Semantic Affect Sensing in User Generated Contents: An Intelligent Machine Informatics to Audit Human Sentiments

1 Asha.K, 2 Dr.T.Devi

1 Assistant Professor, MCA Department, RV College of Engineering, Bangalore, India. 2 Reader and Head-Incharge, Department of Computer Applications, School of Computer Science and Engineering Bharathiar University, Coimbatore, Tamil Nadu, INDIA

Email: 1 sridhar.asha@gmail.com, 2 tdevi5@gmail.com

ABSTRACT

Opinion Mining is an important research area in text mining domain. Modern enterprises host text mining software’s that transform unstructured text into structured data. These software’s extract entities such as people, places, companies and products. Based on wiki survey, text is considered as an unstructured data which comprises somewhere between 31% to 85% of what is stored in any given enterprise. Opinion mining is the process of extraction of sentiments from human authored documents from different written online sources. This paper describes the different opinion mining techniques, importance of opinion mining, issues, applications and also proposes a structural framework for mining opinions from text documents.

Keywords: Text mining, web mining, opinion mining, sentiment analysis, natural language processing, Innovations and collaborations in Technology management.

1. INTRODUCTION

Humans have opinions on every topic in the world and these opinions are frequently expressed in ways that are difficult for machines to understand and translate. Analysis techniques are evidently the key. There are many documents that present author’s subjective views on particular topics. Sentiment analysis is the extraction of attitudes and opinions from human-authored documents [34]. The analysis of such attitudes [1] and opinions in an automated and structured form will put forward a powerful technology to a number of problem domains including business intelligence, marketing, national security, crime prevention and healthcare/wellbeing services. Mining opinion on user generated contents [13] in particular is an interesting and useful source for sentiment mining. This can also be used to filter emails and other messages, or indicate abusive messages in newsgroups.

Knowledge discovery from text [5] is an emerging field connecting the different worlds of text data and information. The information age has enabled many organizations to gather large volume of data and the usefulness of this data is avoided if “meaningful information” or “knowledge” cannot be extracted from it. Data mining otherwise known as knowledge discovery [2], attempts to answer this need. Knowledge discovery is an iterative process that involves many processing steps. Like other traditional methods, knowledge discovery process model [3] can be applied to the large corpus to extract meaningful information. Data repository contains unstructured/semi structured text documents. On-line libraries, search engines, and other large document repositories are growing rapidly that it is difficult and costly to categorize every document manually. In order to deal with these problems, the automated methods of working with web documents are proposed so that they can be more easily organized with minimal human intervention.

Text mining [7], a sub division of the field called data mining [6] tries to find interesting patterns. While data mining tools are designed to handle structured data bases, text mining is carried out on unstructured/semi structured datasets in the form of e-mails, text documents and html form and hence, working with text mining is a difficult task. But finding the opinion from such kind of unstructured data is a major challenge. The problem introduced by opinion mining is very clear: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from understanding the natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that machines cannot easily handle such as colloquial speech, spelling variations and contextual meaning, idioms, and irony. However, although our language capabilities allow us to understand unstructured data, machine fails to achieve the ability to process text in large volumes or at high speeds.

In enterprise perspective, text mining is an answer to the unstructured data challenge. The scope of the challenge is expressed in the truism that 80 percent of enterprise information originates and is locked in “unstructured” form[4]. How complete can organizations business intelligence or its voice of the customer initiative be if it does not accommodate a majority of business relevant information? Sentiment analysis is one of the most exciting applications of text mining today. Because sentiment - with
opinions and attitudinal information - is central to human communications, it is also of key importance for finding the market intelligence.

The remaining structure of the paper is organized as follows: Sections 2 and 3 deal with the discussion on web data mining and the comparison study between text mining and data mining, section 4 presents brief history of opinion mining, and the remaining part of the paper covers different and data mining, section 4 presents brief history of opinion mining, and the remaining part of the paper covers different methodologies, proposed model for evaluation, challenges, issues and different real time application areas. The paper concludes with motivating the readers on capturing the opinion by analyzing the body language.

2. WHY DO WE NEED WEB MINING?

The widespread use of Internet, along with the social web, has led to the birth of new and challenging technologies. The possibility to express opinion "by anyone, anywhere, on anything", in blogs, forums, review sites has made it possible for people all around the world to take better decisions. The companies are more informed on the impact they have on people, because the large amount of opinions expressed on them offers a direct and unbiased, global feedback. The uncontrolled expression of opinions or points of view has given way to negative social behavior. Due to the large volume of such data, automatic systems need to be built to deal with it, both to encounter ways to positively exploit it, at the same time determine appropriate measures to deal with the issues it raises.

