STUDY OF ANALYSIS OF SOCIAL MEDIA DATA AND CLASSIFIERS

Mr. Shishir Kumar Singh.

Miss. Priyanka Ravindra Bagadi.

Miss. Disha Prasad Koparde.

Miss. Shraddha Pradeep Barangale.

Prof. Swapnil Shinde

Abstract

We live in an age of Big Data. With hundreds of millions of people spending countless hours on social media to communicate, connect, interact, share information with our friendsand expressour feelings. Social media data is uncategorized, free format therefore it is very time consuming to access this kind of data for user and hence becomes mandatory to go through every post. Also this social media data may lead to rise of communal riots, social wars and cyber bulling. Hence overcoming these above problems this system introduces classification and analysis of social media data for efficient browsing. This system also helps to reduce todays increasing safety issues of women by providing women safety application.

Introduction

In today’s era, social media has become a unique source of big data. It is a rapidly growing new field. With this the web has become a vibrant and lively realm in which billions of individuals all around the globe interact, share, posts and conduct numerous daily activities. It enables us to be connected with each other anywhere anytime.

With this existing system users have to go through each and every posts even if they are not interested. Also unfortunately sometimes these posts or comments turns out to be harmful towards the society or an individual which may lead to communal riots, social wars. Hence this system introduces a desktop application in which user will get posts displayed in categorized manner such as educational, entertainment, political, history and sports. It also analyses the posts and cancels it which is harmful for the society. Women safety is a challenging issue in India as well as in other countries. It is not safe for a women to travel alone at mid-night or at unknown places. There should be a helping hand for women. There are many android applications based on women safety but they are less efficient or less user friendly. So in order to solve this issue this system develops an android application which is easy to use and efficient to provide help to a victim.

Hence our goal is to enhance women safety, reduce communal riots and social wars.

I. Analysis

Analysis of social media comprises two main parts (1) Data Generated from social networking site or application (2) Sophisticated analysis of that data, in many cases requiring real-time (or near real-time) data analytics, measurements which understand and appropriately weigh factors such as influence, reach, and relevancy, an understanding of the context of the data being analyzed, and the inclusion of time horizon considerations. [1]

1. Key concepts to understand in social data analysis

When we talk about social media data the following factor associate with it:

Sophisticated Data Analysis: what distinguishes social data analytics from sentiment analysis is the depth of the analysis. Social data analysis takes into consideration a number of factors (context, content, sentiment) to provide additional insight.

Time consideration: windows of opportunity are significantly limited in the field of social networking. What's relevant one day (or even one hour) may not be the next. Being able to quickly execute and analyze the data is an imperative.

Influence Analysis: understanding the potential impact of specific individuals can be key in understanding how messages might be resonating. It's not just about quantity, it's also very much about quality.

Network Analysis: social data is also interesting in that it migrates, grows (or dies) based on how the data is propagated throughout the network. It's how viral activity starts--and spreads.

2. Sentiment analysis: ***It*** is becoming a popular capability in the newest social media monitoring tools. It tells you if a post is positive or negative, by how much, and gives you an indication as to why. An entire post can be marked as positive or negative, but more useful is drilling down into the content and discovering the objects of positive or negative sentiment [4]. Sentiment analysis API allows a smart phone manufacturer to understand that 68% of posts are positive about the screen size, but only 20% are positive about the battery life. This provides much more actionable feedback than mistakenly hearing that “80% of posts that mention the product are positive.”

3. Alchemy API for Social Media Monitoring

Only a tiny amount of the rich information in a social media post is structured. Easy to get data like a post’s date, location and username are useful, but the vast majority of what we need to understand is completely unstructured [2]. What a person says, what they really meant, whether they are positive, negative or neutral is hidden from us unless we read each post. Alchemy API is able to analyze the oceans of posts and extract data that formerly could only have been done by humans. By discovering important information like keywords or entities and analyzing sentiment, it is possible to process the data and answer big-picture questions like “Are these tweets about us?” “What’s hot and trending in our communities?” “Which features do they love or hate?” “How do we stack up against our competitors?”

3.1. Various Text Powering functions for Powering Social Media Analysis

1. Entity Extraction

Find the people, companies and organizations that are mentioned in a social media post, and disambiguate them to increase accuracy.

2. Keyword Extraction

Detect the features and components of your company or product, such as "screen size" or "battery life" for smart phones.

3. Sentiment Analysis

Discover the positive and negative social media posts about your company, and drill down into the specific drivers of the sentiment.

4. Relation Extraction

Relations detect important events and signals between entities, such as “Broncos crushed Chiefs” or "Apple just bought Topsy Labs for $200 million."

5. Language Detection

Social media is everywhere, and users post in many different languages. Use language detection to identify and categorize the posts written in the languages that you care about.

6. Author Extraction

Million of bloggers are publishing every day. With author extraction you can monitor which bloggers are friends, which are foes, and who holds the most influence with your audiences.

