Prediction And Classification Of Tweets Without Keywords By Utilizing Machine Learning Techniques in Twitter

R.S.LysaPackiam, M.KasiViswanathan ,K.Gokul ,V.Dhana Vishnu Ram
Department of Information technology,Easwari Engineering College, Chennai, India.

***Abstract*—We demonstrate the effectiveness that machine learning can bring to improve social media platforms through a case study on Twitter trending topics. Social media relies heavily on tagging and often does not take advantage of machine learning advances. Twitter is no exception and as a micro blogging platform,it has vast potential to become a collective source of intelligence. Individual tweets are identified as being part of a trending discussion topic by the presence of a tagged keyword. Our research demonstrates that machine learning techniques can be used identify the top trending tweets up to 90.2% precisionwithoutusing the identifying keyword as a feature.Classifiers in WEKA machine learning software is used as an algorithm to produce outcome by categorizing the topics. We also tried to predict the classification through training set and test set. This can aid in improving the quality of topic categorization by ensuring on-topic tweets that are missing the trend keyword are included, as well as suggest keywords to include in new tweets.**

Keywords— Machine learning tool, Classification , Prediction, twitter.

I. INTRODUCTION

Twitter is a microblogging and social networking servicethat allows users to add friends and send messages to thefriends that follow them in the form of "tweets"– messages of up to 140 characters.One interesting thing about Twitter is that it has a section on its main page entitled "Trending Topics", which displays the top 10 mentioned terms on Twitter at any given moment. This is generated based on Twitter's proprietary algorithm [2]. These trends indicate topics that arebeing discussed heavily on twitter and are tracked by keyword. For example, if "CWC2015" is a trending topic then looking at allthe posts under that trend will reveal any post with the keyword "CWC2015"or"#CWC2015".Commonly used terms,such as "coffee,"are removed and the top 10 trending topics are displayed onthe Twitter homepage.Effectively organizing trends and the tweets that belong tothem is significant because of a wide range of parties thatbenefit from doing so. Recent events and new research showthe increasingly important role of social media (and Twitterspecifically) such as the Oscars2015 and Commemorating International women day in twitter.

Many companies see this service as an important place to monitor and promote their brands. The monitoring of message helps companies understand a specific environment and its constant changes with one basic principle: if something said in the social media,then it can be qualified and quantified[3].The level of interest in Twitter trending topics specifically can be seen through whatthetrend.com, a site recently created by Twitter to help people understand what the trends are and why they are being discussed.Given the importance of identifying trending tweets, a major problem that trend classification faces is a relatively crude method for identifying which tweets are part of a specific trend. Using only the trend keyword to identify a trend will miss relevant and potentially important pieces of discourse[1].

Take, for example, the #bahrain trend which was recentlypopular due to the protests occurring on February 15th, 2011. Here is a tweet identified as being part of the #bahrain trend:

> "#Bahrain 's Pearl Sq looks like Tahrir Square all over again. Police gathered in force but standing back for now."

And here is another tweet that is clearly related to the sametopic but would not be identified as being part of the #Bahrain discussion due to the lack of the proper keyword:

> "just got back from manama and the scenes are overwhelming women are being part of the protest wow just wow".

This illustrates how identifying trend-tweets through keywords alone is insufficient. Important elements of discoursecan be easily overlooked. The recent boom in social mediabeing used both as an organizational tool and a journalismtool shows the necessity for more comprehensive means oforganizing discourse in these mediums.

The motivation behind our work is that social networks andmicroblogs which are products of Web 2.0 are becoming the toolof choice for information dissemination, sharing and interpersonal communication and networking. It is a new platform beingrapidly adopted by all walks of life; from politicians tobusinessmen for use in citizen journalism tobeing a medium to stay close to friends and family [2].

Our research explores this need by demonstrating identification of which trending topic a tweet is a part of withoutknowing if any trend keywords are present in the tweet. Weare able to successfully perform this task with the

ISSN: 2278 – 7798

help of WEKA ,a machine learning software with high accuracy(90.2%), showing the potential for augmenting existing taggingsystems based on collective human intelligence with machinelearning to get a more comprehensive view of discourse relatedto a specific topic.

