IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 44-49

Affect Analysis Of User Review To Determine The Topic Using NLP

S.Sherine Monisha[1], Mr B.Anbarasu, M.E.,[2]
[1](CSE, Sri Venkateswara College of Engineering and Technology, India)
[2](Assistant Professor CSE, Sri Venkateswara College of Engineering and Technology, India)

Abstract: *The rapid increase in the number of e-commerce has given rise to a significant amount of user generated text, which contains good information about consumer preferences and opinion. However, supervised topic modeling remains a challenging problem. First, most supervised topic models require prespecifying the number of topics a priori . Such specification may result in model misspecification when the specified numbers of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features.*

I. Introduction

The proliferation of e-commerce has given rise to a significant amount of user-generated text, which contains salient information about consumer preferences and opinions. Topic models are a major family of text analysis techniques for exploring the underlying semantic themes (i.e., topics) within textual data. However, prior research necessitates not only understanding the semantic themes but also integrating predictive analytics on variables of interest, such as customer sentiment product quality, affect, and more.

Standard topics models (e.g., LDA) are unsupervised and therefore incapable of making such predictions. To this end, the supervised topic modeling techniques have emerged, which can simultaneously discover the underlying semantic themes and leverage these themes for prediction. Both the discovered themes and the predicted response variables provide valuable insights about consumer preferences and opinions. Supervised topic models have a number of important e commerce applications, including customer feedback assessment, online review evaluation, consumer sentiment analysis, product attributes mining, and customer preferences identification.

However, supervised topic modeling remains a challenging problem. First, most supervised topic models require prespecifying the number of topics a priori . Such specification may result in model misspecification when the specified number of topics misrepresent the true underlying topic structure.

For example, customer reviews for new products may contain unseen topics about new features. Prespecifying the number of topics inhibits the incorporation of such unseen topics, leading to unreliable topics and inaccurate predictions. Second, existing supervised topic models treat the proportion of topic mixtures as a reduced dimension representation of the original document and make predictions based on such representations.

It is unclear whether these representations contain sufficient predictive information about the response. Statistically speaking, sufficiency entails that the reduced dimension representation preserves all the information from original documents for making predictions. The missing information in the supervised topic modeling process may diminish the prediction accuracy. Third, large text corpora often span several million documents, leaving many supervised topic models unscalable. Most supervised topic models adopt sampling-based inference algorithms, which require hundreds of iterations over each variable across all documents before convergence. Therefore, the scalability of these models is limited.

Proposed a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR).

1.1 Overview

The rapid increase in the number of e-commerce has given rise to a significant amount of user generated text, which contains good information about consumer preferences and opinion. However, supervised topic modeling remains a challenging problem. First, most supervised topic models require prespecifying the number of topics a priori. Such specification may result in model misspecification when the specified numbers of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features.

