International Journal of Computer Science Engineering & Technology
APPROVED BY
NATIONAL SCIENCE LIBRARY (NSL)
NATIONAL INSTITUTE OF SCIENCE COMMUNICATION AND INFORMATION RESOURCES (NISCAIR)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR)– NEW DELHI INDIA .
ISSN :2455-9091

MICROBLOGGING CONTENT PROPAGATION USING SVM AND SVDANALYSIS

[1] N. Baggyalakshmi, [2] Dr. A. Kavitha and [3] Dr. A. Marimuthu
[1] Computer Science Department,
[1, 2, 3] Assistant Professor in Computer Science,
[1, 2] Kongunadu Arts and Science College,
[3] Government Arts and Science College,
[1, 2, 3] Coimbatore.

ABSTRACT- Twitter is an online micro-blogging platform which allows us to treasure trove about the current circumstance at any juncture in time. In this paper, we analyze the sentiments of huge amount of tweets generated from twitter users which are stored in twitter database.We had extracted data from twitter reviews for sentiment prediction using machine learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), Singular Value Decomposition (SVD). Through tokenization, having several stages of pre-processing and several combinations of feature vectors and classification methods, we are able to achieve an accuracy of 89.61% when analyzing the sentiment of tweets.

Keywords – [Micro-blogging content propagation, Hybrid Classification models, SVM and SVD analysis.]

1. INTRODUCTION

In today's highly advanced technological world, people across the world use online platforms to express themselves.They won't trust on traditional media which conveysinformation via one-way channels. Instead, they directlyinvolve and become "the media" through various onlineplatforms to share and express their opinions, perspectives,insights and experiences with each other. A vast volumeof unstructured data will be generated from these onlineplatforms every minute in text format. Due to the ease ofaccess of micro-blogging platform and unrestricted format ofmessages, internet users deviate from mailing lists and blogsto micro-blogging services. Microblogging is type of blogging which consists of limited number of words. Limitation of words determined by respective microblogging sites. It gives right to share his/her thoughts, opinions and sentiments in less number of words. It is one of the revolutionary thing happened in the world of technology. People in these days depends upon microblogging sites such as twitter, Facebook, tumbler etc. to communicate with both relatives and rest of world. Here sentiments comes into the play which will be shared by anyone in the time they feel and wanted to be shared. Sentiments are nothing but feelings respect to event. Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feel

about it. Users share the things about their ongoing life, discuss current issues and variety of topics. Independent to write in any format without following rules that makes this more popular than older blogging sites. Movies and product reviews easily available now days or thoughts on religious and political issues, so it becomes essential sources of user sentiment and opinion. Data that we using in our experiment are from twitter, it contains vast number of messages by large number of users created by themselves. Twitter is one of the extended popular social microblogging service where users of twitter generate messages called tweets, which expresses their feelings/opinions on thethings that interests you. Hence twitter can be taken as avaluable source of public sentiment and opinions. In this proposed techniques, we analyze the sentiments of huge amount of tweets generated from twitter users which are stored in twitter database. We had extracted data from twitter reviews for sentiment prediction using machine learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), Singular Value Decomposition (SVD). Through tokenization, having several stages of pre-processing and several combinations of feature vectors and classification methods, we are able to achieve an accuracy of 89.61% when analyzing the sentiment of tweets.

2. LITERATURE REVIEWS

A. Topic Modeling

Probabilistic topic models such as LDA were introduced by [6]. [10] Presented the Author-Recipient-Topic (ART) model to learn the distribution specific to author-recipient pairs. [11] Proposed a supervised learning approach to categorize links and quantify influence of web pages.

Neither work considered information propagation. The supervised learning approach requires a training data set that is a link-labeled and link weighted graph. Our work does not require such training data because it works directly on the microblog messages published by users.

B. Influence Maximization

Influence maximization proposed in [2] aims to identify aset of seed users who could influence the most number of other users in a social network. Two popular influence propagation models are Independent Cascade Model and Linear Threshold Model. These models assume influence probability based on simple heuristics, such as uniform probability or probability proportional to the degree of a node. Moreover, this problem does not have a target message nor consider the topics for a link. Most previous works focused on improving the efficiency of greedy algorithms [15], [16], such as the CELF optimization based on the sub modularity of incremental influences.

