The Public Sentiment and Emotional Variations in Social Media using Twitter Dataset

R. Balamurugan, S. Pushpa

Abstract*: The collection of applications (internet followed) that provide way to create communication of user-generated matter by the social media (Twitter, Facebook, Whatsapp, etc.,). Twitter is the micro-blogging platform. Thoughts and opinions about different aspects are shared by users. Analysis on sentiment expressed in a piece of text which expresses opinions, towards a particular topic, product, etc. (positive, negative, or neutral). Primary issues are previous techniques that have biased classification accuracy, due to data distribution in a non-balanced way. The existing methods are applied over small dataset which cannot be extended for generalization with expected accuracy. "Curse of dimensionality" still exists with higher number of attributes in existing methods. Weaker classification in non-linear context. Limited use of transforms (kernels) for linearity in higher dimensional spaces and lack of parameter tuning. In most of the real time dataset the neutral class is very high. The proposed system is the framework for text mining to handle theme extraction from twitter opinion dataset. The Learning models to be built using the Support Vector Tool (SVT) classification method with a kernel trick applied with composition using unigram, bigram and hybrid (unigram + bigram) features. The performance to be obtained by tuning the internal parameters. The result shows that SVT linear kernel with hybrid features are the best classifier when compare to other classifiers with maximum accuracy from the twitter opinion dataset.

Keywords: Sentimental analysis, Twitter, Bayesnet, KNA, SVT, Kernels.

I. INTRODUCTION

In advanced years, because of reputation of gregarious structures organization has essentially broadened and the abundance of data being caused through amiable structures total of Twitter, fb, Google+, etc.,. Gregarious frameworks wind up being well known among a great numerous people who give their phrenic starts in everyday life. The extravagant wellspring of measurements for assessment test is gregarious media sites. To fathom the general people's supposition on remarkable item or solace the conviction examination has been utilized. Twitter, one inside the entire most sizably voluminous and greatest well known pleasant site online which consolidates unstructured records. How the conclusion test highlights. Happy media is a too system for examining enhancements which depend most extreme to a wide association of onlookers and it's miles the reasonable of foundations among people in which they substitute records, set off, and convey starts in structures and advanced systems. Gregarious media enhancements take on a wide scope of frameworks which incorporates gregarious machine, magazines, net social affairs, wiki, weblogs, littler scale running a blog, agreeable sites, webcasts, photographs or pictures, video, gregarious bookmarking and rating.

Sentiment Analysis micro blogging websites have evolved. The utilization of gregarious media is growing day by day. Enlarging impact of gregarious media customers over web has likewise improved their pastime in sundry trades and activities on the equivalent time. Substantial examining this kind of mass total reviews is an astoundingly debilitating errand.

So there's a reason for a customized structure in the event that you need to set off along these lines remove the enormous and negative features of the thing and go to a choice the essential authority strategy well-ordered simple. A couple of locales and gatherings which play out those donning exercises. There are sundry ways to deal with arrangement with setup gadget acing counts. The use of ML counts is to utilize observations as realities and this recognition can be an insights, plan and past experience.

close by these lines gadget language estimations use to reconsider the introduction of events, which must be practical by means of any classifier by endeavoring to transfer the data design into set of groupings or to bunch hard to get models. Since device language computations it improves its introduction from past revel in or by tolerating input. It completely might be dissevered into two preparing directed and unsupervised technique.

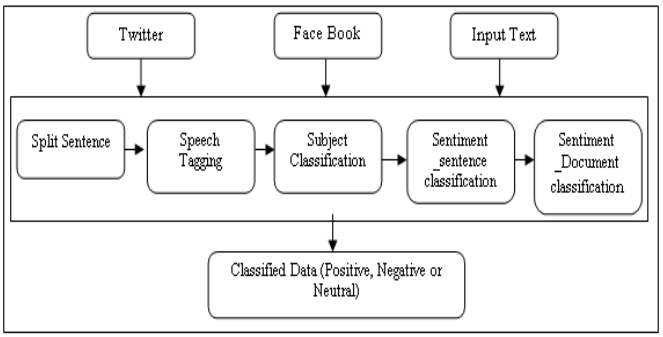


Fig. 1. Steps for Sentiment Analysis process

Revised Manuscript Received on July 22, 2019.