II. System Architecture

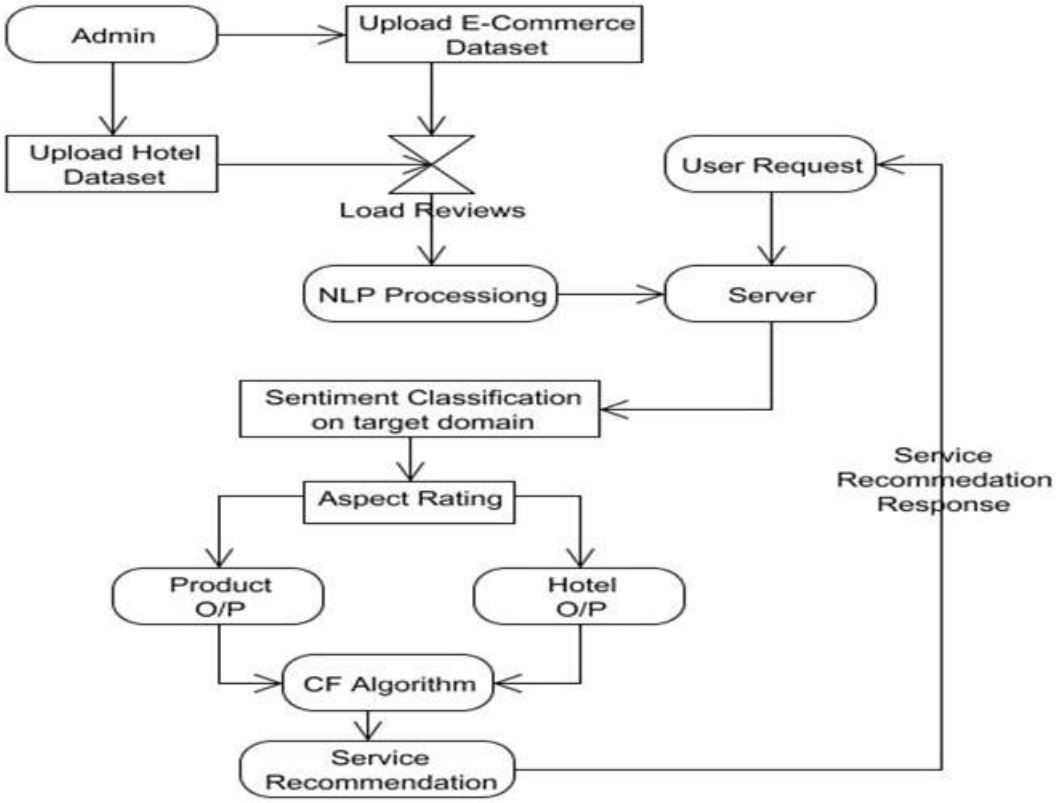


Fig 2.1 Architecture diagram

2.2 Proposed System

We propose a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). Specifically, the Hierarchical Dirichlet Process (HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. Inverse Regression (IR) is a sufficient dimension reduction (SDR) technique that makes predictions with provably sufficient information.

First, HDP-IR combines the advantages of both nonparametric topic modeling and inverse regression, HDP-IR avoids the model selection complications and can capture the uncertainty regarding the number of topics HDP-IR provides a SDR for each document, which can improve the predictive performance. Second, we design a scalable variation inference algorithm for fitting HDP-IR such that it can be applied to large-scale corpora (hundreds of thousands or millions of documents).

Following prior STM literature, we design HDP-IR under a hierarchical Bayesian modeling framework. The user should be adding the reviews of the item based on their intension. Then we collect the data in unstructured datasets over the multiple domains and apply the NLP (Natural Language Processing) to identify the similar kinds of reviews on the products. Then apply the collaborative filtering technique for identifying the suitable topics for specific item in various domains based on the user reviews and sort the items.

2.2.1 Advantages Of Proposed System

- Collecting the data in unstructured datasets over the multiple domains.
- Identifying the competitiveness of items based on the number of users reviews.
- Checking whether the representations contain sufficient predictive information about the responses.

2.3 Technologies Used

- JAVA
- Android

Working Of Java:

For those who are new to object-oriented programming, the concept of a class will be new to you. Simplistically, a class is the definition for a segment of code that can contain both data (called attributes) and functions (called methods)

When the interpreter executes a class, it looks for a particular method by the name of **main,** which will sound familiar to C programmers. The main method is passed as a parameter an array of strings (similar to the argv [] of C), and is declared as a static method.

To output text from the program, we execute the **println** method of **System.out,** which is java's output stream. UNIX users will appreciate the theory behind such a stream, as it is actually standard output. For those who are instead used to the Wintel platform, it will write the string passed to it to the user's program.

Java consists of two things:

- Programming language
- Platform

The Java Programming Language

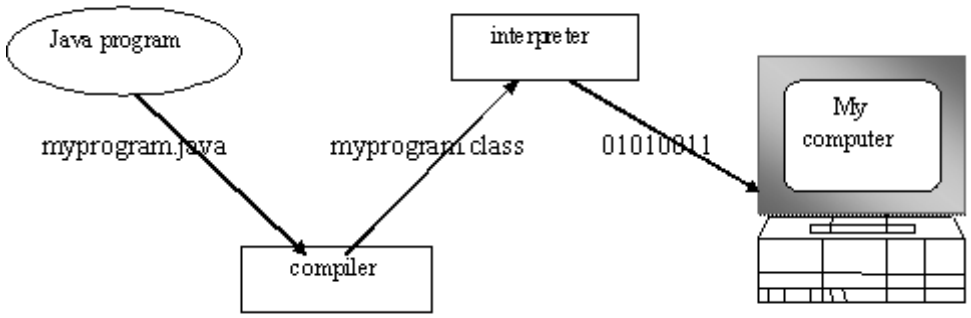
Java is a high-level programming language that is all of the following:

- Simple
- Object-oriented
- Distributed
- Interpreted
- Robust
- Secure
- Architecture-neutral
- Portable
- High-performance
- Multithreaded
- Dynamic

The code and can bring about changes whenever felt necessary. Some of the standard needed to achieve the above-mentioned objectives are as follows:

Java is unusual in that each Java program is both co implied and interpreted. With a compiler, you translate a Java program into an intermediate language called **Java byte codes** – the platform independent codes interpreted by the Java interpreter. With an interpreter, each Java byte code instruction is parsed and run on the computer. Compilation happens just once; interpretation occurs each time the program is executed.

This figure illustrates how it works:



Java Virtual Machine (JVM)

Product Features:

Tomcat 3.x (initial release)

- implements the Servlet 2.2 and JSP 1.1 specifications
- servlet reloading
- basic HTTP functionality Tomcat 4.x
- implements the Servlet 2.3 and JSP 1.2 specifications
- servlet container redesigned as Catalina
- JSP engine redesigned as Jasper
- Coyote connector
- Tomcat 5.x

- implements the Servlet 2.4 and JSP 2.0 specifications
- reduced garbage collection, improved performance and scalability
- native Windows and Unix wrappers for platform integration
- faster JSP paring

History

Tomcat started off as a servlet specification implementation by James Duncan Davidson, a software architect at Sun. He later helped make the project open source and played a key role in its donation by Sun to the Apache Software Foundation.

Davidson had initially hoped that the project would become open-sourced and, since most open-source projects had O'Reilly books associated with them featuring an animal on the cover, he wanted to name the project after an animal. He came up with Tomcat since he reasoned the animal represented something that could take care of and fend for itself. His wish to see an animal cover eventually came true when O'Reilly published their Tomcat book with a tomcat on the cover.

III. Purpose

The main aim of the project is to generate flexible number of topics for predicting the response of interest on the items (product) over the multiple domains based on the user reviews.

Project Scope

We propose a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). Specifically, the Hierarchical Dirichlet Process (HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. Inverse Regression (IR) is a sufficient dimension reduction (SDR) technique that makes predictions with provably sufficient information.

First, HDP-IR combines the advantages of both nonparametric topic modeling and inverse regression, HDP-IR avoids the model selection complications and can capture the uncertainty regarding the number of topics HDP-IR provides a SDR for each document, which can improve the predictive performance. Second, we design a scalable variation inference algorithm for fitting HDP-IR such that it can be applied to large-scale corpora (hundreds of thousands or millions of documents).

Following prior STM literature, we design HDP-IR under a hierarchical Bayesian modeling framework. The user should be adding the reviews of the item based on their intension. Then we collect the data in unstructured datasets over the multiple domains and apply the Nlp (Natural Language Processing) to identify the similar kinds of reviews on the products. Then apply the collaborative filtering technique for identifying the suitable topics for specific item in various domains based on the user reviews and sort the items.

4.3 Product Perspective

The rapid increase in the number of e-commerce has given rise to a significant amount of user generated text, which contains good information about consumer preferences and opinion. However, supervised topic modeling remains a challenging problem.

First, most supervised topic models require pre-specifying the number of topics a priori. Such specification may result in model misspecification when the specified numbers of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features.

4.4 System Features

We propose a novel supervised topic model called Hierarchical Dirichlet Process-based Inverse Regression (HDP-IR). Specifically, the Hierarchical Dirichlet Process (HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. Inverse Regression (IR) is a sufficient dimension reduction (SDR) technique that makes predictions with provably sufficient information. HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain statistically sufficient information for improved predictive performance.