C. Model-based Information Diffusion

Richardson and Domingos [11] proposed a probabilistic method for extracting information from a knowledge-sharing network and put forward a hypothesis about the most effective individuals for viral marketing. Subsequently, Kempe et al. [12] proposed a model to maximize the influence of a social network. First, they showed that finding the most influential people is an NP-hard problem. Then, they proposed two models, the Linear Threshold Model and the Independent Cascade Model, and used them to simulate information propagation in a social network.

D. Information Propagation on Real Data

With the increasing availability of social network data in recent years, researchers have applied different models to analyze the data. Cha et al. [15] exploited Flickr data to construct the relationships between photos and the photographers. They also tried to determine how widely information can be spread and what role word of mouth plays in such a network. Sun et al. [5] investigated the

propagation phenomenon of Facebook's News Feed, for analysis.

3. THEORYFRAMEWORK ANDMODELING

This techniques analyze the sentiments of huge amount of tweets generated from twitter users which are stored in twitter database.We had extracted data from twitter reviews for sentiment prediction using machine learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), Singular Value Decomposition (SVD).

Through tokenization, having several stages of pre-processing and several combinations of feature vectors and classification methods, we are able to achieve an accuracy of 89.61% when analyzing the sentiment of tweets.

The Microblogging propagation model system shows the propagation paths and social graphs, influence scores, timelines, and geographical information among people for the user-given terms. Propagation analysis, based on this framework, it develop a numerical factorization model and another probabilistic factorization variant. The work also develop an efficient algorithm for the models' parameters learning.

3.1 Tokenization

Tokenization involves breaking of words in a tweet into tokens. The words of a tweet are separated by blanks. So, on encountering a space, a new token is formed.

We need to consider several types of dictionaries when analyzing the tweets.

• Lexical dictionary: This consists of maximum of English words that matches the word in a tweet with the list of words in dictionary when analyzing tweets.

• Emoticon dictionary: Emoticons represent some meaning which are textual interpretations of the twitter. This dictionary helps to analyze tweets that consists of emoticons.

• Stop-words dictionary: Some words in a tweet don't have any kind of polarity and those words need not be considered for analysis. Those words are removed and considered as stop words. We conserve a dictionary containing all the stop words.

3.2 Preprocessing (Normalization)

Preprocessing of tweets improves accuracy which involves removal of noisy features. Data is preprocessed in these steps:

• Convert all the collected tweets into lower case

• Eliminate URLs by matching regular expression

• Remove punctuations and extra white spaces

• Remove stop words like 'a', 'the', 'is' etc.

• Eliminate all"@username" and replace it with generic term AT USER

• Remove comma, interrogation marks, double quotations at the starting and at the end of each word

• Remove those words that won't initiate with letters (like 20th, 7:35am etc...)

3.3. Classifiers used

We use supervised machine learning (ML) classifiers namely SVM and SVD. The feature extractors that we use are Unigrams and Bigrams. Here we build a distinguished framework that considers feature extractors and ML classifiers as distinct components, so that we could experiment with different classifiers and feature extractors.

3.4. Extracting Feature Vector

For implementation of a classifier, feature vector is the most vital thing. A successful classifier is actually determined by how good a classifier is. Feature vector builds the set from which the algorithm learns from the data used for training purposes. Feature vector construction involves the removal of irrelevant and redundant words.

In tweets we could utilize the presence or absence of the words that occur in tweet as features. If we have training data, which consists of negative and positive tweets, each tweet is split into words and then adds each word into the corresponding feature vector.

Certain words won't have any effect in indicating the tweet sentiment and hence those words could be excluded from the feature vector. For each one of the tweet, if a word in feature set is absent, it is marked as 0 and if present, marked as 1. The process of adding individual words into the feature vector is known as the 'unigrams' approach. For example, if we have a tweet like"Hey boy, I heard about that contest. Congrats!!!!" Then the feature words extracted into feature vector are 'heard' and 'Congrats'. The tweets are being represented in VSM - Vector Space Model by making use of TF-IDF (Term Frequency - Inverse Document Frequency) weighing score [12], [13].

3.5. Machine learning algorithms

A. Support Vector Machines (SVM)

SVM could be considered as a standard classification technique for any general purpose classification. We use LIBSVM library implemented by Chih-Chung Chang [15] with a linear kernel using [16]. A best survey of SVM is done in the book "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods" [17].

The data that we consider for input has entries, whereeach entry corresponds to the availability of the feature and each occurrence of the word in feature vector is aunique word present in the tweet. If the feature word is absent then the value will be set to 0, otherwise 1. A very detailed explanation is done by Christian I gel [18] on supervised learning of SVM. Since we have not scaled the input, this enhances the processing task [19].

