Novel Approach for Sentiment Classification by Ontology with Swarm Intelligence Optimization

Aditya Bhardwaj[1], Uma Tyagi[2]
[1]*M.Tech, Dept. of Computer Science And Engineering, Meerut Institute of Engineering And Technology, Meerut, India*
[2]*Asst. Prof., Dept. of Computer Science And Engineering, Meerut Institute of Engineering And Technology, Meerut, India*

Abstract – Nowadays, the faster message and ease of use in social media today is Twitter. The messages on Twitter include reviews and sentiment on certain topics such as movie, book, product, politic etc. Based on this condition, this research attempts to use the messages of twitter to review a movie by using sentiment analysis. Sentiment analysis refers to the application of natural language processing, computational linguistics, and text mining to identify or classify whether the movie is good or not based on message opinion. Naïve Bayes is supervised learning method that analyzes data and recognizes the patterns that are used for classification. This research concerns on binary classification which is classified into two classes. Those classes are positive and negative. The hybrid Ant Colony Optimization (ACO) with Naïve Bayes is used to improve the election of best parameter in order to solve the dual optimization problem. The result shows the improvement of accuracy level from 70.

Keywords *- ACO, Sentiment analysis, PSO, Microblogging.*

I. INTRODUCTION

Miniaturized scale blogging is on a fundamental level in view of blogs (i.e. Web logs), where clients can post assessments, encounters and questions on any picked theme. The principle contrast amongst small scale and customary blogs is the strict imperative in content size. Right now, the most prominent on-line small scale blogging administration is Twitter, which empowers its clients to send and get content based posts, known as "tweets", comprising of up to 140 characters [1]. Twitter was made in 2006 and as of now records more than 140 million dynamic clients that create more than 340 million tweets for every day. With their quickly expanding fame, small scale blogging administrations, similar to Twitter, have developed into a reasonable means for imparting insights on all parts of regular day to day existence. The strict character point of confinement of tweets strengths clients to be brief and more expressive than with informal communities and blogs in the end. Consequently, smaller scale blogging posts are permeated with enthusiastic data and are considered as rich feeling mining information sources. Moreover, tweets can be handled more successfully than protracted blog entries and articles [5].

The rise of Web 2.0 has definitely modified the way clients see the Internet, by enhancing data sharing, cooperation and interoperability. In opposition to the original of sites, where clients could latently see content, Web 2.0 clients are urged to partake and team up, shaping virtual on-line groups. Extra Web 2.0 characteristics incorporate information transparency and metadata, dynamic substance, rich client experience and versatility resistance. Among the most well-known Web 2.0 applications, one can go over interpersonal interaction destinations (e.g. Facebook, MySpace), wikis, blogs, multi-media sharing destinations (e.g. YouTube, Flickr), concoction and rich web applications [5]. One of these exercises is small scale blogging, which at first pulled in similarly less consideration, however bit by bit turned into an exceptionally well known specialized device for an impressive level of clients.

The ubiquity of miniaturized scale blogging comes from its unmistakable correspondence administrations, for example, conveyability, quickness, and usability, which enable clients to in a split second react and spread data with constrained or no limitations on content. Twitter is right now the most well known and quickest developing microblogging administration, with more than 140 million clients creating more than 400 million tweets for every day—for the most part versatile. Twitter empowers clients to post announcements, or tweets, no longer than 140 characters to a network of devotees utilizing different correspondence administrations.

Tweets have detailed everything from day by day biographies to most recent nearby and overall occasions. Twitter content reflects ongoing occasions throughout our life and contains rich social data and worldly characteristics. Checking and breaking down this rich and constant stream of client created substance can yield remarkably significant data.

Online web-based social networking locales (Facebook, Twitter, YouTube, and so forth.) have reformed the way we speak with people, gatherings, and groups and modified regular practices. A few late workshops, for example, Semantic Analysis in Social Media, are progressively concentrating on the effect of online networking on our everyday lives. For example, Twitter has changed the way individuals and organizations perform, look for counsel, and make "encompassing mindfulness" and strengthened the frail and solid tie of companionship. Not at all like other media sources, Twitter messages give opportune and fine-grained data about any sort of occasion, reflecting, for example, individual viewpoints, social data, conversational perspectives, enthusiastic responses, and questionable assessments. A noteworthy test confronting occasion identification from Twitter streams is in this way to isolate the ordinary and assessment data from fascinating true occasions. By and by, exceedingly versatile and effective methodologies are required for dealing with and preparing the undeniably