A. Related Work

There has been a great deal of interest in topic modeling and analysis, especially on microblogging platforms like Twitter. One of the most recent examples is the work by Ramageet. al. that uses topic models to characterize informationneeds of Twitter users [4]. This research is effective in better representing content on Twitter to the users that want it, but it still relies on the existing system of identifying trending tweets by keywords. A similarity between Twitter chatter and blog postings have been suggested in [5, 6], leading us to conduct research on Twitter (as a micro blogging platform) to be able to provide an insight into how the Twitter community's chatter accurately reflects the properties of real-world events represented in the form of trending topics.

Phelan et. al. propose a novel technique for identifying news stories of interest based on twitter feeds. Their motivation is that recommender systems take time to develop, requiring a "critical mass" of stories before accurate predictions can be made [7].Bernstein et. al.'s work on Eddi groups tweets based onsubjects explicitly or implicitly mentioned in a tweet [8].While this has real applications for trend classification it's reliance on search engines as a knowledge base means that itwon't have the same effectiveness on Twitter-specific trends.

Machine learning is a branch of artificial intelligence.We use WEKA [9] to find patterns in training datasets and from their own rules which are then used for making forecasts in testing datasets.Chang et. al.'s paper "Reading Tea Leaves: How Humans Interpret Topic Models" present a new method for assigning documents to topics in large document collections [10].

In the previous research [1], the research work shows that Identification of a trending topic in a tweet without using a keyword is performed by evaluating the WCNB and MNB classifier.

II. APPROACH

To determine whether or not it is possible to identify whichtrend, if any, a tweet is a part of we created a data set basedon current Twitter data, extracted a feature set from the tweetsand users in the data, and implemented an improved version of a naive bayesian classifier to apply to it [1].

Keyword	Description
#CWC2015	Trending topic from 14th February 2015.This refers to cricket worldcup 2015 season which had begun in February.
#OSCARS2015	Trending topic on 22nd February 2015.This refers to award ceremony presented by Academy of Motion picture Arts and Science.
#InternationalWomensDay	Trending topic on 8th march 2015. This refers to Commemorating International Womens day tweet.
#Mauritius	Trending topic on 12th march 2015. This refers to India's honourable PM visit to Mauritius.

Table 1.List of trending topics selected

LOL	Non trending but a common word included here for comparison during classification
RT	Non trending but a common word which means Retweet used in many tweets included here

Table 2.List of Non-trending words selected

A. Data and Feature Set

Our system relies on word frequency counts in both theindividual tweets and the information provided in the profile of a tweet's author as well as the time zone provided by the author's profile. Trending topic keywords are not includedin the feature set for any tweet. In addition, the weightsof user profile word frequencies are reduced by 60%. Thisnumber was determined through experimentation that showedequal weighting of profile and tweet word frequencies did notoptimize results. This makes sense as users may choose towrite about a wider range of topics than their profile suggests [1].

The data set used in this research was created specificallyfor this project in offline. The final set included tweets in English from the top 10 current trends on Twitter were obtained on three different days (Feb 14th through Feb 16th, 2015) for a total of 30526 tweets belonging to a combined 20 trends. In addition 1096 tweets were collected on Feb17th from the public timeline, from which all tweets with any reference to the 20 trends were removed leaving 831 public not-trending tweets. The text in each tweet was parsed into word and punctuation tokens which were then used for token frequencies.

ISSN: 2278 – 7798

A "clean" data set was also created from the original data. This set included only tweets with greater than 15 words and punctuation tokens and did not include more than one trend keyword. This data set was reduced to 14553 tweets.

For all data keywords relating to the trending topics wereremoved so as not to influence the classification task (leaving them in would heavily skew the precision towards correctclassification as those words would clearly indicate they belonged to a specific trend). They keywords to remove were taken directly from the trending topic name itself, so in thecase of "coal scam" both the word "coal" and the word "scam" were removed from tweets. If the trending topic was simply"GoodAdvicein4Words" then that is the only token that was removed,the word "Advice" alone was not removed.