***** Correspondence Author

R. Balamurugan *, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: chennaibalamurugan@gmail.com

S. Pushpa, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Chennai, Tamil Nadu, India.

A. Coordinated

In coordinated becoming acquainted with, the cases are set apart with kenned or objective directions names. here up to exchange the dataset kens the goal class. at some point or another it's far reinforcement for the inconveniences that have kenned inputs.

B. Uncoordinated

In unsupervised acing, the count packs the models by utilizing their homogeneous highlights in estimations of features and makes various organizations. this no previous class or packs are given, the estimation itself portrays their organizations consequently and authentically. Feeling examination is a component language overseeing and data extraction mission. This strategy way to remove researcher's feelings conveyed in reviews or comments.

II. LITERATURE SURVEY

The aim of the paper is to categorize the twitter messages as negative, positive or neutral. For this we utilize Tool Learning approach. Tool Learning approach includes the three algorithms

1) Naïve Bayes (NB) 2) Support Vector Tool (SVT), 3) K-Most proximate Neighbors (K-NN).

This includes the training the relegation algorithm. In tool learning methods are coalescing but they utilize different sentiment analysis methods. They are varied than our approach; we first preprocess the data to abstract unwanted data from it. Tool Learning strategies utilized a test set and training set for a relegation. Training set contains their corresponding class labels and information feature vectors by utilizing this training set; a relegation model is engendered which endeavors to relegate the information feature vectors into corresponding class labels.

At that point a test set is utilized to accept the model by soothsaying the class labels of unseen feature vectors. Sundry tool learning methods like Support Vector Tools, Bayesnet (NB), and K-Most proximate Neighbors (K-NN) are acclimated to relegate reviews. Term Presence, Term Frequency, negations, n-grams are some features that can be utilized for opinion relegation.

To identify the semantic exordium of words, expressions, sentences and that of documents, these features can be utilized. Semantic prelude is the polarity which may be either negative or positive.

III. PROPOSED METHODOLOGY

To reach the following aim the methodology is followed as shown in the flowchart.

A framework for text mining to handle theme (topic) extraction from twitter opinion data set.

SVT classification method with a kernel trick applied with composition of kernels. (Linear, Polynomial, Radial Basis Program (RBP) & Sigmoid)

Maximum accuracy and the performance to be obtained by tuning internal parameters.

A. Data Accumulation

Process of amassing and quantifying information on targeted variables in an established systematic fashion, which then enables one to reply pertinent questions and evaluate outcomes, is called data accumulation.

B. Genetic search

A method for clearing each constrained and unconstrained reducing quandaries find out on a natural cull process that mimics biological evolution. The algorithm perpetually changes a population of individual solutions.