substantial measure of Twitter information (particularly for constant occasion identification). Different difficulties are natural to Twitter's outline and use. These are basically because of the short length of tweet messages, the incessant utilization of (progressively developing) casual, unpredictable, and abridged words, the substantial number of spelling and syntactic mistakes, and the utilization of dishonorable sentence structure and blended dialects.
Machine learning is a strategy for information investigation that computerizes systematic model building. Utilizing calculations that iteratively gain from information, machine learning enables PCs to discover shrouded bits of knowledge without being unequivocally modified where to look. In software engineering and operations inquire about; the Ant Colony Optimization (ACO) is a probabilistic strategy for tackling computational issues which can be lessened to discovering great ways through diagrams. This calculation is an individual from the subterranean insect settlement calculations family, in swarm knowledge techniques, and it constitutes some met heuristic advancements. Naive Bayes classifiers are exceptionally versatile, requiring various parameters direct in the quantity of factors (highlights/indicators) in a learning issue. Most extreme probability preparing should be possible by assessing a shut shape articulation, which takes direct time, instead of by costly iterative estimation as utilized for some different sorts of classifiers.

II. LITERATURE REVIEW

In [1], proposed an ontology outline towards semantification arrangement of Twitter social examination. The ontology is sent over a freely accessible administration that measures how persuasive a Twitter account is, by consolidating its social action and communication over Twitter circle. Aside from persuasive amount and quality measures, the administration gives a SPARQL endpoint where clients can perform progress semantic inquiries through the RDFized Twitter elements (notices, answers, hashtags, photographs, URLs) over the semantic diagram.

In [2], A versatile and mechanized procedure is presented for client profiling by removing his URLs from openly accessible tweets data and utilizing a semantic ontology in which client interests and expectations are portrayed. Keeping in mind the end goal to improve the execution of our strategy, classification of sites offered by OpenDNS and DBpedia aggregate knowledge databases are utilized to discover the interests and goal classifications of the client profile ontology. Test comes about in light of client's tweets affirm firmly that the proposed strategy enhances the programmed securing of interests and aims of a client profile.

In [3], an ontology based portrayal has been proposed for speaking to the API of the well known SNS Google+. The API of Twitter SNS and make an ontology based portrayal of its auxiliary and useful properties is considered. The proposed Twitter REST API ontology reuses classes of the current Google+ API ontology and portrays important auxiliary and useful points of interest of the API, in a machine processable arrangement valuable for understanding the API and proper for coordinating into ontology based Mashups.

In [4], they exhibited their endeavor in bringing the Semantic Web Knowledge Management worldview at the accessibility of various individual desktop instruments (Mail customers, Web Browser, Agenda and so forth ...), by developing Web Browser Semantic extension Semantic Turkey to an extensible system giving RDF information access at various levels: java access through OSGi extensions, HTTP get to or devoted JavaScript API for the entire scope of devices from the open source suite of Mozilla applications.

In [5] the organization of unique ontology-based strategies is proposed towards a more effective slant investigation of Twitter posts. The oddity of the proposed approach is that posts are not just portrayed by a conclusion score, similar to the case with machine learning-based classifiers, yet rather get an assessment review for each unmistakable idea in the post. Generally, this proposed design brings about a more point by point examination of post assessments with respect to a particular theme.

In [6], recommended that knowledge about basic and behavioural properties of programming can be shared over the product building group as configuration designs communicated in the web ontology language (OWL). The intrinsic preferred standpoint of their approach is that it yields depictions that are machine process-capable, yet in addition reasonable for a group to share knowledge exploiting the decentralized foundation of the Internet.

In [7], they presented a model that uses a world ontology developed from the Dewey decimal classification to gain client profiles. Via seek utilizing particular and thorough client profiles, data combination methods never again depend on the insights gave by specialists. The model has been effectively assessed utilizing the huge INEX informational index reproducing the dispersed Web condition.

In [8], presents a framework for quickening the ontology building process by means of semi-consequently taking in a hierarchal ontology given an arrangement of space particular web reports and an arrangement of seed ideas. The techniques are tried with Web records in the space of Agriculture. The ontology is built using two integral methodologies. The principal approach uses the structure of expressions showing up in HTML headings while the second uses the various levelled structure of the HTML headings for recognizing new ideas and their taxonomical connection between seed ideas and between each other. The exhibited framework has been utilized to assemble ontology in the agrarian space utilizing an arrangement of Arabic extension archives. The subsequent ontology was assessed against an adjusted rendition of the AGROVOC ontology.

