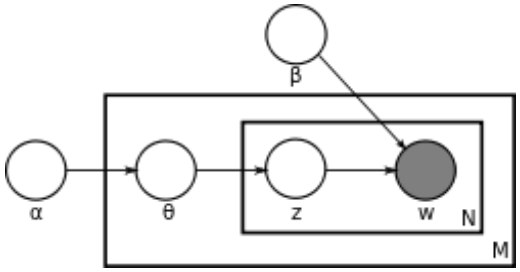

Fig 1: Plate notation of LDA

α is the Dirichletparameter on the per-document topic distributions,

β is the Dirichletparameter on the per-topic word distribution,

θ_i is the topic distribution for document i,

ϕ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.

LDA takes the output of feature extraction as input to it. It predicts the topics from each and every document. It is a mixture of topics that emit words with certain probabilities.

ALGORITHM

For each and every document d go through each word w in d and for each topic t, compute:

For each document w in a corpus D:

- Choose N (words) ~ Poisson(ξ)
- Choose θ (topic distribution) ~Dir(α)
- For each of the N words w_n
- Choose a topic z_n~ Multinomial(θ)
- Choose a word w_n from $p(w_n |z_n,\beta)$, a probability is conditioned on the topic z_n

3.3.Support Vector Machine

Support Vector Machines have been applied to classify the opinions. In SVM two sets namely training set and tests sets are used. In the field of text categorization SVM gets good results. It is based on the concept of decision planes.In this technique it finds an optimal hyperplane to separate two classes. SVM performed best when using binary features. It is more robust that can effectively deal with all kinds of noises and errors involved in the text classification.

3.4.Performance Evaluation

On implementing sentiment classification system using topic modeling (LDA), performance of both LDA and SVM classifier were analyzed. Performance analysis was made based on the evaluation measures such as precision and recall.

Evaluation Measure:

*Precision:*It is the number of relevant documents identified; precision is given in eqn (3.1)

$$\frac{\text{Number of true positives}}{\text{Number of true positives} + \text{False positives}} \ldots\ldots\ldots (3.1)$$

Recall: It is the percentage of all topics that are correctly classified. It evaluates the performance of the particular system which is used in classification. Recall is given in equ (3.2),

$$\frac{\text{Number of true positives}}{\text{Number of true positives} + \text{False negatives}} \ldots\ldots\ldots\ldots (3.2)$$

*F-Measure: P*recision p and Recall r were considered to compute the score which is expressed in eqn (3.3)

$$\text{FMeasure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \ldots\ldots\ldots\ldots\ldots\ldots (3.3)$$

4. SYSTEM DESIGN

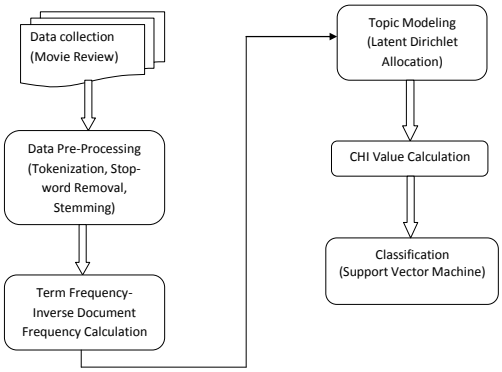

Fig 2:Framework of data processing

5. RESULT AND ANALYSIS

The movie review dataset is collected from the website ai.stanford.edu. It includes the IMDB database which is a large database with relevant and comprehensive information on movies- past, present and future. Dataset consists of 50k unsupervised reviews.

The Data is pre-processed to remove the stop words and to generalize the words for removing the null value and noisy data present in the actual dataset.The information is then extracted by calculating TF-IDFcan be obtained once the document has been given as input for the particular one.

The tf-idf matrix has to be generated which has taken output of both term and inverse document frequency. Both the values have to be multiplied and then the output will be given as in vector form.(i.e) the values will be given as input to the topic modeling techniques.