Moreover, we develop an efficient inference algorithm for model estimation that is capable of examining large-scale corpora with millions of documents. Evaluation of HDP-IR in comparison with the state-of-the-art baseline techniques reveals that both increasing the topic structure flexibility and using sufficient dimension reduction could improve the predictive performance on user-generated review text in e-commerce applications, and the proposed inference algorithm is highly effective in terms of its scalability.

4.5 Design And Implementation Constraints

4.5.1 Constraints In Analysis

- Constraints as Informal Text
- Constraints as Operational Restrictions
- Constraints Integrated in Existing Model Concepts
- Constraints as a Separate Concept
- Constraints Implied by the Model Structure

4.5.2 Constraints In Design

- Determination of the Involved Classes
- Determination of the Involved Objects
- Determination of the Involved Actions
- Determination of the Require Clauses
- Global actions and Constraint Realization

4.5.3 Constraints In Implementation

A hierarchical structuring of relations may result in more classes and a more complicated structure to implement. Therefore it is advisable to transform the hierarchical relation structure to a simpler structure such as a classical flat one. It is rather straightforward to transform the developed hierarchical model into a bipartite, flat model, consisting of classes on the one hand and flat relations on the other.

Flat relations are preferred at the design level for reasons of simplicity and implementation ease. There is no identity or functionality associated with a flat relation. A flat relation corresponds with the relation concept of entity-relationship modeling and many object oriented methods.