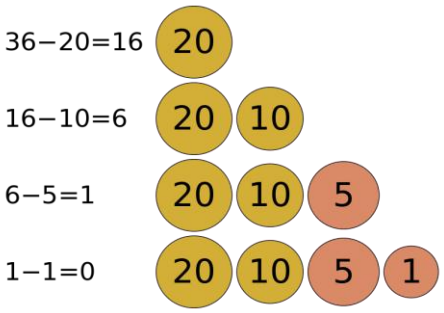


Fig. 3. Genetic search

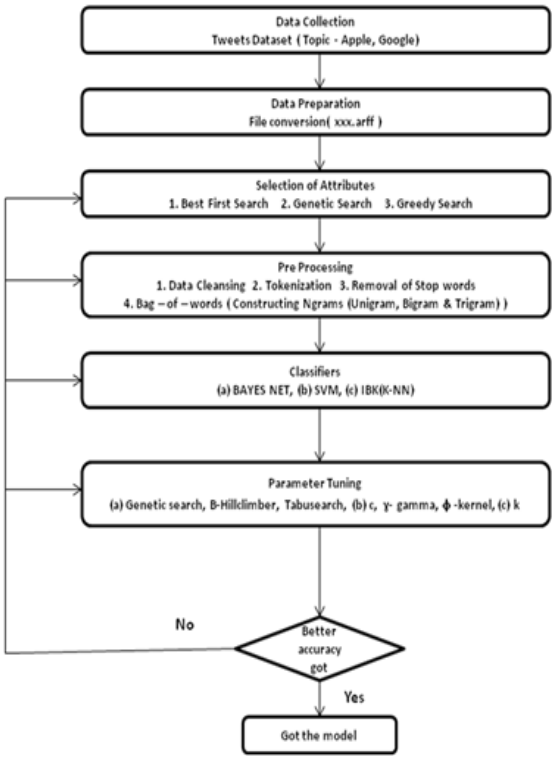


Fig.3. Proposed methodology

C. Greedy search

An algorithmic paradigm that follow the quandary solving heuristic of making the domestically most excellent cull at every stage with the desire of finding an ecumenical optimum

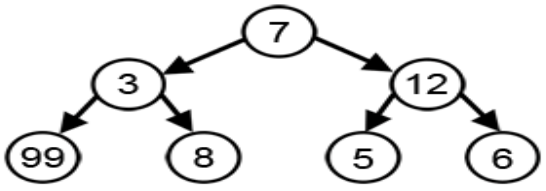


Fig.4 Greedy search

IV. PRE-PROCESSING

If you are using *Word*, use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). “Float over text” should *not* be selected.

A. Data cleansing

The system of detecting and redressing corrupt or erroneous records from a record set, table, or database and refers to identifying incomplete, erroneous, erroneous or extraneous components of the data and then superseding, modifying, or expunging the dirty.

B. Tokenization

The method of superseding sensitive data with unique find out symbols that retain all the consequential information about the data without compromising its security.

C. Replace of Stop words

The most mundane words in a language, there is no single ecumenical list of stop words utilized by all natural language processing implements, and in fact not all implements even utilize such a list. Few implements categorically eschew abstracting these stop words to fortify phrase search.

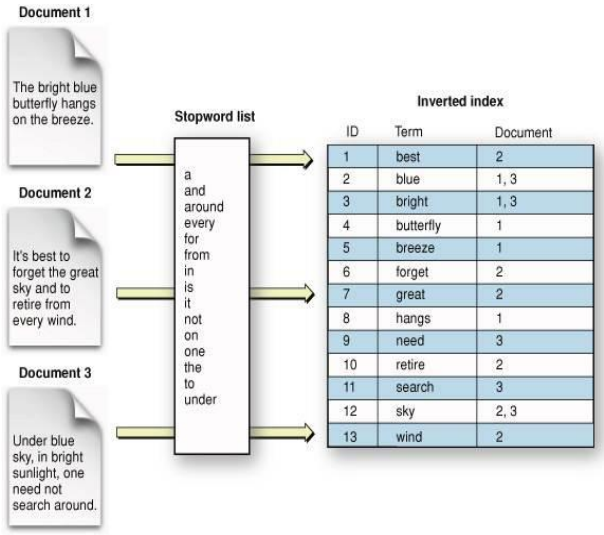


Fig. 5. Removal of Stop words

D. Container-of-words

A simple demonstration accept in natural language and information recovering. In this, a text is represented as the container (many set) of its words, avoid grammar and also word order but placing multiplicity.

E. n Grams

Within the fields of computational linguistics and possibility, an n-gram is a contiguous continuation of n items from a given content of text or verbalization. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or verbalization corpus. Words rudimentary, paramount elements with the faculty to represent a different meaning when they are in a sentence. By this point, we recollect in mind that sometimes word groups provide more benefits than only one word when explicating the construal. Here our sentence "I read a book about the history of America." The tool wants to get the meaning of the sentence by separating it into small pieces.

