JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

ANALYSIS OF DATA-MINING TECHNIQUE APPLIED IN SOCIAL MEDIA

Janardhan G, Dr.S. Suresh Raja

[1]Research Scholar, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, MadhyaPradesh, India
[2]Research Guide, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road,Madhya Pradesh, India

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: Today, the use of social networks is developing ceaselessly and quickly. More disturbing is the way that these networks have become a significant pool for unstructured data that belong to a large group of areas, including business, governments and health. The increasing reliance on social networks calls for data mining techniques that is likely to facilitate reforming the unstructured data and place them inside a systematic pattern. Data mining techniques are capable of taking care of the three predominant research issues with social media data which are size, noise and dynamism. This paper analyze data mining techniques currently being used on social media.

I. INTRODUCTION

Data mining is the collection of techniques for efficient discovery of previously obscure, substantial, novel, useful and understandable patterns in large databases. It is generally used by business intelligence associations, and monetary investigators, however it is increasingly used in the science to get data from the enormous data sets generated by modern experimental and observational methods.

Data mining introduces fundamental concepts and principle algorithm suitable for exploring enormous social media data, it discusses theories and methodologies from different direction, for example, computer science, machine learning, social network examination, data mining, optimization and mathematics. Social media asserts to the development of these social networks wherein individual collaborate with one another through friendship, emails, BlogSpot and many other mechanisms. Social media mining intents to make sense of these people enclosed in networks.

There are many different types of techniques have been developed to overcome the problems, for example, size, noise, and dynamic nature of social media data. Due to different types of data and massive volume of data in the social media, it requires a programmed data processing so as to analyze it inside a given time length. Different types of data mining techniques are as per the following.

Unsupervised classification

We can easily decide a review as 'approval' or 'disapproval' by utilizing unsupervised learning. This type of checking should be possible finding the phrases including an adjective or adverb. We can estimates the semantic orientation of every phrase by utilizing PMI-IR followed by the gathering of the review by utilizing the mean semantic orientation of the phrase

Sentiment orientation

Sentiment orientation can be positive, negative, or neutral (no opinion).It may be immense for the future buyers to make the decision regarding the purchase of a product by following usable reviews which are attracted by the widespread products. Semantic orientation is additionally used by the application developers for their application positioning so they could see the reviews presented by the users. Where the rating is represented as 5-star scale with 5 indicating the best ranked while one denotes the helpless positioning.

Sentiment lexicon

Sentiment lexicon is a collection of sentimental words that are used by reviewers in their expressions. Sentiment lexicon is an index of the normal words that intensify data mining techniques. Different aggregation of sentiment lexicon can be created for assortment of subject matters. For example sentimental words used in legislative issues are often different those used in sports. Expanding the occurrence of sentiment lexicon helps to zero in more on dissecting topic specific occurrence, however with the use of high manpower, Lexicon-based approaches require parsing to take a shot at simple, comparative, compound, contingent sentences and questions.

Feeling definition and outline

These are the significant techniques conceding opening. Assessment definition can be discovered in a text, sentence or the document's topic, and it can likewise involve the whole document. Supposition extraction is hard for rundown and following of any document. Utilizing this technique, the biased (fixed views) part is explored in the texts, and documents. It is required to aggregate the assessment since all the sentiments fetched in the document are not as a direct result of consequence concerning the topic under examination. It assumes a fundamental role in the business associations and government offices by helping in improving the products and policies respectively.

Fundamental clustering technique

Clustering can be considered the most significant unsupervised learning problem; it deals with finding a structure in a collection of unlabelled data. A cluster is therefore a collection of objects which are comparative between them and are unlike the objects belonging to other clusters. Clustering techniques can be applied in many fields, for instance: Marketing, Biology, Libraries, Insurance, City-arranging, Earthquake studies, and WWW (World Wide Web). Clustering techniques involves four most used clustering algorithms; K-means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians. So that, Kmeans is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is clear and finally, Mixture of Gaussian is a probabilistic clustering algorithm.

Sentiment extraction

This technique is necessary so as to point that piece of the document including genuine sentiment. A person's assessment regarding a skilled subject does not matter unless that specific individual has mastered that specific area. However, the use of both sentiment extraction and outline is essential because of the conclusion from many people. The massive number of people offering their input regarding a certain subject, it will be more critical to take out that specific. Other types of unsupervised learning which are being used these days are POS (Parts of Speech) labeling. Sentiment extremity is the twofold classification technique that classifies the opinionated document into predominantly positive or negative feeling.