4.6 Other Nonfunctional Requirements

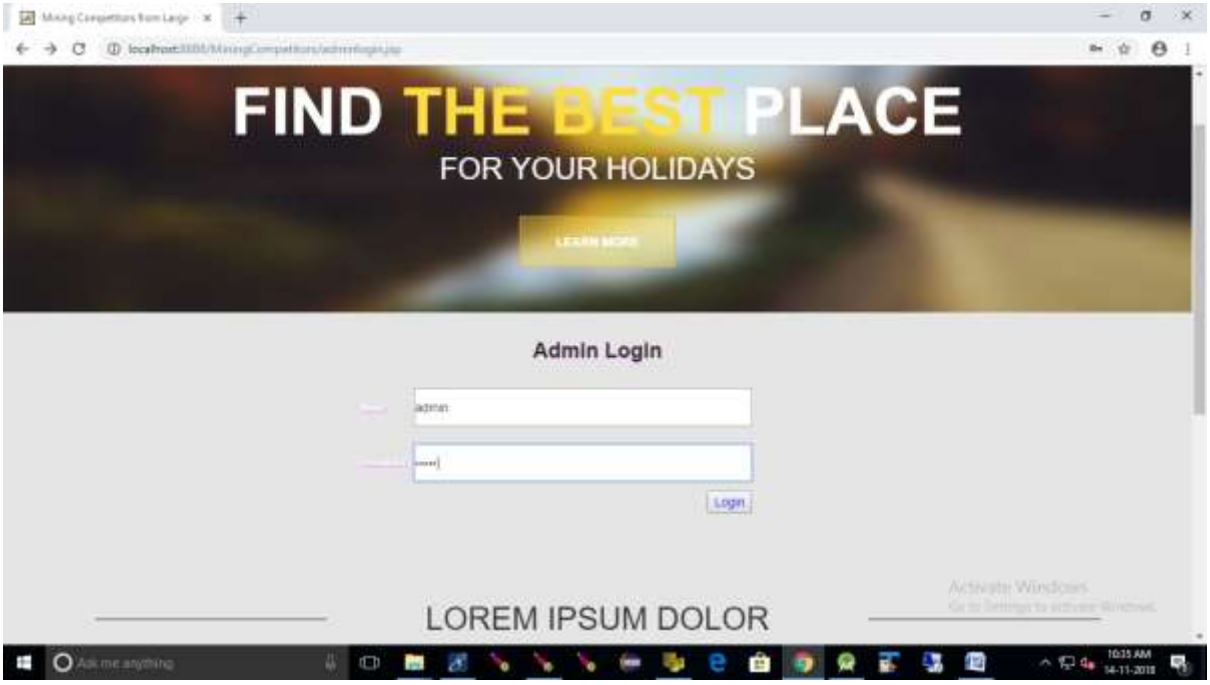
4.6.1 Performance Requirements

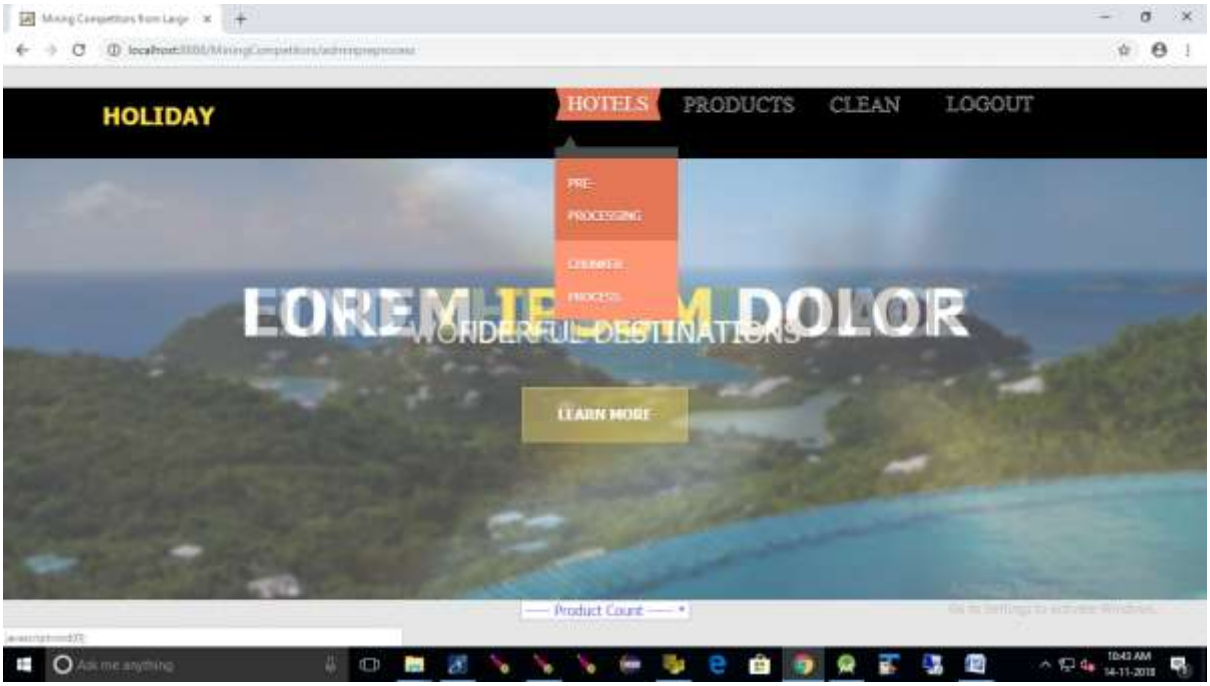
The application at this side controls and communicates with the following three main general components.

- Embedded browser in charge of the navigation and accessing to the web service;
- Server Tier: The server side contains the main parts of the functionality of the proposed architecture. The components at this tier are the following.

Web Server, Security Module, Server-Side Capturing Engine, Preprocessing Engine, Database System, Verification Engine, Output Module.

Screenshots





IV. Conclusion

HDP-IR topic modeling determines the topic structure from the data. makes predictions with sufficient dimension reduction of the document. Able to examine large-scale corpora containing millions of documents. Experimental results revealed that the proposed HDP-IR model significantly outperformed. Proposed HDP-IR model is the first nonparametric topic model leveraging SDR to improve prediction accuracy.

References

[1]. *D. Duan, Y. Li, R. Li, R. Zhang, X. Gu, and K. Wen, "LIMTopic: A framework of incorporating link based importance into topic modeling," IEEE Trans. Knowl. Data Eng., vol. 26, no. 10, pp. 2493–2506, Oct. 2014.*

[2]. *A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," IEEE Trans.Knowl. Data Eng., vol. 20, no. 9, pp. 1168–1180, Sep. 2008.*

[3]. *A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.*

[4]. *Y. Gao, Y. Xu, and Y. Li, "Pattern-based topics for document modelling in information filtering," IEEE Trans. Knowl. Data Eng., vol. 27, no. 6, pp. 1629–1642, Jun. 2015.*