Table 1. Topics from documents

doc_no	topics_lda
1	video-2 horror-24
2	video-1 mild-23 funny-17
3	actor-71 movie-38
4	love-84 horror-42
5	movie-86

Based on the topics assigned to the documents, chi value will be calculated to make the classifier to classify the topics according to the terms.

Table 2. Topics Classified for terms

doc_no	topics_lda	class_name
1	video-2 horror-24	movie_function
2	video-1 mild-23funny-17	movie_function
3	actor-71 movie-38	movie_function
4	love-84 horror-42	movie_type
5	movie-86	movie_function

SVM classifier takes the topics assigned to the documents with their chi values as input and classifies the terms associated with each topic based on the chi value calculated.Fig 3 shows the comparison of LDA and SVM:

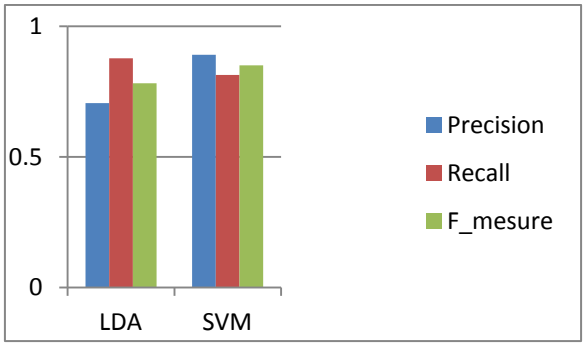

Fig 3: Performance Comparison of LDA and SVM

6. CONCLUSION

Explosion of digital content to manage on the internet requires new tools for automatically mining, searching, indexing and browsing large collections of text data. This system provides the mining results of different movies based on data mining and natural language processing methods. The proposed Topic modeling algorithm identified the topics associated with the documents which is given to the classifier and generates the appropriate topics for the documents. This model achieved either better or comparable performance compared to existing semi-supervised approaches despite using no labeled documents, which demonstrates the flexibility of LDA in the sentiment classification task. The results provided by the system are easily understandable by the user and helpful for the user in decision making process. Moreover, the topics and topic sentiments detected by LDA are indeed coherent and informative. In the future, it can be further extended to discover a set of topics with shared features in a hierarchical structure and more scalable extraction and classification techniques can be used to improve the performance.

7. REFERENCES

[1] Ying Fu, Meng Yan, Xiaohong Zhang, "Automated classification of software change messages by semi-supervised Latent Dirichlet Allocation", Information and Software Technology. A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park, Mar, 2014

[2] AyoubBagheri,MohamadSaraee, "LatentDirichlet Markov Allocation for Sentiment Analysis", Information Technology and Quantitative Management, ITQM Modeling and Simulation Design. AK Peters Ltd, Jun, 2013

[3] Wayne Xin Zhao, Jing Jang, "Jointly Modeling Aspects and Opinions with a Max-Ent LDA Hybrid", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages.1289-1305, Oct, 2010

[4] Jiguang Liang, Ping Liu, "Sentiment Classification Based on AS-LDA Model", Information Technology and Quantitative Management, Journal of Systems and Software, Feb, 2005

[5] Suin Kim, Jianwen Zhang, "A Hierarchical Aspect-Sentiment Model for Online Reviews", Association for the Advancement of Artificial Intelligence, Oct, 2013

[6] RavendraRatan Singh Jandail, "A proposed Novel Approach for Sentiment Analysis and Opinion Mining", International Journal of UbiComp (IJU), Vol.5, Apr, 2014

[7] RichaSharma,Shweta Nigam and Rekha Jain, "Mining Of Product Reviews At Aspect Level", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, May, 2014

[8] Tan C, Lee L, Tang J, et al., "User-level sentiment analysis incorporating social networks", Proceedings of SIGKDD, pages.1397-1405, Sep, 2011

[9] Oh J H, Torisawa K, Hashimoto C, et al. "Why question answering using sentiment analysis and word classes". Proceedings of EMNLPCNLL, pages.368-378, Apr, 2012

[10] ArtiBuche, Dr. M. B. Chandak, AkshayZadgaonkar, "Opinion Mining and Analysis: A Survey", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013