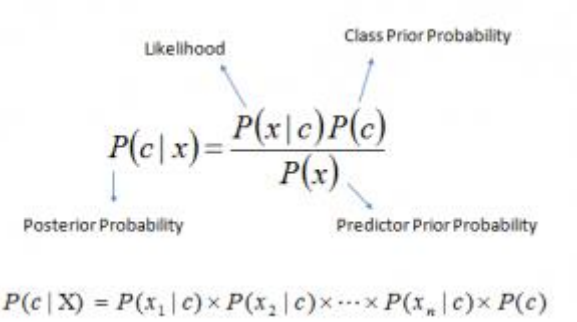
How should it do that? 1. It can regard words piecemeal. This is unigram; each word is a gram."I", "read", "a", "book", "about", "the", "history", "of", "America" 2. It words two at a time. This is bigram; each two neighbor words engender a bigram. "I read", "read a", "a book", "book about", "about the", "the history", "history of", "of America" 3. It words 3 at a time.

This is trigram; each three adjacent words create a trigram. "I read a", "read a book", "a book about", "book about the", "about the history", "the history of", "and history of America"

V. CLASSIFICATION ALGORITHMS

A. Bayes Net

Model is a Bayesnet. It reflects the states of some part of a world that is being modeled.. might be the model A Bayesnet can be modeled by anything absolutely. It is a relegation method predicated on Bayes' Theorem with a posit of independence among prognosticators. A Verdant Bayes classifier surmises that the presence of a particular feature in a class is not related to the presence of any other feature.



B. IBK (K-NEXT ADJACENT)

In pattern apperception, the K-Most proximate Adjacent Algorithm (K-NN) is a non-parametric approach utilized for regression and relegation. In each case, the input includes of the 'k' most proximate training examples in the feature space.

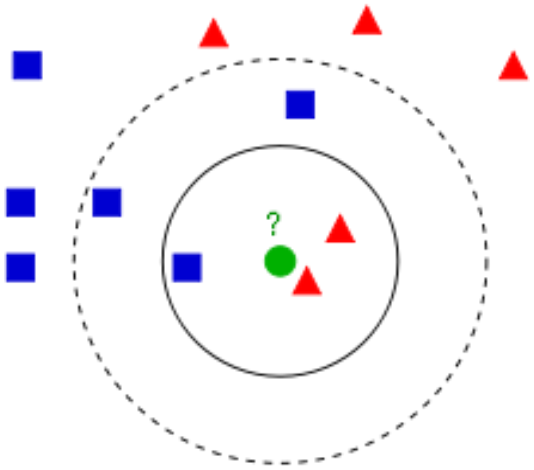


Fig. 6. IBK

C. *Equations* Support Vector Tool (SVT)

Support vector tools are supervised learning models with associated learning algorithms that analyze data utilized for relegation and regression analysis. SVT is a technique utilized for relegating the linear data. The main principle of SVTs is to decide linear separators in the search space which could excellent disunite the different classes. The SVT method utilizes a nonlinear mapping to transform the training data set into high dimensions. The SVT finds the hyper plan utilizing the fortification vector. SVT has been used prosperously in text relegation and in a variety of sequence processing application.