III. PROPOSED ALGORITHM

A. Ant Colony Optimization

Ant colony optimization is fundamentally roused by the

genuine ant settlements conduct and called artificial framework. Through the charts the Ant colony optimization calculation (ACO) is utilized for the taking care of computational problems and discovering great way. Like ant conduct, looking for way between food source and their colony to look through an ideal way comparative is the principle point of this calculation. To take care of the problem of travelling salesman problem (TSP) the principal ACO was created. Prior to the pheromones are refreshed along their food source trail on change probability bases a probability decision is made in the standard ACO. Before refreshing the pheromones along their trail to a food source in the standard ACO, which depends on the progress probability, ants settles on a probabilistic decision. For the k^{th} ant the change probability at the time step t from city x to city y in the TSP problem:

$$PROB_{xy}^{k}(t) = \begin{cases} \frac{[\tau_{xy}(T)]^{\alpha}.[\eta_{xy}]^{\beta}}{\sum_{y\in I_x^k}[\tau_{xy}(T)]^{\alpha}.[\eta_{xy}]^{\beta}} \text{ if } j \epsilon I_x^k \\ 0 \qquad\qquad Otherwise \end{cases}$$

Where

η_{xy}←priority heuristic information,

τ_{xy}←pheromones trail amount on the edge (x, y)at the time T,

The pheromone trail and heuristic information relative effects are identified by two factors i.e., $\alpha$and β. And the city's neighborhood set that are reasonable is denoted by I_x^k.

After a visit is finished by every ant, a constant dissipation rate at first bringing down them which refreshed the pheromone trail. Inferable from which every ant is permitted effective pheromone affidavit on curves which is its visit part as appeared in the condition underneath:

$$\tau_{xy} = (1-\rho).\tau_{xy} + \sum_{k=1}^{N} \Delta\tau_{yx}^{k}$$

Where

ρ←pheromones rate of trail evaporation,

N←no. of ants,

The pheromone trail that is boundless aggregated is averted by the utilization of parameter ρ which empowers the awful choices to be overlooked by the calculation. The no. of cycles declining the pheromone quality related on circular segments which ants don't choose. $\Delta\tau_{yx}^{k}$, the trail substance quality per unit length which lays nervous (y, x) is given as takes after:

$$\Delta\tau_{yx}^{k} = \begin{cases} \frac{Q}{L_k} \; if \; ant \; k \; in \; its \; tour \; uses \; edge \; (y,x) \\ 0 \qquad\qquad Otherwise \end{cases}$$

Where

Q←constant that is predefined,

L_k←length of the tour.

B. *Naïve Bayes Classifier*

Naïve Bayes classifier is the one among the group of probabilistic classifiers in machine learning, between features which is generally based on naïve independence assumption with applied Bayes' theorem. Number of features is required to be parameter linear in learning problem where Naïve Bayes classifier is highly scalable.

C. *Naïve Bayes with ACO*

Ant Colony Optimization technique and Naïve Bayes Classifier combination introducing a new novel approach.
In this approach we use ACO for the optimization purpose and Naïve Bayes for the classification of the features.

NB_ACO ALGORITHM

Step 1: Initializing ants, where for each ant_n, n=1,2,3……..N.

Step 2: In ant_n, each variable $\boldsymbol{a_n^d}$, d=1,2,3…….D.

Step 3: Updating pheromones by choosing $\boldsymbol{\mu_x^d}$ from the pheromone
table with probability, where i∈{1,2,3……K}.

Step 4: If minimum error is obtained, then it has higher probability.

Step 5: Generating a standard deviation $\boldsymbol{\sigma_x^d}$, if rv ≤ a_1 with the use of uniform distribution U (0,1), where rv is the random value lies between x_1, the predefined threshold 0 and 1.

Step 6: Generating a new value for variable $\boldsymbol{a_n^d}$: if rv ≤ a_2, by normal distribution N ($\boldsymbol{\mu_x^d}$,$\boldsymbol{\sigma_x^d}$).

Step 7: Else, uniform distribution generates random value, and generating random solution for $\boldsymbol{a_n^d}$.

Step 8: Obtained variable $\boldsymbol{a_n^d}$ denoting the observed attribute values to certain class label c.
Step 9: Computing probability for each class:

$$\boldsymbol{p(b_y|a_n^d)} = \frac{p(b_y)p(\boldsymbol{a_n^d}|\boldsymbol{b_y})}{\sum_{i=1}^{c} p(b_x)p(\boldsymbol{x_n^d}|\boldsymbol{b_x})}, \text{ y = 1,2……..c}$$

Where

$p(b_x)$is the y_i prior probability,

$p(x_n^d | b_y)$←conditional class probability density function.