Web documents have well-defined structures such as letters, words, sentences, paragraphs, sections, punctuation marks and HTML tags. In 2007, a business study has proven that 85% of all digital business information, most of it web-related, is stored in unstructured formats i.e. Non-tabular formats and spreadsheets [4]. Developing improved methods of performing machine learning techniques on this vast amount of non-tabular, semi-structured web data is therefore highly desirable. Recent advances have led to the newest and hottest trends in data mining- text mining and web mining. These two technologies open a rich vein of particular dataset in the form of textual comments from warehouse. The knowledge discovery from text involves not only plenty of semi structured and unstructured text data but also the complicated semantic relation and hence, the existing datamining technology cannot be applied directly.

Web mining is a simplified form of KDD and is often associated with information extraction and retrieval. Information extraction has the goal of transforming a collection of documents, makes the structure, and represents those documents and also aims to extract relevant facts from them while information retrieval aims to select relevant documents [11]. Manual creation of an information extraction system is not practical and scalable for dynamic medium such as web contents [12]. Due to this nature of web, most of the extraction system focus on specific websites and the tremendous growth of information available on the web [9] and the high interest in e-commerce makes this area of research, a potential one.

3. DATA MINING VS TEXT MINING

Text mining seems to be an extension of the better known data mining. Data mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in different business situations. But, only about 20% of the data on enterprise [8] and on the World Wide Web are numbers - the rest is text. The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations [35].Text mining tries to apply these same techniques of data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning from unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis and this process is called as feature extraction and is a crucial step in text mining.

Text mining (TM) is a comprehensive technique and it relates to data mining, machine language, information extraction, natural language understanding, and knowledge management. Its object type is not only structural data but also semi-structured data or unstructured data. TM is also used to retrieve the unnoticeable concepts explicitly and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to work with large quantities of text data. Knowledge discovery from text, while deeply rooted in NLP, sketch on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. It also plays a significant role in emerging applications such as text understanding.

4. OPINION MINING: A BRIEF HISTORY

Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text [15]. Opinions are subjective statements that reflect people’s sentiments or perceptions about the entities and events. Much of the existing research on text information processing has been focused on mining and retrieval of factual information, e.g., information retrieval, web search, and many other text mining and natural language processing tasks. Little work has been done on the processing of opinions [14]. Opinions are so important that whenever one needs to make a decision, others’ opinions are to be heard.
These techniques are appropriate in times of high turbulence and when markets are emerging and the sources of data cannot provide clear information or present more ambiguous information. This could be especially the case when emotions are to be extracted.

a. Why Traditional method Fails

Text lived for years on stored documents, every word and enforced consistency, summarily dismissing changes from well-intentioned copywriters. So getting the right opinion from it was important. Sentiment analysis/opinion mining is a new methodology to distinguish the market strategies internationally by its survey research and experience. Through the help of this method, access respondents and provide timely and relevant information to clients. Extraction of emotion from text documents research has a huge impact on today’s enterprise. Understanding which amount of text messages work and why it helps to the company connect with its audiences in a way that motivates them to take action. To gain this understanding, it is ideal to conduct qualitative research to refine the text message. The problem is that companies are faced with doing more with less time and fewer resources. So market strategies created a web-based quantitative survey technique that uses best practices from qualitative research. This also can be considered as an integrated research that aims at time and money saving objectives that reveals strongest opinion on text elements, explains why certain themes and messages resonate best, and explores these reactions across multiple segments or subgroups of buyers. Unlike traditional text mining research, merges this feedback with follow-up questions to pinpoint whether customer reactions are a matter of content or just wording and phraseology by adding verbatim responses to reliable, quantitative data. Some selected and most prominent use cases for opinion mining techniques are explained here. Some have been already related in various research experiments previously mentioned while others still remain a goal to achieve in the future.

b. Product benchmarking and market intelligence

The key to selling a product is responding to customer’s demands in proper time and in the right location. Many companies spend huge money on market analysis and hire external specialized consulting companies. The opinion mining techniques could aid this effort and potentially minimize costs. Market analysis done by specialized companies is needed to take certain amounts of time and effort, while in many cases getting fast access to accurate market data can be a key factor. The right opinion mining tools could create a business advantage for a company to get ahead of its competitors and swiftly react to customer needs. Additionally opinion mining opens new frontiers. With the immense amounts of community created data on the Internet its analysis becomes impossible or very difficult and expensive without some automatic methods. This domain is huge and the amount of appliances possible is vast.