B. *Naïve Bayes Classifier*

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

Constructing a classifier from the probability model

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k). \qquad (1)$$

C. *Multinomial naive Bayes*

With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial $(p_1, \dots, p_n)$ where p_i is the probability that event i occurs (or K such multinomials in the multiclass case). A feature vector $\mathbf{x} = (x_1, \dots, x_n)$ isthen a histogram, with x_i counting the number of times event i was observed in a particular instance. This is the event model typically used for document classification, with events representing the occurrence of a word in a single document (see bag of words assumption). The likelihood of observing a histogram x is given by

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i} \qquad (2)$$

Newer machine learning approaches such has Support Vector Machines (SVMs) are heavily favored over bayesian approaches but the viability of the bayesian approach.

Why Not A Support Vector Machine?

The biggestachievement of TWCNB is that its effectiveness is comparableto that of a support vector machine (SVM) [1]. We confirmedthis in our own initial findings early in the study, demonstratingthat the 1% to 4% decrease in precision when using MNB over SVM that Rennie et. al. find extends to our data. With ourvery large data set we found that training an SVM can take upto one hundred times longer than the bayesian approach. Thisis important in Twitter because the generation of trending topicfeeds needs to be done in near real-time. Given the scale that asocial media platform like Twitter we feel that small differenceprecision between SVM and MNB approaches is not worththe trade off in training time.

III. EVALUATION

We evaluated our data in a variety of ways to thoroughlyanalyze the strengths and weaknesses of our approach. Wefirst established a baseline using standard unsupervised topicmodeling using the original tweet data without the trendkeywords. We then analyze our approach, breaking down the Effectiveness on different days, which trends were easier to pick up, and the improvement of our MNB approach overtraditional Naïve Bayes.

A. *Topic-Modeling Baseline*

To derive a baseline for our analysis, we also analyzedhow well basic, unsupervised topic modeling [11] fared at the task of detecting and classifying tweets. We used the topicmodeling functionality in Stanford NLP tool [12]. We ran Stanford topic modeling toolbox with default parameters, creating three 10 topic models of 10topics each. The results can be seen in Figure 1.

The topics generated by the topic modeling software did not always correlate with the trending topics. For example, in one of the models, topic 3, "amp unlimited http iphone tethering ipad makes announces customers big iphones coming users kills tiered end drops plan read" and topic 7 "http ly bit news wireless phone stop usage att caps leak eu pricing give url blog eat cap offering" corresponded to the trending topic "DataPlans" whereas no topics in that model corresponded to the trend #CWC2015. We assigned a topic to a trend based on

ISSN: 2278 – 7798

theclass that the plurality of tweets assigned that topic belongedto (thus, adding an element of supervision to the problem).The topic model represents a document (in this case a tweet)as a combination of topics. We considered the topic assignedthe highest proportion to be the class assignment. It is likelythat a supervised topic model [13], and/or a model using acombination of multiple topics might do better. However, theseresults are intended to provide a base level of performance.

Each model was evaluated by precision, recall and f-measure. For each tweet, we assigned it to a trend if two ofthe three models agreed upon where it should be assigned andleft it unassigned otherwise, resulting in increased precision.The results were characterized by high variance. They formed much better on certain trends such as "RachelCorrie" and "Data Plans" than others like "#everlastingfriends" and #isupport. News-oriented topics tended to garnerbetter results than social topics.

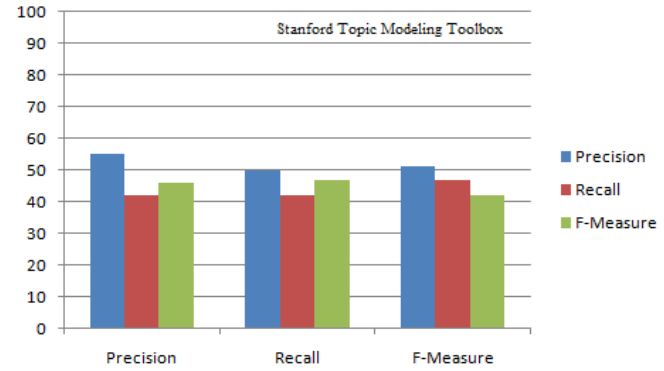


Fig. 1. Precision, Recall and F-Measure for the top 10 trends of each day in our data set using Stanford Topic Modelling.