For training a classifier in a supervised learning approach, it requires hand labeled training data. For training a sentiment classifier, it will be very difficult to collect enough training data manually since it is very time-consuming and labor-intensive. So one simple solution is to adopt distant supervision in which training twitter data contains emoticons so that we can run our trained classifiers consisting of emoticons against a set of test tweets (where emoticon data may be present or not). Here we solve this problem using classifiers in a target-independent manner, while the problem could also be solved in a target-dependent classification technique where instead of utilizing information related to current tweet, related tweets are also taken into account.

B. Singular-value decomposition (SVD)

SVDFeature is built based on the traditional matrix factorization approach, and it considers factorization in three aspects, namely global features (also called as dyadic features), user features and item features.

SVD Feature, a machine learning toolkit for highlight based collaborative filtering.SVD Feature is intended to tackle the feature-based matrix factorization. The feature based setting permits us to construct factorization models consolidating side data, for example, temporal dynamics, neighbor -hood relationship and various level information.

The toolkit is equipped for both rate forecast and collaborative ranking, and is intended for preparing on vast scale information set. Recommender system, which recommends things in view of users' interests, has turned out to be increasingly main stream in some circumstances. Collaborative filtering (CF) techniques, as the central purpose behind recommender systems, have been produced for a long time and keep to be a hot region in both scholarly world and industry.

In this system, we concentrate on building collaborative filtering based recommendation toolkit which can successfully influence the rich data of information gathered and actually scale up to extensive information set. Matrix factorization (MF) is a standout amongst the most prominent CF strategies and variations of it have been proposed in particular settings. Be that as it may, customary methodologies outline particular models for every issue, requesting great efforts in building [18] [20].

4. MICRO-BLOGGING ATTRIBUTES

Proposed system solution to micro-blogging feature learning consists of following attributes:

Demographic Attributes: A demographic profile of a user includes, for example gender, age and education can be utilized by e-commerce companies to give better customized services. We extract users' demographic attributes from their profiles on Facebook. Demographic attributes have been appeared to be imperative in promoting, particularly for product recommendation. Main demographic attributes that are used: Gender, Age, Marital status, Education, Career and Interest [1].

Text Attributes: Users' micro-blogs often writes regularly mirror their suppositions and interests towards particular themes. Thus, a potential correlation between text attributes and users' buy preferences [1].

Network Attributes: In the online web-based social networking space, it is frequently observed that users associated with each other are probably going to share similar interests. Accordingly, we can parse out latent user groups by the user's patterns expecting that users in the same group share similar buy preferences [1].

Temporal Attributes: There might be exists connections between temporal activities patterns and users' purchase preferences. Temporal activity attributes, consider two types of temporal activity attributes, namely daily activity attributes and weekly activity attributes. The daily activity attributes of a user is described by a distribution of 24 ratios, and the i-th ratio demonstrates the normal proportion of Facebook published within the i-th hour of a day by the user; similarly weekly activity attributes of a user is described by a distribution of seven ratios.

5. EXPERIMENTALRESULTS

Each work is implemented and simulated under various configuration parameters to know their performance measure values. Here compare proposed model to existing process.

We applied supervised machine-learning algorithms like support vector machines (SVM), Singular Value Decomposition (SVD) with Hybrid Propagation Models and Analysis (HPMA).

The performance measures that are considered for evaluating the improvement of the proposed research methodologies are, "Accuracy, Precision, Recall," The comparison results of this performance metrics are illustrated and explained in the following sub sections.

Accuracy (%)

Accuracy is determined as the overall correctness of the model and is computed as the total actual classification parameters ($T_p + T_n$) which is segregated by the sum of the classification parameters ($T_p + T_n + F_p + F_n$). The accuracy is computed as like :

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)}$$

Where T_p - True positive, T_n-True negative, F_n-False negative, F_p-False positive

