International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 9, September-2018

Improved Opinion Mining System to Enhance Accuracy using SVM with RBF Kernel

[1]Umakant Sharma, [2]Abhilash Mishra

[1]Dept. of Software Engineering, [2]Asst. Prof. Dept. of CSE/IT
NITM College, Gwalior, India

***Abstract*— Opinions in social media platforms provide worldwide access to what people think about daily life topics/issues. Thus exploiting such a source of data to comprehend popular opinion can be exceptionally valuable in numerous situations, for instance by political gatherings keen on observing attitudes towards their policies. The opinion mining research field aims to develop automated approaches to accurately analyze such opinion data. Although there is much previous work on opinion mining, the majority of early studies analyzed text documents such as product and movie reviews. Only a limited number of studies have attempted to analyze public opinion in a political context. Twitter, which is now and again called a buy in and- publish social network, gives coordinated connections among users' and hosts feeling rich data over a wide arrangement of users' and points. In this manner, it is clear that mining client opinions and estimations from Twitter will be exceptionally valuable for some applications.**

Keywords*— *opinion mining; model; application; SVM; RBF kernel.

I. INTRODUCTION

Opinions in social media platforms provide worldwide access to what people think about daily life topics/issues. In this manner misusing such a wellspring of data to comprehend general opinion can be extremely helpful in numerous situations, for instance by political gatherings keen on checking attitudes towards their policies. The opinion mining research field aims to develop automated approaches to accurately analyse such opinion data. Although there is much previous work on opinion mining, the majority of early studies analysed text documents such as product and movie reviews. Only a limited number of studies have attempted to analyse public opinion in a political context. The nature of political discourse which often includes sarcasm and irony makes the analysis more challenging. Most political opinion mining studies analysed election datasets with the aim of election result prediction. The approaches employed range from lexicon based methods to supervised machine learning methods using algorithms such as SVM and Deep Learning approaches in recent years. However, most work in this area focuses on overall sentiment classification, though Maynard and Funk considered identification of the opinion target and Vijayaraghavan et al. addressed classification of the topics along with the sentiment. An opinion is defined as a tuple of four components sentiment orientation, sentiment target, opinion holder, and time. The sentiment target is defined as an entity or an entity with possibly an aspect of the entity that the sentiment has been expressed upon. As an example, if a tweet expresses a negative sentiment towards the Labor Party regarding Employment/Jobs, Labor Party is the target entity and Employment/Jobs is the part of the element. Be that as it may, the dominant part of early examinations in opinion mining endeavored to recognize just the general notion, paying little respect to the elements said and their viewpoints. After Hu and Liu's examination, fine-grained opinion mining/viewpoint based opinion mining ended up noticeable. However, most of the aspect-based opinion mining studies identified only aspect and sentiment by assuming a known target.[1].

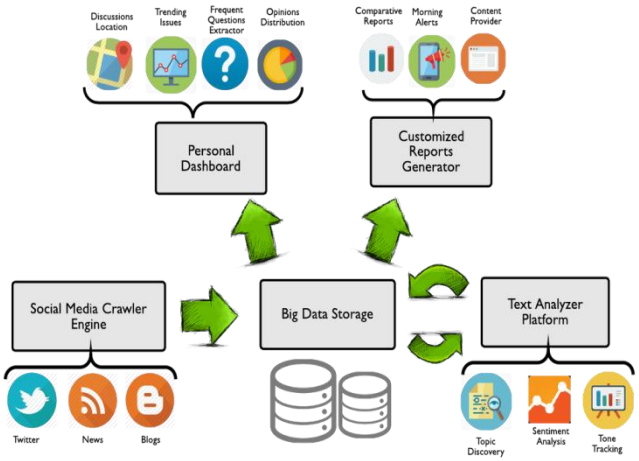

Fig.1 Opinion Mining

1.1 OPINION MINING MODELS:

In this segment, we portray two models that accomplished best in class exhibitions in opinion mining. The first model won the "Sentiment Analysis in Twitter", and uses feature engineering to train an opinion classifier. The second model depends on displaying compositionality utilizing profound learning systems. We assess these models for opinion mining in Arabic tweets.[2]