Step 10: Calculate probability distribution over the set of features:

$$p(a) = \prod_{i=1}^{k} p(c_x)p(a_n^d | c_x)$$

Where

k←number of classes,

c_x←x^{th} class.

Step 11: Calculate precision, accuracy, and recall.

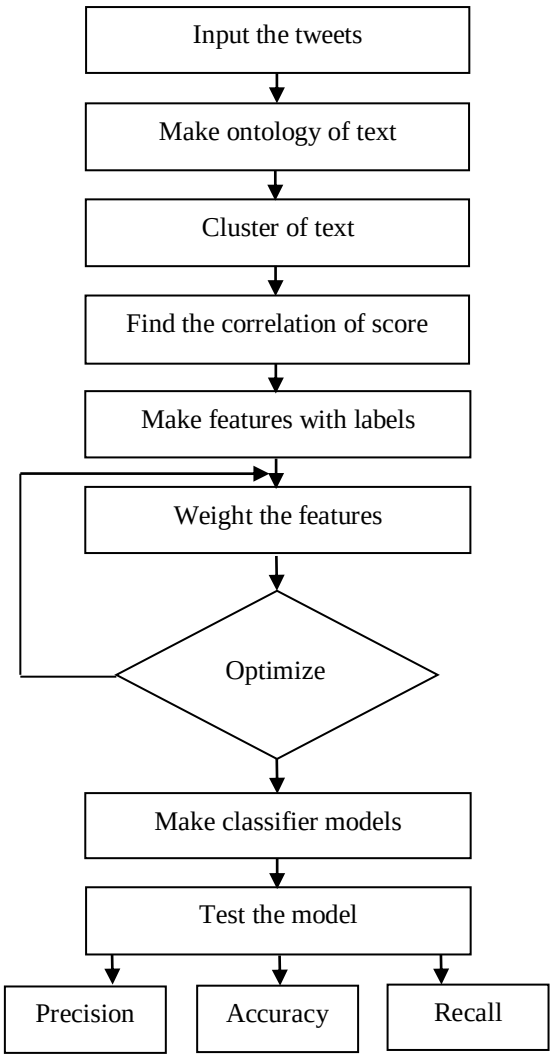

Figure 1: System model.

IV. EXPERIMENTAL RESULTS

Table 1: Comparison table of various parameters (accuracy, precision and recall) in case of given approaches.

Algorithms /parameters	Accuracy	Precision	Recall
ANN	67.17	67.667	67.667
SVM	62.62	51	81
ACO-Naïve Bayes	70.303	86.333	58.555

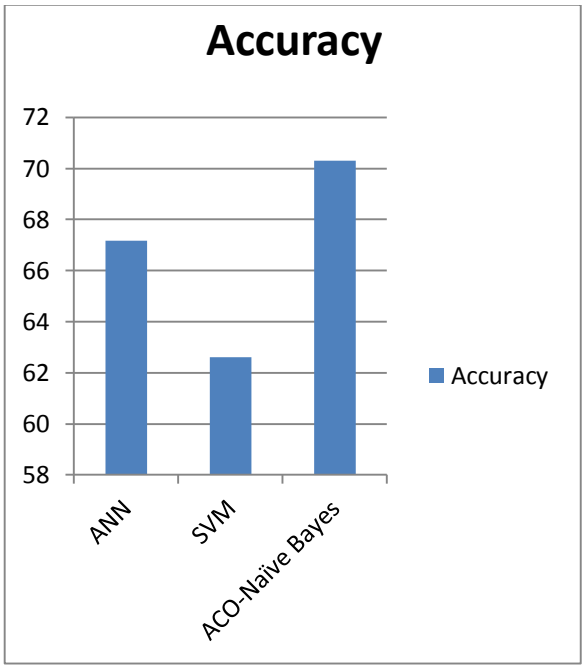

Figure 2: Performance comparison of various algorithms (ANN, SVM and ACO-Naïve Bayes) based on accuracy.