Semi-supervised Classification

Semi-supervised learning is an objective targeted action however unlike unsupervised; it very well may be specifically evaluated. Equivalent word and antonym comparatives were added to the seed sets in an online word reference. The methodology was meant to produce the extended sets P' and N' that makes up the preparation sets. Other learners were employed and a double classifier was fabricated utilizing every glosses in the word reference for both term in P'∪ N' and making an interpretation of them to a vector [27], [51], their methodology discovers the starting point of data which they reported was absent in earlier techniques used for the task . Semi-supervised lexical classification proposed by [68] integrating lexical knowledge into supervised learning and spread the way to deal with comprise unlabelled data. Cluster supposition that was engaged by gathering two documents with the same cluster essentially supporting the positive - negative sentiment words as sentiment documents. It was noted that the sentiment extremity of document decides the extremity of word and vice versa.

In semi-supervised learning, used extremity detection as semi supervised label engendering problem in a diagram. Each node representing words whose extremity was to be discovered. The results shows label spread progresses exceptionally above the baseline and other semisupervised techniques like Mincuts and Randomized Mincuts. Crafted by compared diagram based semi-supervised learning with regression and [60] metric labeling which runs SVM regression as the first label preference work comparable to comparability measure they proposed. Their result shows that the chart based semi-supervised learning algorithm as per PSP correlation (SSL+PSP) proved to perform well.

Supervised Classification

While clustering techniques are used where premise of data is established however data pattern is obscure, classification techniques are supervised learning techniques used where the data association is already identified. It is deserving of mention that understanding the problem to be solved and selecting the correct data mining tool is very essential when utilizing data mining techniques to solve SM issues. Pre-processing and considering security privileges of individual (as mentioned under research issues of this paper) ought to likewise be taken into account. Nonetheless, since SM is a unique platform, impact of time must be reasonable in the issue of topic recognition, however not generous on account of network enlargement, bunch behavior/influence or marketing. This is because this attributes will undoubtedly change now and again. Data updates in some SM, for example, twitters and Facebook present Application Programmers Interfaces (APIs) that makes it possible for crawler, which gather new data in the site, to store the data for later usage and update.

A supervised learning algorithm used the blend of multiple bases of realities to label couple of adjectives having comparative or unique semantic orientations. The algorithm resulted in a diagram with nodes and connections which represents adjectives and closeness (or uniqueness) of semantic orientation respectively. Then again employed naive Bayes, Maximum entropy classification and Support Vector Machines to examine whether it is enough to treat sentiment classification exclusively as a topic-based categorization of positive or negative, or presumably special sentiment-categorization methods must be assembled.

II. CLASSIFICATION ALGORITHMS

Naïve Bayes, SVM and Maximum Entropy classification techniques have been the three significant ones normally being used when dealing with sentiment examination Clarke et al use WEKA tokenizer with standard setting, and other classification tools like Decision Tree, K-nearest neighbor land Bayesian networks to create Error Reduction (JRip). Mix of Support Vector Machine, Multinomial Naive Bayes and Maximum Entropy, was employed by method of pipeline cascade style to discover the emotions expressed by people in Dutch, French and English language. The technique was used to discover to their experience in utilization of certain products. The methodology endorsed the use of active learning in labeling preparing examples giving evident enhancements in the all out results. Different attribute selection comparism including MI, IG, CHI and DF together with learning methods, for example, K-NN, Naive Bayes, SVM, Centroid and Window classifier in feeling mining. Their experiment proved IG as the premium for sentiment phrases collection while SVM performs best for sentiment classification. In their work, clearly sentiment classifier rely solely on topics and space. Features of supervised learning are perhaps the most widely studies problem.

Support Vector Machine

Support Vector Machine (SVM) is one of the significantly experimented data mining techniques in sentiment examination of SM. It is a kernel-based supervised learning technique. SVM is as yet an unresolved research problem however a very quick technique in preparing the evaluation. The use of SVM in sentiment examination can be supposed to be mainstream because of its non-linear nature which makes it simple to evaluate both theoretically and computationally. SVM is a model of information yield efficient relationship with the yield variable capable of being uttered nearly as a linear blend in its information vector components. SVM has the greatest efficiency at traditional text categorization when compared with other classification techniques like Naïve Bayes and Maximum Entropy.

Naïve Bayes

Naïve Bayes algorithm uses contingent probabilities by checking the occurrence of values and mixes of values in the chronicled data. It is therefore called the probabilistic method. Naïve Bayes is likewise an efficient technique of mining weather forecast. Naïve Bayes is one of the three significantly used supervised learning in sentiment examination. In Naïve Bayes algorithm and paired keyword were simultaneously used to produce a single dimensional degree of sentiment entrenched in tweets from twitter network. Creators used a mix of contextual lexical knowledge with Naïve Bayes. A generative model of sentiment lexicon was created with another model trained on labeled article. A multinomial Naïve Bayes was generated with the circulation from the two models pooled together to incorporate the two bases of data. Work of recommended Adapted Naïve Bayes (ANB) in attempt to deal with the problem of space transfer regular to supervised sentiment classifier. In their experiment they made use of the old-space data and the labeled new-area data suggesting an effective measure to control data from the old-area data. The result of their experiment shows that recommended method can enhance the performance of base classifier essentially and produce greater performance than the naïve Bayes Transfer Classifier (NTBC) which is transfer-learning baseline.