B. *Our Results*

The effectiveness of classifying tweets as part of a trend was tested across different classifiers, feature sets, and data sets. The "standard" classification setup which is used as a basis

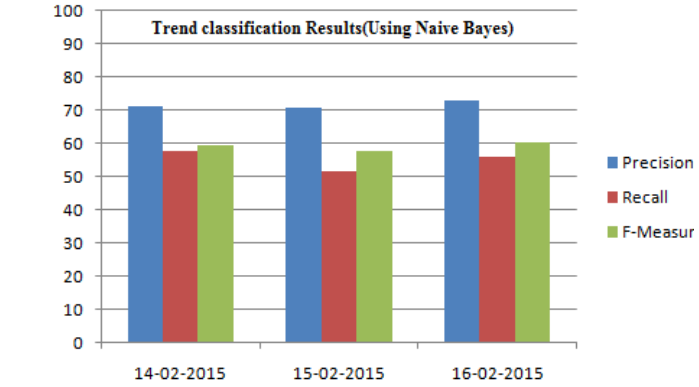
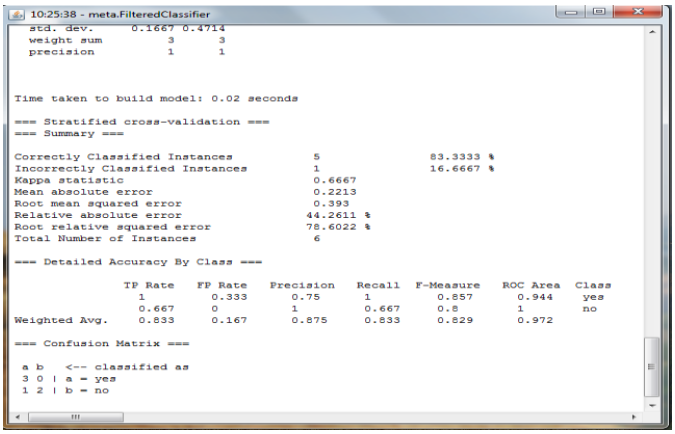


Fig. 2. Precision, Recall and F-Measure for the top 10 trends of each day in our data set using the Naïve Bayes classifier.



for comparison against all of the other methods is the MNB classifier, 3 days of data with 10 individual trends for each day (no public "no-trend" timeline included), all three major features (tweet text, user description text, and time zone), and the original data set that has not been scrubbed in any way.

As we can see in Figure 2, the classifier is quite effectivein correctly classifying the tweets correctly. Depending on theday the precision ranges from 70% to 73%, the recall from56% to 58% and the F-Measure from 58% to 60%

In thesefigures we see that the variance in accuracy between trendscan be significant with some trends, such as "CWC2015"exhibiting great differences in precision and recall and others like "Dhoni captaincy" and "India" showing almost perfect recall orprecision. The graphs primarily show that accuracy remainsrelatively uniform with onlyafewpoorly classified trends. The MNB classifier showed noticeable improvements overa standard naive bayesian classifier [1].

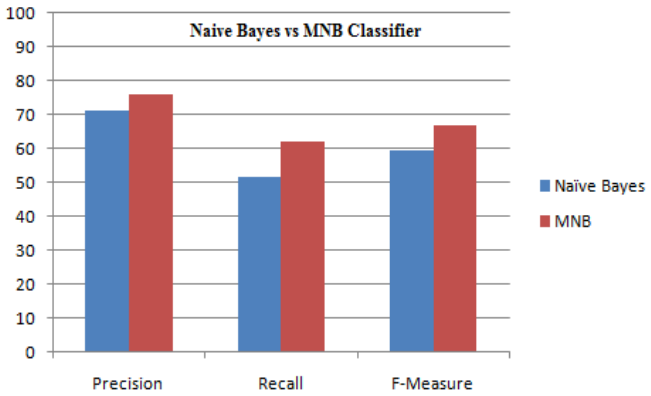


Fig.3. Results for Naïve Bayes vs MNB Classifier , averaged across all three days.

In fig. 3, We can able to obtain the difference between Naïve bayes and Multinominal Naïve bayes classifier for Precision, Recall, F-measure.