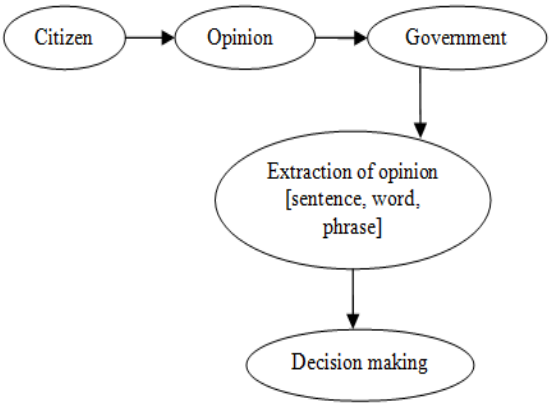

Fig. 2 Opinion Mining Model

1. Opinion Mining with Feature:

Engineering According to, training a SVM with a collection of surface, syntactic, semantic features achieved state-of-the-art results on opinion mining in tweets. Beneath, we depict the proportional arrangement of highlights that we removed to prepare a comparable model for opinion mining in Arabic tweets.[2]

- Character n-grams; n 2 [3, 5].
- Word n-grams; n 2 [1, 4]. To represent the many-sided quality and sparsity of Arabic dialect, we separate lemma n-grams since lemmas have preferred speculation capacities over crude words.
- Counts of exclamation script, question marks, and both exclamation and question marks.
- Count of elongated words.
- Count of negated contexts, defined by phrases that occur between a negation particle and the next punctuation.
- Counts of positive emoticons and negative emoticons notwithstanding a parallel component demonstrating if emojis exist in a given tweet.
- Counts of each part-of-speech (POS) mark in the tweet.
- Counts of positive and negative words founded on ArSenL, AraSenti and ADHL lexicons.

2. Opinion Mining with Recursive Neural Networks:

Most profound learning models for opinion mining depend on the idea of compositionality, where the importance of a content can be portrayed as a component of the implications of its parts and the principles by which they are consolidated. Specifically, the Recursive Neural Tensor Networks (RNTN) demonstrate has demonstrated fruitful for opinion mining. RNTN to anticipate the notion of a threeword sentence, where words are spoken to with vectors that catch distributional syntactic and semantic properties. Preparing a RNTN show requires an assessment treebank; a gathering of parse trees with assumption explanations at all levels of body electorate. For English, the Stanford slant treebank was produced to prepare the RNTN.[2]

1.2. OPINION MINING APPLICATION:

Opinion mining and sentiment analysis cover an extensive variety of uses.

1. Argument mapping software helps sorting out legitimately these strategy articulations, by express ting the consistent connections between them. Under the examination field of Online Deliberation, instruments like Compendium, Debatepedia, Cohere, Debate chart have been created to give a sensible structure to various strategy proclamation and to interface contentions with the proof to back it up.

2. Voting Advise Applications help voters understanding which political gathering (or different voters) has nearer positions to theirs. For example, SmartVote.ch requests that the voter announce its level of concurrence with various strategy articulations, and afterward coordinates its situation with the political gatherings.

3. Atomated content examination helps processing vast measure of subjective information. There are today available numerous instruments that consolidate factual calculation with semantics and ontologies, and additionally machine learning with human supervision. These arrangements can recognize pertinent remarks and relegate positive or negative implications to it (the alleged assessment). [4]

II. LITERATURE SURVEY

Mika V. Mäntylä et al [2018] as of late, slant examination has moved from breaking down online item surveys to web based life writings from Twitter and Facebook. Numerous themes past item surveys like securities exchanges, races, fiascos, solution, programming building and digital tormenting broaden the usage of feeling analysis1.[5]

Deruo Cheng et al [2018] This paper proposes a hybrid K-means clustering and support vector machine (HKCSVM) approach to recognize the places of vias and metal lines from delayered IC pictures for resulting netlist extraction. The fundamental commitments of the proposed strategy include: 1) completely mechanized discovery of by means of and metal line positions with no need of human mediations; and 2) novel half breed philosophy to epitomize K-means clustering and bolster vector machine for recovering exact places of vias and metal lines as opposed to the person techniques which can only provide a region for the detected elements.[6]