5. BASIC NOTION OF OPINION MINING FROM TEXT DATA

Opinion mining is the computational study of opinions, sentiments and emotions expressed in text. Machine learning was applied in information retrieval long before the recent advances of the Web. One of the main reasons for the lack of study of opinions is the fact that there was little opinionated text available before the World Wide Web. In the past few years the world has been transformed with the explosive growth of internet forums, discussion groups, and blogs. Web drastically changed the way that people express their views and opinions. Finding the opinion sources and monitoring them on the web can still be a formidable task because there are large numbers of diverse sources. It is difficult for a human reader to find relevant sources, extract related sentences with opinions, read them, summarize them and organize them into usable forms. Thus, automated opinion discovery systems are needed. The system can be categorized into different processing steps. The steps are named as classification and information extraction, named entity recognition, review and summarization, and clustering.

a. Classification

Text classification is a supervised learning problem that classifies textual documents according to their predefined categories as positive and negative. Machine Learning techniques have been applied extensively to text classification and text clustering. Five text classification methods that have been widely examined are Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Naïve Bayes, Neural Network (NNet) [18], and Rocchio method [19].

b. Entity Recognition

An object in a document can be represented as a tree, hierarchy or taxonomy. The root of the tree is the object itself and each non root node is a component or sub-component of the object. Each link is a part of relation and each node is also associated with a set of attributes or properties. An opinion can be expressed on any node and any attribute of the node.

c. Review

It is the process of concluding the document, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative
methods must be considered. One of the strategies most widely used by text summarization, and sentence extraction, used for summarizing the contents from sentence and important statements.

d. Clustering

Text clustering is unsupervised learning that groups documents into categories dynamically according to their similarities. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents. For text clustering, the Estimation Maximization (EM) algorithm has become the basis for many unsupervised clustering algorithms [20].

6. PROPOSED FRAMEWORK FOR EVALUATION

Any application should be implemented with the help of evaluation metrics and these metrics show the accuracy and performance efficiency of the schema that is used to examine the learning algorithms. The framework consists of three major stages: preprocessing, classification and analysis. The sub-stages beneath each major stage are as given in Fig.1.

![Figure 1: Structural Framework](image)

The data preprocessing part describes definitions and methodology of the research, review of some existing work on the automatic text classification and gives a brief introduction to the learning algorithms. Classification describes the characteristics and processing of the data sets, classifiers and the system environment, and a method to measure the accuracy. Testing and analysis describe the result of the experiments and the evaluation. One can compare the performance of the learning algorithms, and identify the most appropriate business intelligence application.

7. ISSUES IN DEVELOPING OPINION MINING SYSTEMS

The beginning of Web 2.0 and social media content has motivated much excitement and created rich opportunities for understanding the opinions of the general public and consumers towards social events, political movements, company strategies, marketing campaigns, and product preferences. Various business related research questions can be answered by analyzing the millions of comments and responses expressed in different blogs, forums, and social network sites. Opinion mining solves the computational techniques for extracting, classifying, understanding, and assessing opinions expressed in various online news sources, and other user generated contents.

There are still several promising new directions for developing and advancing opinion mining research. Much work in this research area focused on English, Chinese, Arabic and other European languages. Most of the advanced work has been developed especially in English. Framework and methods for integrating sentiments and opinions expressed with other computational representations such as product features. Current OM research has focused on business and e-commerce applications, finance wal-mart forum and Market intelligence 2like Ya0. Liu [21] introduces some of the problems and suggests several technical challenges including object identification, feature extraction, and synonym grouping, opinion orientation classification, and integration. Many real-life applications require more detailed analysis because user often wants to know the subject of opinions. To explore the generic issues in developing the opinion mining systems, a review of multiple segments has been carried out by Liu [21].The summary proves that people can express opinions on any target entity-products, services, individuals, or events. In this context, the term object is used to denote the target entity [22], opinion holder [23], object and time extraction [24, 25, 26], direct opinion [27] and comparative opinion [28] has been observed.

The objective of opinion mining gives us a good clue of the main tasks involved and technical challenges. In object identification [29], similar to entity recognition problem, the user wants to find the opinions on some competing objects such as competing products or services. Product feature extraction [26, 30, 31] like the clarity, voice and quality of product remains to be a major challenge. Also need to group synonym features because people often use different words or phrases to describe the same feature. Opinion orientation identification [32, 33] and information integration are also commented on.
8. OPINION MINING APPLICATIONS

Opinion Mining (OM) techniques are used to manage customer relations and market analysis strategically, mainly with applications aiming at analyzing customer’s opinions as well as monitoring the level of investor satisfaction. In the context of market analysis, OM techniques are often utilized to monitor and analyze the state of health of a company by means of the systematic analysis of fluctuations in the share market.