ISSN: 2278 – 7798

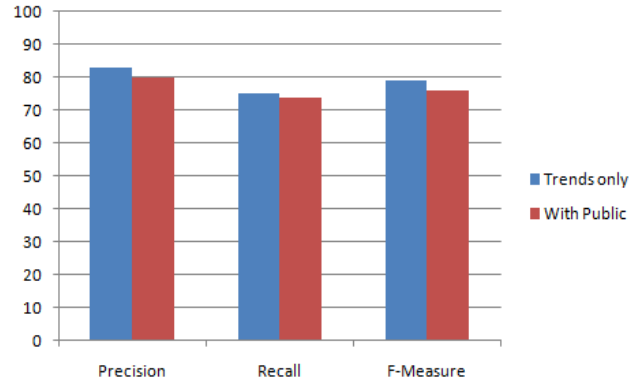


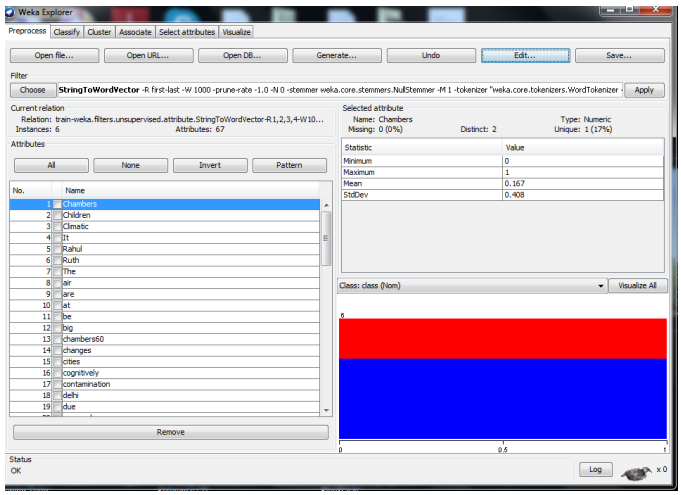
Fig. 4.Comparing the addition of public non-trending tweets. Averaged across all three days.

IV. PREDICTION OF CLASSIFICATION

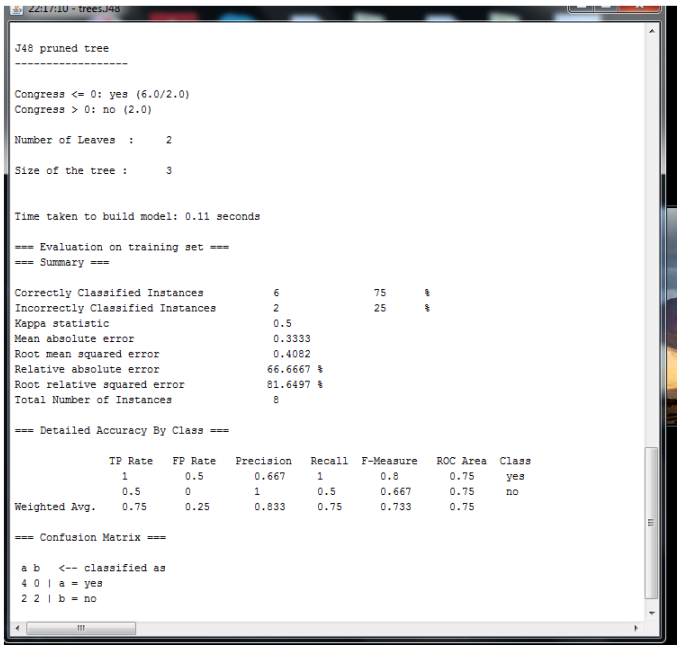
It is a section after comparing traditional Naïvebayes versus Multi-nominal Naïve bayes (MNB) we tried to have a machine learning approach for automatic prediction of classification instead of classifier for given data set. For prediction we need to train the machine with training set provided that a classification is already given for training set. Because it is always necessary to give a machine how to find pattern by providing classification in order to make it predict for future data set.To make this prediction happen it depends on how much we train the machine in order to achieve high accuracy for predicting classification for any given text input.