4. Sentiment Lexicon:

It is a dictionary of sentimental words user often used in their expression. The common words are listed into Sentiment Lexicon that enhances data mining techniques when used mining sentiment in document. Different corpus of sentiment lexicon can be created for variety of subject matters. For Example sentimental words used in sport are different from those used in politics. Expanding the occurrence of sentiment lexicon helps to focus more on analyzing topic-specific occurrence, but with the use of high manpower [3]. Lexicon-based approaches require parsing to work on simple, comparative, compound, conditional sentences and questions .Sentiment lexicon can be expanded by use of synonyms. Nonetheless, lexicon expansion through the use of synonyms has a drawback of the wording loosing it primary meaning after a few recapitulation. Sentiment lexicon can also be enhanced by 'throwing away' neutral words that depicts neither positive nor negative expression. Neutral expression is common especially in products ratings and reviews.

II. Classifiers

Data mining from social media is the process of extracting relevant information from the web documents, which can be done either manually or automatically by using different classifiers. This collection of extracted data may be in the different forms of content. This social media data mining involves data extraction, data analysis, data classification and data clustering which results in formatted labeled data. There are researches going on for retrieval methods web information and processing of natural language. Applying different types of classifiers for extracting the data from social media sites classification is done. This system uniquely classifies the social media data by learning the information given in the data.

Classification is the process of separation of objects into different categories or classes. In text classification text automatically is assigned into predefined categories. In machine learning, classifiers learns the features of set of training data which is extracted from the web to classify it into different categories. Social media data mining also can be done on unstructured data like text or images. Application of data and text mining techniques are used for social media data mining. For classification the methods which are commonly used are as follows:-

i. **Bayesian Classifier:** It is the most commonly used classifier for text classification. Naïve Bayes uses the principle of Bayesian theorem with independent feature selection. It is used for anti-spam filtering technique. This classification involves two phases-first phase is for training data set and the second phase is for classification [5].

In Naïve Bayes classifier Prior Probability is the ratio of number of single objects and number of total objects. It is based on previous experience. Likelihood is used to specify the category of the object. Posterior probability is the combination of Prior probability and Likelihood which is used for the final categorization [6].

ii. **K-Nearest Neighbor*:*** It is case based learning algorithm which calculates similarities between a text document and each neighbor. It is non parametric, easy to implement and very effective[7]. For classification of observation it uses distance or similarity functions. However, it is difficult to find optimal value for k and the time for classification is very long.

iii. **Support Vector Machine:** SVM requires both positive and negative training set for maximizing the margin it requires to find out linear separating hyperplane. Support vector are the representatives which are closest to the decision surface [8]. The hyperplane equation is WX+BX=0, where X is arbitrary object,W is vector,B is constant.For separation of two different classes hyperplanes are used. SVM can be used for pre-classified documents.

iv. **Neural Network:** Neurons have ability to extract relevant information from large set of data[8]. It is used for pattern recognition, feature extraction and noise reduction. The authority of one neuron on the other can be determine from the connection between them and the weight of connection determines the strength of authority. In neural network two types of learning methods used are 1) Supervised and 2) Unsupervised. For the logical management of text from social websites neural network can be used.

Classifiers	Advantages	Disadvantages
Bayesian Classifier	• Effective for numeric as well as textual data. • Efficient implementation and computation.	• Conditional independence assumptions affects the posterior probability estimate. • Low performance for small data.

K Nearest Neighbor	• Non-parametric. • Effective.	• Time consuming classification. • Difficult to find optimal value for k.
Support Vector Machine	• Better feature selection of data.	• Kernel selection is difficult.
Neural Network	• Can be used for discrete and continuous data. • Fast testing.	• Causes overfitting. • Slow training. • Results are difficult to interpret.

Conclusion:

Enumerous Technologies are being developed for fetching meaningful data from huge collection of textual data using various text mining algorithms. Classification and analysis becomes more challenging when the data is unstructured. Hence, through this survey we further conclude that various algorithms perform differently depending upon the collection of data. In this survey we studied different classifiers and analyzers with their advantages and disadvantages. Some of them give satisfactory results and some do not perform well.

References:

[1] http://www.ibm.com/analytics/watson-analytics/solution/add-ons?cm_mmc=search-gsn-_-branded-watson-social-media-_-analyze%20social%20media-Broad-_-IND-WW-WA-mkt-oww

[2] http://www.alchemyapi.com/

[3] Mariam Adedoyin-Olowe, Frederic Stahl, “A Survey of Data Mining Techniques for Social Network Analysis”.

[4] Boiy, E. and Moens, M.: “A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. Information Retrieval ”, 12(5):526-558, 2009.

[5] Xin Chen, Student Member, IEEE, and Krishna Madhavan, “Mining Social Media Data for Understanding Student’s Learning Experiences”, vol. 7, No.3, July-Sept 2014.

[6] G. Tsoumakas, I.Katakis, "Mining Multi-label Data," Data Mining And Knowledge Discovery Handbook, pp. 667-685, Springer, 2010.

[7] Min-Ling Zhang, Zhi-Hua Zhou "A k-nearest neighbor based algorithm for multi-label classification" 25-27 July 2005.

[8] Ms. Priyanka Patel, Ms. Khushali Mistry, "A Review: Text Classification On Social Media Data," pp 80-84, Jan-Feb 2015.

[9] L. Jing, M.K. Ng, and J.Z. Huang, "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data", IEEE Trans. Knowledge and Data Eng., 2007

[10] Nadir Omer Fadl Elssied and Othman Ibrahim, "K-means Clustering Scheme for Enhanced Spam Detection", March 15, 2014.