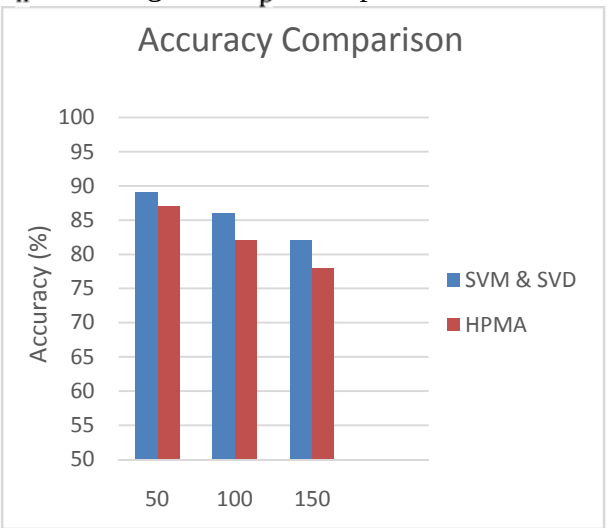

Figure 4.1 Accuracy Comparison

No of input Samples	Accuracy (%)	
	SVM & SVD	HPMA
50	89	87
100	86	82
150	82	78

Table 4.1. Accuracy Measure

From the above Figure 1, it can be observed that the comparison metric is evaluated using existing and proposed method in terms of accuracy. For x-axis the algorithms are taken and in y-axis the accuracy value is plotted.

PRECISION (%)

Precision (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

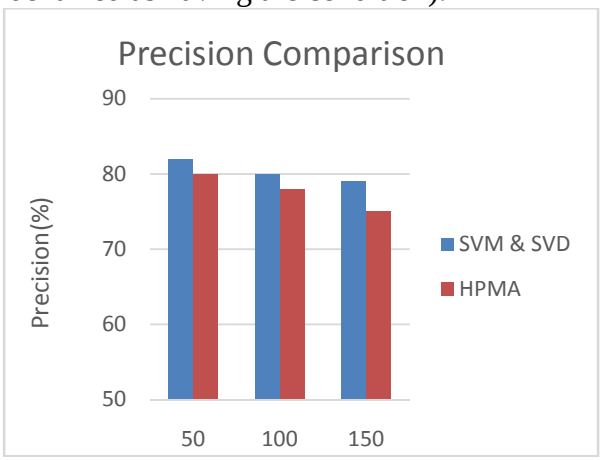

Figure 4.2 Precision Comparison

No of input Samples	Precision (%)	
	SVM & SVD	HPMA
50	82	80
100	80	78
150	79	75

Table 4. 2. Precision Measure

In figure 4.2Precision measure comparisons of the proposed research methodologies is given. This graph proves that the proposed research method can accurately predict the faults present in the software efficiently with improved performance.

RECALL (%)

Recall (e.g., the percentage of healthy people who are correctly identified as not having the condition). Specificity relates to the test's ability to correctly detect classifier without a condition. Mathematically, this can also be written as:

The graphical representation of the recall measurement values of the proposed research methodology is given in figure 4.3.

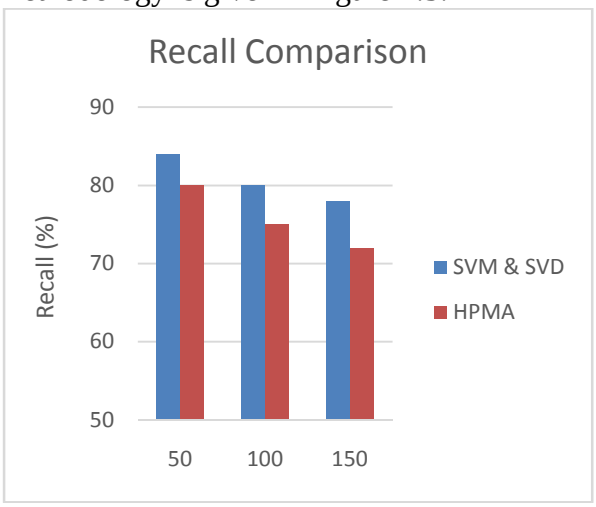

Figure 4.3 Recall Comparison

No of input Samples	Recall (%)	
	SVM & SVD	HPMA
50	84	80
100	80	75
150	78	72

Table 4.3RecallMeasure

In figure 4.3 Recall measure comparisons of the proposed research methodologies is given. This graph proves that the proposed research method can accurately predict the faults present in the software efficiently with improved performance. From this comparison analysis, it can be predicted that the proposed method shows better outcome than previous techniques.

CONCLUSION

The proposed method of sentiment analysis on tweets involved several steps including tokenization of tweets, thorough preprocessing the tokens generated for removal of noise likeURLs, punctuations, stopwords etc. Then we used ML classifiers (SVM and SVD) with different combinations of featureextractors and represented. We have used SVD for dimensionalityreduction related to SVM only. Also we have incorporatedNLTK for sentiment classification. SVM classifier gives

a better accuracy of 89.61%for tweets considering all the possible domains.