Steps involved in prediction of data set classification:
Training set:

- First we have to give the trained dataset(manually classified input lines which are classified as "Yes or No" as per logic) in preprocess tab in WEKA.
- Apply "StringToWordVector" filter in order to change the text input into tokens

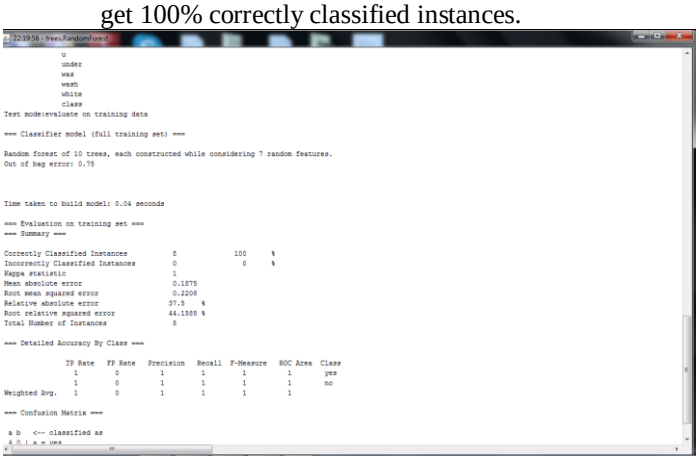


- After applying filter we try to classify the data set with "J48" algorithm to check its classification precision with given output and predicted output.
- For our given dataset it gives 75% correctly classified instances and 25% incorrectly classified instances.



```
J48 pruned tree
------------------

Congress <= 0: yes (6.0/2.0)
Congress > 0: no (2.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0.11 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           6               75      %
Incorrectly Classified Instances         2               25      %
Kappa statistic                          0.5
Mean absolute error                      0.3333
Root mean squared error                  0.4082
Relative absolute error                 66.6667 %
Root relative squared error             81.6497 %
Total Number of Instances                8

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.5        0.667      1         0.8        0.75     yes
                 0.5       0          1          0.5       0.667      0.75     no
Weighted Avg.    0.75      0.25       0.833      0.75      0.733      0.75

=== Confusion Matrix ===

 a b   <-- classified as
 4 0 | a = yes
 2 2 | b = no
```

- In order to make it more accurate we try "Randomforest" tree algorithm for classification we get 100% correctly classified instances.



```
Test mode:evaluate on training data

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 7 random features.
Out of bag error: 0.75


Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances           8              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.1875
Root mean squared error                  0.2208
Relative absolute error                 37.5    %
Root relative squared error             44.1588 %
Total Number of Instances                8

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1         1          1        yes
                 1         0          1          1         1          1        no
Weighted Avg.    1         0          1          1         1          1

=== Confusion Matrix ===

 a b   <-- classified as
```

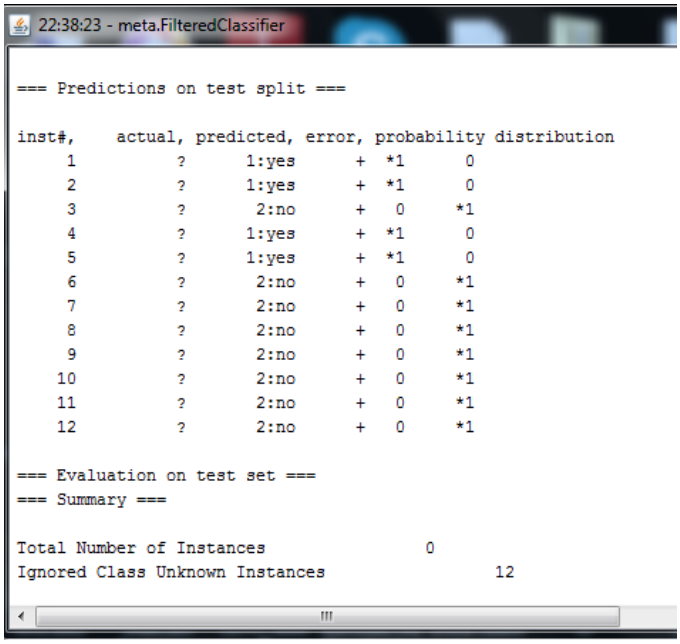
Supplying Test set for prediction

After training the data set with many lines of inputs we can provide the machine with proper understanding of pattern to classify the data set in future to predict the result.