Yuan-Hai Shao ET AL [2018] in the paper, we propose a sparse $L\ q$ -norm least squares support vector machine ($L\ q$ -norm LS-SVM) with $0 < q < 1$, where feature selection and prediction are performed simultaneously. Different from traditional LS-SVM, our $L\ q$ -norm LS-SVM minimizes the $L\ q$ -norm of weight and releases the least squares problem in primal space. The effectiveness of the proposed $L\ q$ -norm LS-SVM is validated via theoretical analysis as well as some illustrative numerical experiments.[7]

Heng-Li Yang et al [2018] In this investigation, we concentrated on audits in view of profoundly feeling inserted items/administrations, for example, motion pictures, music, and dramatization. Fur- thermore, we tried to solve the multiple polarities problem for the same review word for multiple types of product/service. First, we collected text written in Chinese from a Taiwanese movie forum. In our proposed approach, we applied an evolutionary strategy algorithm to optimize the weight tables corre- sponding to two different types of movies: horror and drama movies.[8]

Gaurav dubey et al [2017] this paper applies content mining on tweets created on Twitter locales for two well known Indian political diploments: Arvind Kejriwal and Narendra Modi. This study could really help these diplomats to improve their political strategies.[9]

Kai Yang et al [2017] In this paper, we build the space slant word reference utilizing outside printed information. Additionally, numerous order models can be utilized to arrange records as indicated by their opinion. In any case, these single models have qualities and shortcomings. We propose a profoundly successful cross breed demonstrate consolidating distinctive single models to defeat their weaknesses.[10]

[17] ALAA M. El-HALEES et al [2017] In this study, we used distributed representations for Arabic opinion mining and compare it with Bag of Words (BOW) representation. We applied them on four benchmark datasets. Then, we used four machine learning methods which are Support Vector Machine, Logistic Regression and Random Forest. Using f-measure metric, we found that, in all datasets and all methods we used in our experiment, the distributed representations have better performance than bag-of- words representation.[11]

Sana Parveen et al [2017] In this work we propose a strategy to identify mockery in Twitter that makes utilization of the diverse parts of the tweet. Work proposes four classifications of highlights that cover diverse sorts of mockery we characterized, and that will be utilized to arrange tweets into snide and non- sarcastic. To evaluate the performances of our work study the importance of each of the proposed sets of features and evaluate its added value to the classification.[12]

Rashid Kamal et al [2017] This paper presents a framework to visualize raw tweets in a scalable and optimal fashion. The main objective of the research work is to get sentiment of the people and visualize it for better understanding. Spring XD has been used to fetch tweets on a real time basis. These raw tweets are then transformed to Hadoop Distributed File System (HDFS). Hadoop Scripting Language (HIVE) is used to refine and label the tweets for their respective sentiments. Finally, these sentiments are classified as positive, negative and neutral using an algorithm which is simulated over HIVE.[13]

Bakhtiar Feizizadeh et al [2017] This investigation thinks about the prescient execution of GIS-based landslide susceptibility mapping (LSM) utilizing four diverse portion works in SVMs. Nine conceivable causal criteria were viewed as in view of prior comparative examinations for a zone in the eastern piece of the Khuzestan area of southern Iran. Distinctive models and the subsequent avalanche defenselessness maps were made utilizing data on known avalanche occasions from an avalanche stock dataset. The models were prepared utilizing avalanche stock dataset. A two-step accuracy assessment was implemented to validate the results and to compare the capability of each function.[14]

III. PROPOSE MTHODOLOGY

LEAST SQUARES

The technique for least squares is a standard way to deal with the estimated arrangement of over decided issues, i.e., the quantity of conditions in which there are a larger number of conditions than questions. The best fit at all squares sense limits the whole of squared residuals, a lingering being the distinction between a watched esteem and the fitted esteem given by a model. In particular, a simple data set comprises of m focuses (information pairs) (xi ,yi) I = 1, ..., m where xi is an autonomous variable and yi is a reliant variable whose esteem is found by perception. The model capacity has the shape f(x, β), where the n movable parameters are held in the vector β. The objective is to discover the parameter esteems for the model which "best" fits the information. The slightest squares strategy discovers its ideal when the whole E(β) of squared residuals

$$E(\beta) = \sum_{i=1}^{m} s_i^2 \text{ is a minimum}$$

where si = yi − f(xi , β) is the residual (or bias) between the actual value of the dependent variable and the value predicted by the model.