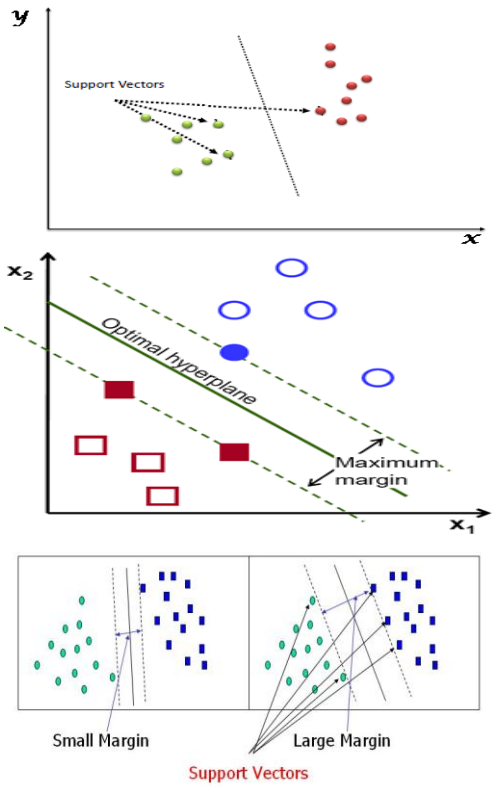


Fig.7. Support vector tool

VI. SUPPORT VECTOR TOOL - KERNEL PROGRAMS

A. Linear kernel:

Often linearly separable is Text. The majority of text classification problems are linearly separable. When we practice a SVT with a kernel, we only want to reduce the C regularization parameter.

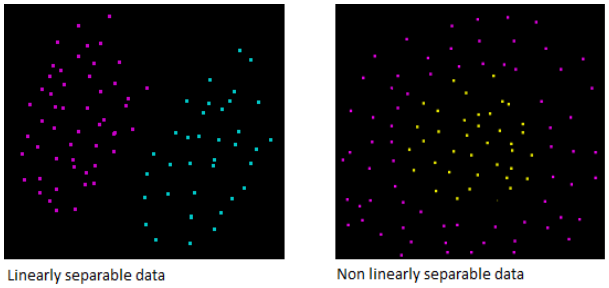


Fig. 8. Linear kernel

B. Polynomial kernel:

The polynomial kernel is a kernel program usually used with SVT and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models. We have to tune the parameters d of SVT polynomial kernels.

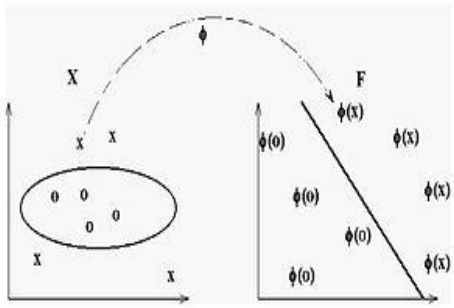


Fig. 9. Polynomial kernel

C. Gaussian (Radial-Basis Program (RBP)) kernel:

The (Gaussian) radial substructure program (RBP) kernel is a popular kernel program utilized in sundry kernelized learning algorithms. In particular, it is commonly utilized in support vector tool relegation.

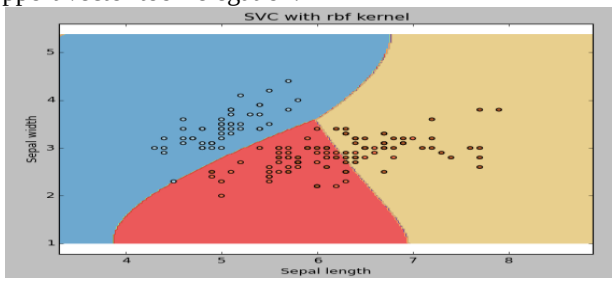


Fig. 10. Radial-Basis Program kernel

Tuning the parameters(C and gamma) of SVT kernels. Higher the value of gamma, will endeavor to exact fit the as per practice data set. Example: varied ϒ values like 0, 10 or 100. (kernel='RBP', C=1 ϒ,=0)

D. Sigmoid:

The Sigmoid Kernel comes from the Neural Networks (NN) field. SVT model need a sigmoid kernel program is equivalent to a double-layer, perception neural network. We have tuned the parameters (C, coeff and gamma) of SVT sigmoid kernels.