The main objective is to create the programs/algorithms that could check to meet the rivals of customers as well as the market volatility. The necessity and tracing of the opinion/sentiment must be to analyze and utilize the buying/selling patterns of the customers at any given point of time. The algorithms should be used to analyze, understand and react to information contained within a variety of content sources. With the help of OM technology, old technology can be improved to provide the best in text and sentiment analysis solutions and the investors can do market analysis from the unstructured trading portals. The development goes far beyond standard digital era like most read and e-mailed lists. The computers are actually parsing writers words, sentence structure and even the emotions. Without human intervention, one can use these algorithms to interpret the news and trade on it. The technology has the strength to analyze the entire blogosphere, the subtlety enough to look for the slightest trends in confidential records and the flexibility to solve completely novel problems in future. This application focuses on customers of enterprise search companies, social media firms, market research companies and financial services companies.

a. Opinion Mining Applications in CRM and Market analysis

The following are few applications of opinion mining in the area of CRM and market analysis.

Customer Relationship Management: In CRM [16] domain, the most widespread applications are related to the management of the contents of clients’ messages. This kind of analysis often aims at automatically rerouting specific requests to the appropriate service or at supplying immediate answers to the most frequently asked questions. Services research has emerged as a green field area for application of advances in computer science and IT. CRM practices, particularly contact centers (call centers) in our context, have emerged as hotbeds for application of innovations in the areas of knowledge management, analytics and data mining. Unstructured text documents produced from a variety of sources in today’s contact centers have exploded in terms of the sheer volume generated. Companies are increasingly looking to understand and analyze this content to derive operational and business insights. The customer, the end consumer of products and services, is receiving increased attention. Analytics and business intelligence (BI) applications revolving around the customer has led to emergence of areas like customer experience management, customer relationship management and customer service quality. These are becoming critical to competitive growth, and sometimes even, survival. Applications with such customer focus are most evident in services companies, especially CRM practices and contact centers. Customers interacting with contact centers, typically fill out feedback surveys with free form text comments which are directly indicative of their (dis)satisfaction with products, services, or interactions. Currently, only manual analysis is done, if at all, on a small sample of available data. Various problem buckets identified by the analysis, text classification can help automate this making it consistent and exhaustive.

b. Market Analysis

Market Analysis uses OM technique mainly to analyze competitors and/or monitor customers’ opinions to identify new potential customers, as well as to determine the companies’ image through the analysis of press reviews and other relevant sources. For many companies, tele-marketing and e-mail activity represent one of the main sources for acquiring new customers. The opinion mining makes it possible to present also more complex market scenarios. Traditional marketing had a positive impact due to technology over the past few decades. Database technologies transformed storing information such as customers, partners, demographics, and preferences for making marketing decisions. In the 90s, the whole world saw economy boom due to improvements and innovation in various IT-related fields. The amount of web pages ameliorated during dot-com era and search engines were found to crawl web pages to throw out useful information from the heaps. Marketing professionals used search engines, and databases as a part of competitive analyses.

Data mining technology helps to retrieve useful information and find chunks from various databases. Data warehouses turned out to be successful for numerical information, but failed when it came to textual information [15]. The 21st century has taken us beyond the limited amount of information on the web. Opinion mining technology could help marketing professionals to use this information for finding nuggets. Market Analysis deals with different applications by finding the data sources for analysis like credit card transactions, loyalty cards, customer complaint calls, target marketing, determine customer purchasing patterns, cross-market analysis, and asset evaluation. Resource planning also acts as an application, dealing with review and compares the resources.
9. CONCLUSION

Web is a great source of information where the information itself is buried under the visual markups, texts, and links of the web pages; discovering relationships between entities in text document is an interesting problem. The repeated occurrence of loosely defined structures in text and their relationships helps to define these entities with increased confidence. The scope of further improvement lies in the context of capturing the opinion related body language expressed by virtual worlds (second life participants) such as thumbs up, thumbs down, and applause for opinion mining. In this survey, iterative process of text mining for finding opinion and sentiments, their techniques, issues and how practical applications can help in today’s enterprise are discussed. This paper will be a fine grained medium for researchers who work with extracting emotions from text documents.

REFERENCES