Proposed work:

In the proposed work, different websites are used to collect the tweets from the various blogging websites and twitter. Various details are contained in the data such as name of citizen, their age and profession. The data has been trained in the form of

< Category, aspect, opinion word, sentiment, political interest government employee>

For example for the below tweets-

“narendra modi took a great decision on black money”

Category->decision, aspect->black money, opinion word->great, sentiment->positive, political interest->yes, government employee->yes

After gathering all tweets from various sites, our following stage is sorted opinion in various citizen classes. We are classified opinions in view of citizen’s profession. Explanation for perform classification of opinion is-If government needs to settle on choice for particular gathering than that of gathering opinion of citizen are most critical for solid making of decision.

When we arranged all opinion our subsequent stage is preprocessing because of the fact that data removed from various sites is in raw form that comprise distinctive abnormalities in information. Keeping in mind the end goal to utilize additionally preparing, we have to clean and change it in more usable organized type of information. Tweet cleaning is the initial move towards information change. Subsequent to cleaning the information, the collection of tweets is then parsed.

For instance words like 'a', 'the' won't not be so noteworthy. These words are known as Stop Words, and are eliminated during pre-processing. Likewise, it doesn't make a difference whether it is 'awesome' or 'Awesome', so all the content can be conveyed to bring lowercase. Behind cleaning the information, the collection of tweets is then parsed. After information change we have organized information which is prepared for additionally handling. After information change we prepare organized information which is for additionally handling.

SVM is correctly work on linear Separable elements in which all data points are plotted in n-dimensional space or plane where n is the sum of all features and we have to select correct decision boundary that classify all data point in different classes. Major problem with SVM is it can't works in high dimensional feature space because large number of feature can't be classified linearly. To solve this problem we use RBF kernel function. It solves the problem that occurs in linear separable opinion classification. The reason behind using radial basis function is increase the accuracy of SVM algorithm. RBF is an artificial neural network function which is used for classification. SVM (RBF) function assign class label to each opinion.

In subsequent step of preprocessing of all the opinion, we perform classification on every group. For classification, we utilize SVM with Least Square Method (LSM). SVM give better outcome regarding complexity and accuracy than the other classification techniques. SVM is effectively take a shot at straight Separable components in which all information focuses are plotted in n-dimensional space or plane where n is the whole of all highlights and we need to choose adjust choice limit that arrange all information point in various classes. Real issue with SVM is it can't works in high dimensional element space since substantial number of highlight can't be characterized directly. To take care of this issue we utilize RBF portion work. It tackles the issue that happens in straight distinct feeling order. The purpose for utilizing outspread premise work is increment the precision of SVM calculation. RBF is a manufactured neural system work which is utilized for order. SVM (RBF) work dole out class mark to every conclusion.

In our approach sentiment label assignment is done in five categories. It classifies opinion in positive, negative, neutral, strong positive and strong negative. We are define all classes below-

- Positive-simple positive views {I love this decision very much}
- Negative- simple negative views {I don't agree with government decision}
- Neutral-not clear views {hop for best}
- Strong positive-positive view with positive reason { This move will force PeoPle to take this money to banks if they want to keeP it, so this is right decision}
- Strong negative-negative view with valid reason {There are many who get their salaries in cash and do not have bank accounts so what they people do}

IV. RESULT ANALYSIS

In previous opinion classification method all opinion are classified by applying support vector machine but in our approach we are using SVM with RBF kernel capacity to enhance the consequence of content arrangement.

The major drawback has overcome, SVM performs better on limited opinions because we already categorized all citizen sentiments in different categories and we are applying SVM independently on each category. For analysis we had collected citizen reviews and tweets on black money or demonetization decision. Here we will compare classification results of previous approach and our proposed approach.