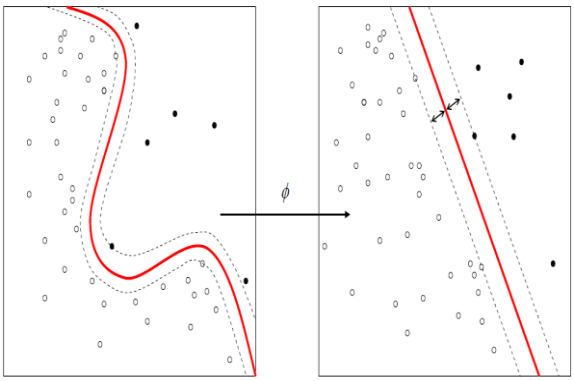


Fig. 11. Sigmoid

VII. INFORMATION GAIN ALGORITHM USED

Algorithm for selection of attributes based on information gain

Input :

< Θn, IG (Θn) >

// Θn = Original set of attributes a1,a2,..,an ∧ IG(Θn) = {IG(a1),IG(a2),..,IG(an)} with IG(a1)≤IG(a2)≤…≤IG(an)

Process:

i = n; τ = desired accuracy

Ai = Classify (Θi);

// Get accuracy with Θ being the training set , 10-folded CV is applied, and Mi is the associated model

While (Ai < τ)

{

i = i-d

// d =length part of Θ or equivalently Θ is divided into d equal parts

Ai = Classify (Θi);

}

Output Mi

// Mi is the optimal model with maximum information gain and accuracy

A. Datasets

The Sanders (Twitter dataset) is used as dataset for over whelming analysis. It consists of 2524 tweets. Each ingression contains: Tweet id, Tweet text, Tweet engenderment date, Sentiment utilized for Topic, over whelming label: 'positive', 'neutral', 'negative', or 'impertinent'. A neutral class is very high in the majority of the real time dataset.

B. Tools

Weka tool - It provides many tool learning and different algorithms for data mining. It is freely available and open source. It is platform-independent. Who are not data mining specialists, they can easily useable this tool. For scripting experiments, it provides flexible facilities. As they appear in the research literature, it has kept up-to-date, with new algorithms being added.

Table. 1. Dataset

Topic	#Positive	#Neutral	#Negative	#Irrelevant	#Twitter search term	Total Tweets
Apple	163	518	315	138	@apple	1134
Google	205	621	57	507	#google	1390

Table.2. SVT kernel Dataset

Topic	Kernel type	Cross validation	Feature	Performance / Accuracy %
Apple	**Linear**	**10**	**unigram & bigram**	**57**
Google		**10**		**57**
Apple	Radial basis program (RBP)	10	unigram & bigram	46
Google		10		45
Apple	Polynomial	10	unigram & bigram	46
Google		10		45
Apple	Sigmoid	10	unigram & bigram	46
Google		10		45