ISSN: 2278 – 7798

Steps To Give Test Set For Prediction

- First we have to give training set to WEKA but without applying filter
- Now in the classify tab in WEKA we have to select "FilteredClassifier" in which we can select the classifier to be "RandomForest" and select filter to be "StringToWordVector" filter.
- Supply the test set for prediction
- So as per the result we obtained it gave us classification of dataset with predicted "yes or No" results.



Thus Weka provides a predicted result for given data set, but the classification will not always be correct since the machine work on a pattern so it will always have a glitch.

V. DISCUSSION

While these results are specific to the Twitter, one couldeasily see how they can be extended to other domains. Thehighly effective means of classifying tweets without utilizingthe corresponding keyword is an important discovery as itmeans that real improvement upon such systems are easilywithin reach. We believe research like this will help demonstrate the effectiveness of machine learning while at the same time helping to demystify the concept as a whole by showinglearn, straightforward applications that have been overlooked.The previous section laid out a number of positive findings,but one somewhat surprising result was that the use of aclean data set did not have a greater impact on precisionand in fact reduced recall. But what ismore important in this context, precision or recall? The answermight change depending on the topic at hand, both in termsof content and volume of discussion.Similarly, we were surprised at the relatively small decreasein accuracy when the public non-trending tweets were includedin the data set. We thought those tweets would be difficult toclassify but the precision for classifying non-trending tweetsconsistently stayed between 75% and 85%.

VI. CONCLUSION

Our research demonstrates both the benefit and feasibility of incorporating machine learning into social media in order to ease information overload and organize discourse. Twitter is an example of how machine learning is underused in social media. It also provides a fruitful test bed for machine learning approaches to easily be implemented with great effect.Despite limited use in modern machine learning research,efficient bayesian approaches like MNB can be implemented easily and are effective enough to be useful in textbased domains that are a large part of social media platforms.Further refinement of the approach outlined in this researchcould add a lot to parsing and understanding discourse ofimportant topics in social media. Tools to better handle andunderstand the massive amount of social communication onthe Internet are essential as networks like Twitter and Facebook grow and machine learning has a big role to play in thedevelopment of those tools.

REFERENCES

[1] Michael Brennan, Rachel Greenstadt, "Coalescing Twitter Trends: The Under-Utilization of Machine Learning in Social Media," 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing .

[2] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval anddecision-making from the twitter trends knowledge base," inProceedingof the 2nd ACM workshop on Social web search and mining. New York, NY, USA: ACM, 2009, pp. 1–8.

[3] Ana C.E.S Lima,Leandro N.de Castro, "Automatic Sentiment Ananlysis of Twitter messages", 978-1-4673-4794-5/12/,2012 IEEE.

[4] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogswith topic models," inInternational AAAI Conference on Weblogs andSocial Media, 2010.

[5] A. Java, X. Song, T. Finin, and B. Tsen, "Why We Twitter: An Analysis of a Microblogging Community," in 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis: Springer-Verlag, 2009, pp. 118-138.

[6] E. Mischaud, "Twitter: Expressions of the Whole Self," Master's Thesis. Department of Media and Communications (Media@LSE)., University of London, 2007.

[7] O. Phelan, K. McCarthy, and B. Smyth, "Using twitter to recommend real-time topical news," inRecSys, L. D. Bergman, A. Tuzhilin,R. Burke, A. Felfernig, and L. Schmidt-Thieme, Eds. ACM, 2009, pp.385–388.

[8] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi,"Eddi: interactive topic-based browsing of social status streams." inACM Symposium on User Interface Software and Technology, 2010, pp. 303–312.

[9] WEKA ," http://www.cs.waikato.ac.nz/ml/weka/".

[10] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in Neural Information Processing Systems, 2009.

[11] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation,"Journal of Machine Learning Research, vol. 3, January 2003.

[12] The Stanford Natural Language Processing Group," http://nlp.stanford.edu/index.shtml".

ISSN: 2278 – 7798