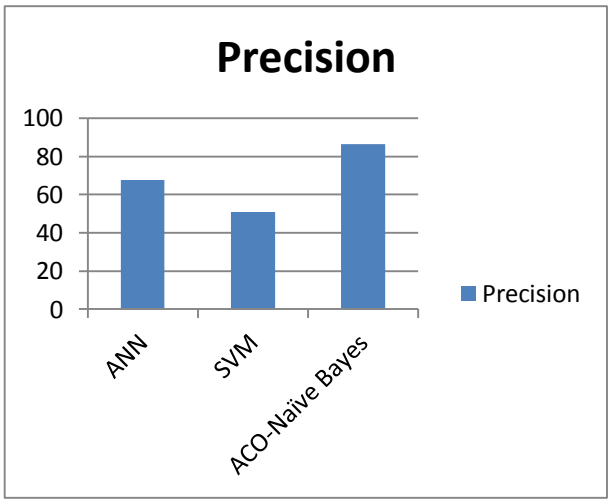

Figure 3: Performance comparison of various algorithms (ANN, SVM and ACO-Naïve Bayes) based on precision.

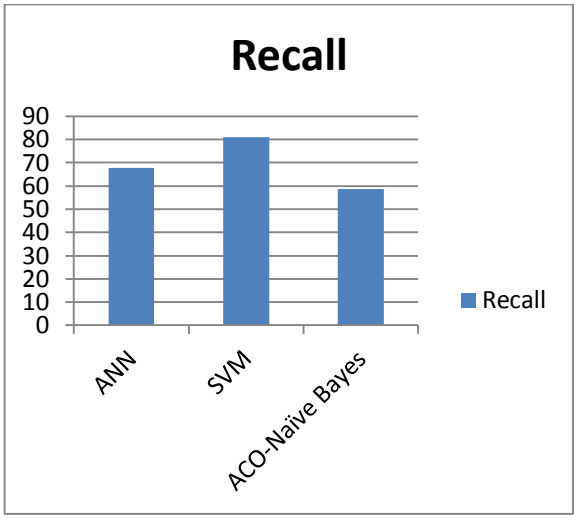

Figure 4: Performance comparison of various algorithms (ANN, SVM and ACO-Naïve Bayes) based on recall.

V. CONCLUSION

This study has shown that ACO affect the accuracy of Naïve Bayes after the hybridization of ACO-Naïve Bayes. The best accuracy level that gives in this study is 70% and has been achieved by ACO-Naïve Bayes after data cleansing. On the other hand, the accuracy level of ACO-Naïve Bayes still can be improved using enhancements of Naïve Bayes that might be using another combination or variation of Naïve Bayes with other optimization method. The ACO-Naïve Bayes without cleansing achieved Accuracy 70, Precision 86.33 and Recall 58.55.

VI. REFERENCES

[1]. Carbon, Seth Ami-GO, "Online access to ontology and annotation data", International Journal of Computer Science and Information Technology & Security, Vol. 2, No. 1, 2012.

[2]. Go, Alec, Lei Huang, and Richa Bhayani, "Twitter sentiment analysis, Entropy", International Journal of Computer Applications, Vol. 139, No.11, pp. 17-21, April 2009.

[3]. Kontopoulos, Efstratios, "Ontology-based sentiment analysis of twitter posts, Expert systems with applications", International Journal of Engineering Research and Applications, Vol. 4, Issue 2, pp.69-74, February 2014.

[4]. Pazienza, Maria Teresa, "A Desktop-Integrated Semantic Platform for Personal Information Management SPIM", Vol. 1, 2010.

[5]. Penas, Paula, "Collective Knowledge Ontology User Profiling for Twitter--Automatic User Profiling, Web Intelligence (WI) and Intelligent Agent Technologies (IAT)", IEEE/WIC/ACM International Joint Conferences on. Vol. 1,2013.

[6]. Razis, Gerasimos, and Ioannis Anagnostopoulos, "Semantifying twitter: The influence tracker ontology, Semantic and Social Media Adaptation and Personalization (SMAP)", 9th International Workshop on IEEE, 2014.

[7]. Tao, Xiaohui, "Ontology-based specific and exhaustive user profiles for constraint information fusion for multi-agents, Web Intelligence and Intelligent Agent Technology (WI-IAT)", IEEE/WIC/ACM International Conference on. Vol. 1, 2010.

[8]. Togias, Konstantinos, and Achilles Kameas, "An ontology-based representation of the Twitter REST API", IEEE 24th International Conference on Tools with Artificial Intelligence, 2012.

[9]. T. R. Gruber, 1995, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", In International Journal of Human-Computer Studies, Vol 43 Issue 5-6, pp. 907-928, Nov./Dec. 1995.

[10]. A.Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp.1320-1326, 2010.

[11]. R. Parikh and M. Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009.