Neural Network

Neural Network (NN) unlike the SVM is a non-linear technique which makes it not very famous data mining technique to use in sentiment examination. There are two principle setbacks associated with non-linear techniques namely; 1) problem of dissecting theoretically and 2) problem of deciphering them computationally (Zhang, 2001). However, as a kernel-based technique, it is positive that with required features the prediction strength of linear techniques can be transformed to be as effective as that of non-linear techniques. NN is usually used for predicting monetary performance and settling on budgetary decision. The back-engendering algorithm of NN was employed to incorporate essential and technical investigation for money related performance prediction which demonstrates NN integration thrives more during economic recession. The experiment result reports that NN performed above the base benchmark on diverse investment approach.

K-NN Algorithm

Based on research done up until now, k-NN has not received elevated level of attention in sentiment investigation as much as SVM, Naïve Bayes and Maximum Entropy which are widely explored in large number of experiments in sentiment examination of social media. In experiment on favorability investigation and sentiment examination included kNN as one of the classification algorithm while testing the pseudo-subjectivity task. Their result revealed the advantage of utilizing attribute selection and pseudo-sentiment task. Conversely, the experiment result didn't report the performance of k-NN classifier or use it as trade-off as interpretability improvement. This makes the relevance of its consideration in the experiment inadequate.

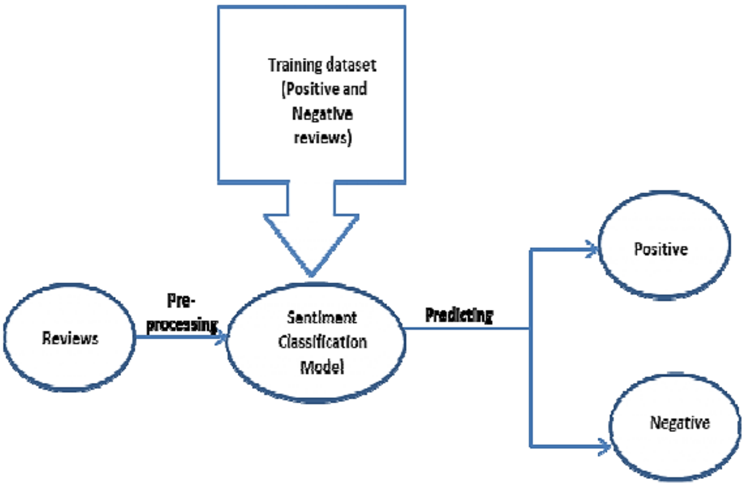
Training dataset (Positive and Negative reviews)
Reviews
Pre-processing
Sentiment Classification Model
Predicting
Positive
Negative

Figure 1. Flow chart of supervised machine learning algorithms

III. DECISION TREE

Like kNN, Decision Tree (DT) has not been considered significantly as a data mining technique in sentiment investigation. However Jia et al (2009) in experiment on candidate score (CS) DT was used to determine the extremity of CS utilizing SVM-Light as the classifier. The technique was use with selected terms as tree and review sets of positive and negative as leaf nodes. The most huge review formed the base of the tree. DT was used as trade-off for profoundly predictive techniques (SVM and Naïve Bayes). Its incorporation in the experiment was worthwhile as it help with understanding predictive performance profile in sentiment examination. C4.5 DT was employed in learning semantic limitations while discovering part-whole relations in text mining. However, it performance was not reported in the survey with respect to whether the technique yielded any useable outcome.

Text Mining

Aspect-rating is numerical evaluations in relation to the aspect highlighting the level of fulfillment portrayed in the comments gathered toward this aspect and the aspect rating. It makes use of diminutive phrases and their modifiers (Liu, 2011), for example 'great product, excellent price'. Each aspect is extracted and collated utilizing Probabilistic latent semantic investigation (pLSA). It tends to be used in place of structure of the phrase. The already realized complete post is exploited to ascertain the aspect rating. Aspect cluster are words that collectively represents an aspect that users are concerned in and would comment on. Latent Aspect Rating Analysis (LARA) approach attempts to analyze sentiment borne by different reviewers by doing a text mining at the purpose of topical aspect. This enables the determinance of every reviewer's latent score on each aspects and the relevant influence on them when coming to an affirmative end result. The revelation of the latent scores on different aspects can quickly continue aspect-base conclusion outline. The aspect influences are relative to breaking down score performance of reviewers. The combination latent scores and aspect influences is capable of continuing personalized aspect-level scoring of entities utilizing only those reviews originated from reviewers with comparable aspect influences to those considered by a specific user. An aspect-based synopsis make use of set of user reviews of a subject as info and creates a set of significant aspects contemplating the combined sentiment of each aspect and supporting textual sign.