Obtainedresults depict that the proposed scheme gives a better accuracywhen compared with existing schemes in the literature. Dataset considered is a vast vocabulary. If we limit our domainto some particular area it is assured that theproposed model will perform better. The reason is that if wetrain our model related to a particular domain and if the testdata is also related to the same domain, then the predictionaccuracy of the classifier will be higher

In the future, we aim to thoroughly examine which features are most influential for an effective sentiment classification and based on that we could modify our feature set.Here we have considered English tweets only. It is possibleto do sentiment classification across multiple languages. As now a days multilingual messages are posted in twitter, so we will able to predict the sentiment for any language.

REFERENCE

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: quantifying influence on twitter,” in WSDM,2011.

[2] S. A. Macskassy and M. Michelson, “Why do people retweet? antihomophily wins the day!” in ICWSM, 2011.

[3] Z. Liu, L. Liu, and H. Li, “Determinants of information retweetingin microblogging,” Internet Research, 2012.

[4] S. Stieglitz and L. Dang-Xuan, “Political communication and influence through microblogging–an empirical analysis of sentiment intwitter messages and retweet behavior,” in HICSS, 2012.

[5] T.-A. Hoang, W. W. Cohen, E.-P. Lim, D. Pierce, and D. P.Redlawsk, “Politics, sharing and emotion in microblogs,” inASONAM, 2013.

[6] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted?large scale analytics on factors impacting retweet in twitter network,” in SocialCom, 2010.

[7] J. A. Berger and K. L. Milkman, “What makes online contentviral?” Journal of Marketing Research, 2012.

[8] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of socialnetworks in information diffusion,” in WWW, 2012.

[9] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a socialnetwork or a news media?” in WWW, 2010.

[10] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, “Twitterpower: Tweets as electronic word of mouth,” JASIST, 2009.

[11] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury,“Information resonance on twitter: watching iran,” in SOMA,2010.

[12] J. H. Parmelee and S. L. Bichard, Politics and the Twitter revolution:How tweets influence the relationship between political leaders and thepublic. Lexington Books, 2011.

[13] P. Achananuparp, E.-P. Lim, J. Jiang, and T.-A. Hoang, “Who isretweeting the tweeters? modeling, originating, and promotingbehaviors in the twitter network,” ACM TMIS, 2012.

[14] C. Castillo, M. Mendoza, and B. Poblete, “Information credibilityon twitter,” in WWW, 2011.

[15] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonalves, A. Flammini,and F. Menczer, “Detecting and tracking political abuse in socialmedia,” in ICWSM, 2011.

[16] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” in WWW, 2004.

[17] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topicsensitive influential twitterers,” in WSDM, 2010.

[18] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, “Mining topic-levelinfluence in heterogeneous networks,” in CIKM, 2010.

[19] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” inICWSM, 2010.

[20] G. Szabo and B. A. Huberman, “Predicting the popularity of onlinecontent,” Comm. ACM, August 2010.

[21] D. Romero, W. Galuba, S. Asur, and B. Huberman, “Influence andpassivity in social media,” 2011.
[22] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, and L. Sun, “Who shouldshare what?: item-level social influence prediction for users andposts ranking,” in SIGIR, 2011.
[23] J. L. Iribarren and E. Moro, “Affinity paths and informationdiffusion in social networks,” Social networks, 2011.
[24] R. Bandari, S. Asur, and B. A. Huberman, “The pulse of news insocial media: Forecasting popularity.” in ICWSM, 2012.
[25] T.-A. Hoang and E.-P. Lim, “Virality and susceptibility in information diffusions,” in ICWSM, 2012.
[26] S. Aral and D. Walker, “Identifying influential and susceptiblemembers of social networks,” Science, 2012.
[27] L. Weng, F. Menczer, and Y.-Y. Ahn, “Virality prediction andcommunity structure in social networks,” Scientific reports, 2013.
[28] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, “Who sayswhat to whom on twitter,” in WWW, 2011.
[29] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonc¸alves, A. Flammini, and F. Menczer, “Political polarization on twitter,” inICWSM, 2011.
[30] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microbloggingduring two natural hazards events: what twitter may contributeto situational awareness,” in CHI, 2010.