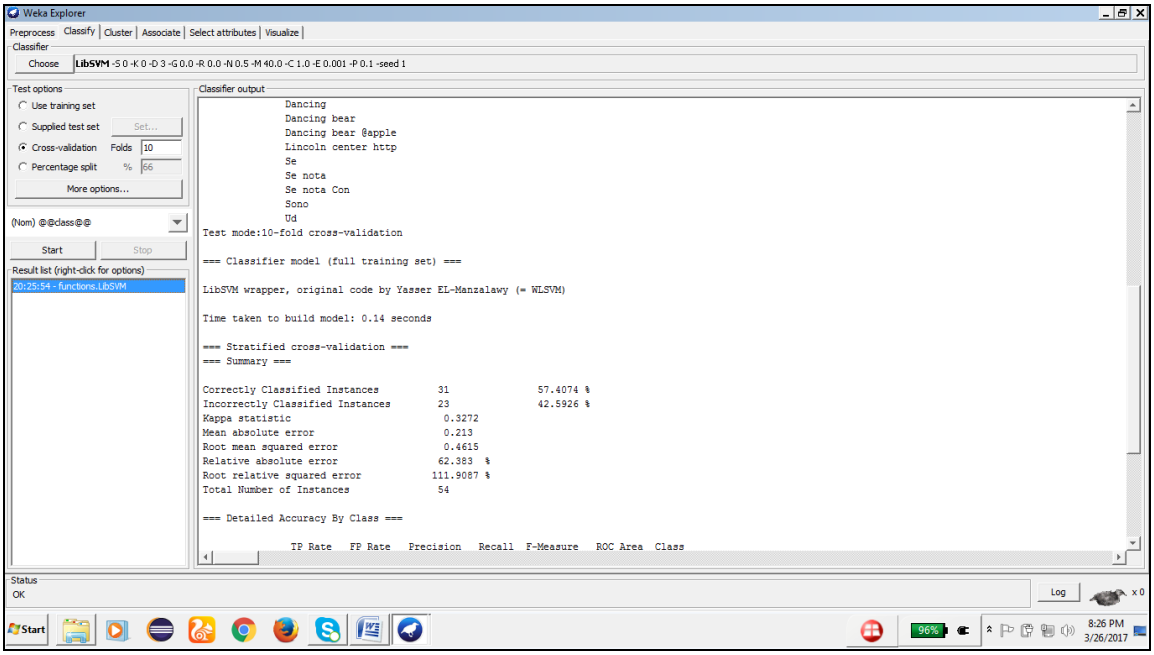


Fig 12(a). Theme "Google" with SVT Linear kernel using hybrid feature (unigram and bigram) by Weka tool.

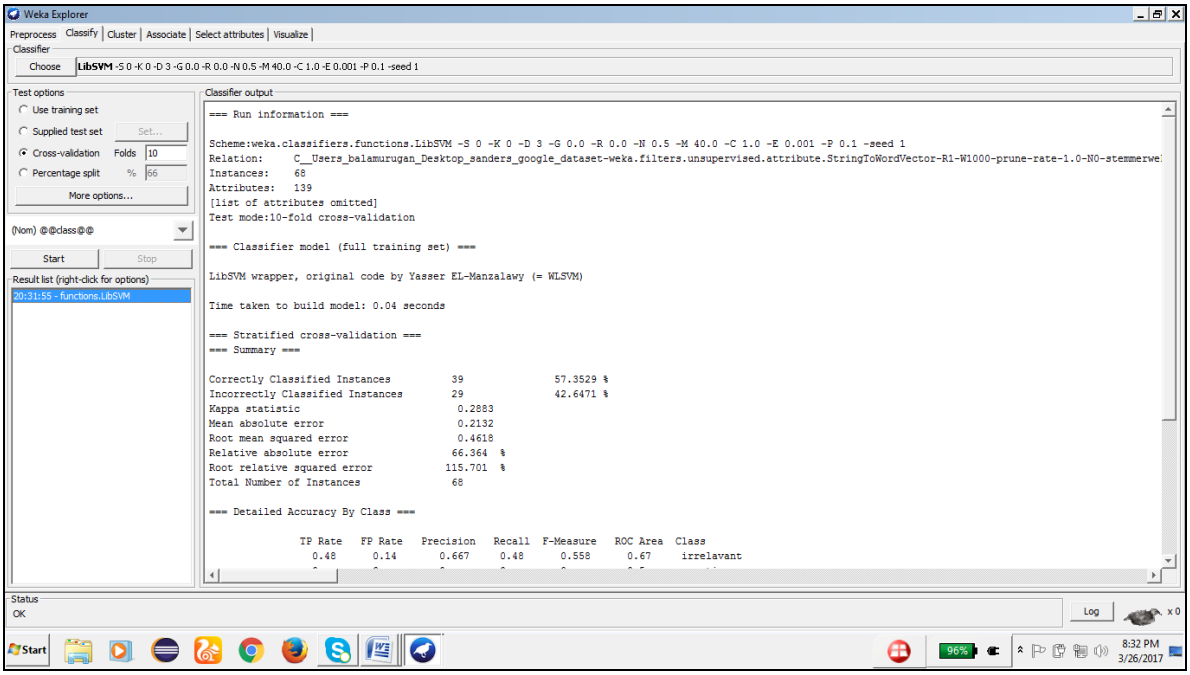


Fig 12(b). Theme "Google" with SVT Linear kernel using hybrid feature (unigram and bigram) by Weka tool