IV. CONCLUSION

Data mining techniques has proved effective and useful considering the research carried out so far in the field. This is so because of the limit data mining possess in taking care of uproarious, large and dynamic data. Different creators have come up with several algorithms that can be used to mine the assessments of online users of the SM. Large number of works reviewed significantly utilized Support Vector Machine (SVM), Naive Bayes and Maximum Entropy. Rule Mining, Decision Tree, KNN and Neural Network, these techniques have not gained notoriety as much as SVM, Naïve Bayes and Maximum Entropy. However their reports have been helpful for trade-off interpretability purpose. It is expected that future work will make use of both currently used but to-be-explored data mining techniques to delve deeper into mining the ever increasing online data generated every day on SM.

V. REFERENCES

[1] Sankar K. Pal, et.al "Web Mining in Soft Computing Framework:,IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO.5, SEPTEMBER 2002.

[2] Tomoyuki NANNO, et.al "Automatically Collecting, Monitoring, andMining Japanese Weblogs",

[3] Ralph Gross et.al "Information Revelation and Privacy in Online SocialNetworks ACM Workshop on Privacy in the Electronic Society(WPES), 2005.

[4] Huan Liu et.al "Toward Integrating Feature Selection Algorithms forClassification and Clustering",IEEE Transactions 2005.

[5] Dhuha Khalid Alferidah, NZ Jhanjhi, A Review on Security and Privacy Issues and Challenges in Internet of Things, in International Journal of Computer Science and Network Security IJCSNS, 2020, vol 20, issue 4, pp.263-286

[6] Almusaylim Z, Alhumam A, Mansoor W, Chatterjee P, Jhanjhi NZ. Detection and Mitigation of RPL Rank and Version Number Attacks in Smart Internet of Things. Preprints.org; 2020 https://doi.org/10.20944/preprints202007.0476.v1

[7] Vahini Ezhilraman S., Srinivasan S., **Suseendran G.** (2020) "Gaussian Light Gradient Boost Ensemble Decision Tree Classifier for Breast Cancer Detection", In: Peng SL., Son L., Suseendran G., Balaganesh D. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, Vol 118. Pages 31-38 Url : https://link.springer.com/chapter/10.1007/978-981-15-3284-9_4

[8] Josephine Isabella S., Srinivasan S., **Suseendran G.** (2020) An Efficient Study of Fraud Detection System Using Ml Techniques. In: Peng SL., Son L., Suseendran G., Balaganesh D. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, Vol 118 Pages 59-67.Url : https://link.springer.com/chapter/10.1007/978-981-15-3284-9_8

[9] Andrea Esuli_ et.al, "SENTIWORDNET: A Publicly Available LexicalResource for Opinion Mining", Proceedings of the 5th Conference,2006.

[10] Marcelo Maia, et.al "Identifying User Behavior in Online SocialNetworks", SocialNets'08, April 1, 2008

[11] Ai Ho, AbdouMaiga, et.al "Privacy Protection Issues in SocialNetworking Sites 2009

[12] L. Dey et.al "Opinion mining from noisy text data,"2009

[13] Mohammad Al-Fayoumi, Soumya Banerjee, Jr.,et.al"Analysis of SocialNetwork Using Clever Ant Colony Metaphor", PROCEEDINGS OFWASET 2009

[14] David Ediger Karl Jiang et.al"Massive Social Network Analysis: MiningTwitter for Social Good", ICPP 2010

[15] RojalinaPriyadarshini; "Functional Analysis of Artificial NeuralNetwork for Dataset Classification IJCCT August 2010.

[16] Selman Bozkır1, et.al Identification of User Patterns in Social Networksby Data Mining Techniques: IMCW 2010.

[17] Alexander Pak et.al "Twitter as a Corpus for Sentiment Analysis andOpinion mining", Proceedings of the LREC 2010.

[18] S. Baccianella, et.al enhanced lexical resource for sentiment analysis andopinion mining," in Proceedings of LREC '10 2010

[19] SitaramAsur et.al "Predicting the Future With Social Media" Mar2010.

[20] Kuan-Yu Lin et.al, "Why people use social networking sites: Anempirical study integrating network externalities and motivationtheory", Computers in Human Behavior 27 (2011).

[21] M. Taboada et.al "Lexicon-based methods for sentiment analysis .IEEE2011

[22] E.Raju K.Sravanthi, "Analysis of Social Networks Using the Techniquesof Web Mining" IJARCCSE Volume 2, Issue 10, October 2012.

[23] K. Vengatesan, A. Kumar, R. Naik and D. K. Verma, "Anomaly Based Novel Intrusion Detection System For Network Traffic Reduction," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Palladam, India, 2018, pp. 688-690.