Base work:

accuracy = 44.79%

Confusion Matrix:

131 20 13 129 7

151 117 11 39 0

128 37 327 302 70

0 11 4 198 3

53 0 11 12 39

Proposed work:

accuracy = 61.45%

Confusion Matrix:

101 14 17 126 42

27 173 12 36 70

35 48 555 109 117

0 0 7 209 0

2 7 15 15 76

Accuracy= sum of all diagonal elements/total elements in matrix

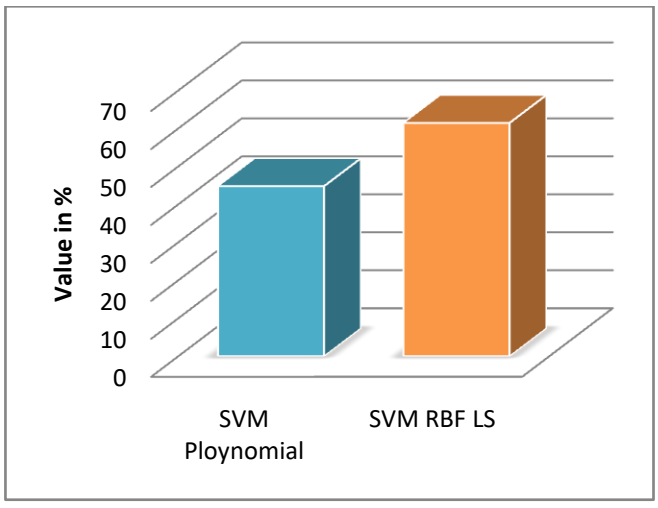

Fig. 3 Accuracy

Conclusion

In this proposed approach data analysis approach is presented in which we extract group of opinions from different micro blogging web sites like twitter, political blogs that are used to make decision for any object. SVM apply with RBF function to increase result accuracy than previous opinion classification algorithms. In our presented approach each group of citizen shows different opinion on any government decision or policies. Because it is important that if government wants to make any decision for specific group of citizen then that group of citizens opinion are very important for decision making. Aim of presented approach is to make strong decision for development of citizen and government empowerment.

References

[1] Sandeepa Kannangara, "Mining Twitter for Fine-Grained Political Opinion Polarity Classification, Ideology Detection and Sarcasm Detection", WSDM'18, February 5-9, 2018, Marina Del Rey, CA, USA.

[2] Ramy Baly, Gilbert Badaro, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Wassim El-Hajj, Nizar Habash, Khaled Bashir Shaban, "A Characterization Study of Arabic Twitter Data with a Benchmarking for State-of-the-Art Opinion Mining Models", Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 110–118, Valencia, Spain, April 3, 2017.

[3] Mário J. Silva, Paula Carvalho, Luís Sarmento, Pedro Magalhães and Eugénio Oliveira, "The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics", 22 May 2014.

[4] David Osimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis",2010.

[5] Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers", Volume 27,February 2018.

[6] Deruo Cheng, Yiqiong Shi, Tong Lin, Bah-Hwee Gwee, Kar-Ann Toh, "Hybrid *K*-Means Clustering and Support Vector Machine Method for Via and Metal Line Detections in Delayered IC Images",IEEE 2018.

[7] Yuan-Hai Shao, Chun-Na Li, Ming-Zeng Liu, Zhen Wang , Nai-Yang Deng, " Sparse $L\ q$ -norm least squares support vector machine with feature selection", 2018 Elsevier Ltd. All rights reserved.

[8] Heng-Li Yang , Qing-Feng Lin, " Opinion mining for multiple types of emotion emb e dde d products/services through evolutionary strategy", Expert Systems With Applications,2018.

[9] Gaurav dubey,shilpi chawla,kirandeep kaur, "social media opinion analysis for indian political deploments", 2017 IEEE.

[10] Kai Yang, Yi Cai, Dongping Huang, Jingnan Li, Zikai Zhou, Xue Lei, "An Effective Hybrid Model for Opinion Mining and Sentiment Analysis", 2017 IEEE.

[11] ALAA M. El-HALEES, "Arabic Opinion Mining Using Distributed Representations of Documents", 978-1-5090-6538-7/17 $31.00 © 2017 IEEE.

[12] Sana Parveen, Sachin N. Deshmukh, "Opinion Mining in Twitter – Sarcasm Detection", International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 10 | Oct -2017.

[13] Rashid Kamal, Munam Ali Shah, Asad Hanif, J Ahmad, "Real-time Opinion Mining of Twitter Data using Spring XD and Hadoop", Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017.

[14] Bakhtiar Feizizadeh & Majid Shadman Roodposhti & Thomas Blaschke3 & Jagannath Aryal, "Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping", Saudi Society for Geosciences 2017.

.