VIII. CONCLUSION

Our go through represents that ML strategies are less arduous and more efficient in this paper. There are varied procedures to relegate the sentiments from content. The techniques were applied for twitter sentiment analysis. Here there are problems while dealing with identifying sentiment keyword from tweets having multiple keywords. It is also challenging to solve miss spelt and slang words. To solve the difficulties in data, to manage this issue, an efficient feature vector is made by doing feature extraction after opportune preprocessing. By utilizing different classifiers like Bayesnet, K-Most proximate Neighbors (K-NN), and SVT kernels the relegation precision of the feature vector were tested. This feature vector performs subsidiary for theme. The message contact in Twitter was found with the human demeanor, nature, personality and posture. The relegation of tweets indicates the views of people on theme in to 'positive', 'negative', or 'neutral sentiments'. The SVT linear kernel with hybrid feature performs the best classifier with maximum accuracy with tool Weka. This helps people to select the best theme and they easily decide that which theme is like by majority of people.

REFERENCES

1. Z. Hong, X. Mei, and D. Tao, "Dual-force metric learning for robust distracter-resistant tracker," in Proc. ECCV, Florence, Italy, 2012.
2. X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," ACM Trans. Multimedia Comput.Commun. Appl., vol. 8, no. 3, article 26, Jul. 2012.
3. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.
4. Y. Hu, A. John, F. Wang, and D. D. Seligmann, "Et-lda: Joint topic modeling for aligning events and their twitter feedback," in Proc. 26th AAAI Conf. Artif. Intell., Vancouver, BC, Canada, 2012.
5. X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in Proc. 20th ACM CIKM, Glasgow, Scotland, 2011.
6. J. Weng and B.-S. Lee, "Event detection in twitter," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
7. J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in Proc. 4th ACM Int. Conf. Web Search Data Mining, Hong Kong, China, 2011
8. F. Liu, Y. Liu, and F. Weng, "Why is "SXSW" trending? exploring multiple text sources for twitter topic summarization," in Proc.Workshop LSM, Portland, OR, USA, 2011.
9. J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc.5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
10. Dharmarajan, K., and M. A. Dorairangaswamy. "Discovering User Pattern Analysis from Web Log Data using Weblog Expert." Indian Journal of Science and Technology 9.42 (2016).
11. D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
12. H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in Proc. 3rd ACM WSDM, Macau, China, 2010.
13. C. X. Lin, B. Zhao, Q. Mei, and J. Han, "Pet: A statistical model for popular events tracking in social communities," in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.
14. B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
15. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inform. Retrieval, vol. 2, no. (1–2), pp. 1–135, 2008.
16. T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in Proc. 19th Int. Conf. WWW, Raleigh, NC, USA, 2010.
17. Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, 2010.
18. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," J. Amer. Soc. Inform. Sci. Technol., vol. 61, no. 12, pp. 2544–2558, 2010.
19. D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 260–274, Feb. 2009.
20. D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 260–274, Feb. 2009.

21. J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. 15th ACM SIGKDD, Paris, France, 2009.
22. G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Darmstadt, Germany, Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2009.
23. [23] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Rep., Stanford: 1–12, 2009.
24. Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Interpreting the Public Sentiment Variations on Twitter,‖ IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.
25. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe,"Predicting elections with twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
26. J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011
27. A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Rep., Stanford: 1–12, 2009.
28. L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in Proc. 49th HLT, Portland, OR, USA, 2011.