International Journal of Electronics Engineering (ISSN: 0973-7383
Volume 11 • Issue 1 pp. 965-970 Jan 2019-June 2019 www.csjournals.com

Sentiment Analysis using Decision Tree

Meenu, Sunila Godara
Guru Jambheshwar University of Science & Technology, Hisar, Haryana.

Abstract: Sentiment analysis is a procedure of characterization of a content that might be positive or negative. Decision Tree, which is a machine learning technique is used for Sentiment analysis using Twitter dataset in this paper. Performance is compared through accuracy, precision, recall and F measure using 2gram, 3gram and 4 gram features.
Keywords: Sentiment Analysis, Machine learning, Decision Tree, n-gram features.

I. INTRODUCTION

The expression "Feeling Analysis" itself portrays that it is examination of the different slants communicated by people over the web, or the conclusions of/input given by clients to different business associations. A basic model in our everyday life ‘where assessment examination comes in to picture is, the point at which you search for the film audits before watching it, there are committed methods accessible just to investigate the motion picture surveys [1] . On a more extensive note, nostalgic investigation or conclusion mining use information mining and common language preparing (NLP) procedures to find, recover and distil data and sentiments from the World Wide Web's immense printed data. [2] Sentiment examination allows us to follow demeanors and emotions on the web. Entities create blog entries, remarks, analyses and tweets pretty much a wide range of various themes. We can follow items, trademarks and personages for instance and decide if they are gotten definitely or adversely on the web. We can examine: certainties "the antique was sold at a lot more affluent rate than assessed” sentiments "the initial segment of the motion picture was superior to the second one". Fig 1 shows tasks of sentiment analysis process.

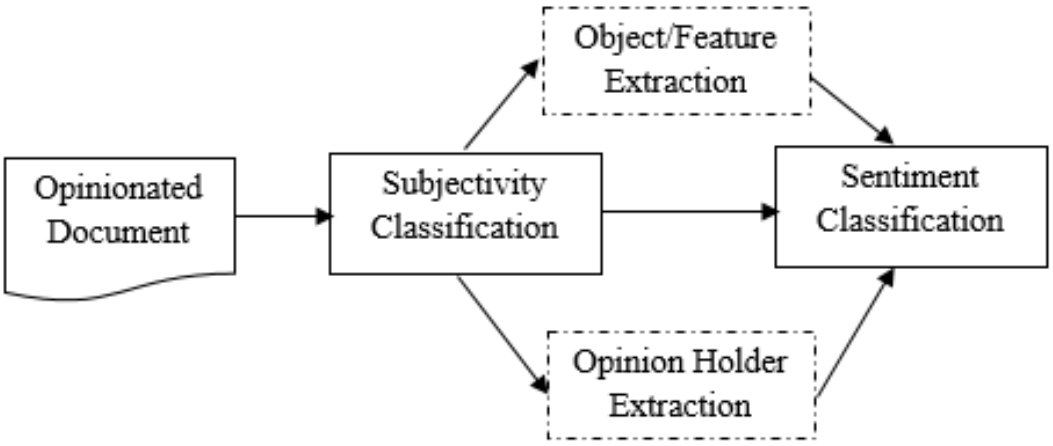

Figure 1: Tasks of Sentiment Analysis

II. RELATED WORK

Abu-Nimeh et al. [3] used 2889 phishing and genuine messages to analyze the Logistic Regression , Classification and Regression Trees , Bayesian Additive Regression Trees, Support Vector Machines , Random Forests (RF), and Neural Networks using precision,recall and accuracy parameters.

Kolariet al. [4] concluded the use of SVM to identify slogs. Evaluation of various models and their utility to log web frameworks that utilized web indexes was performed.

Crawford et al. [5] performed survey of spam discovery and the execution of various techniques for arrangement and location of audit spam. The larger part focused on administered learning techniques, which used information with regard to online audit spam. Research on techniques for Big Data are of interest, since there are a big number of online surveys till date. But few papers were discovered that reviewed the impacts of Big Data investigation for audit spam identification.

Wang et al. [6] proposed machine learning technique to deal with the spam. Three chart based highlights are used to encourage the spam boot discovery for example, removed the quantity of companions and quantity of supporters, to investigate the devotee and companion connections among various clients on Twitter. Three highlights are likewise extracted from client's latest 20 tweets. A genuine informational collection is gathered from Twitter's open accessible data utilizing two different techniques. Assessment tests demonstrate that the location of framework is efficient and precise to find spam bots in Twitter.

Viktorov et al.[7] proposed different grouping calculations of Naïve Bayes, Decision Tree , Logistic Regression, Classification and Regression Trees and Sequential Minimal Optimization (SMO). Framework was proposed to recognize the phishing messages in a coordinating between the managed and unsupervised method. Further, the study compared the guidebook and automatic feature selection groups for the Email.

P.Rohini et al. [8] concluded that the most interesting species of Internet fraud is Phishing. Email Phishing is an activity in which a web link was used to access confidential information of users such as passwords, account details. The email will be named phishing email. Different Anti phishing Mechanisms and instruments were used for client's assurance against this deceitful demonstration by utilizing heuristics technique and machine learning calculation by (SVM) bolster vector machine classifier. The phishing issue is very powerful and no single arrangement exists to moderate every one of the vulnerabilities successfully.

III. Decision Tree FOR TWITTER SENTIMENT ANALYSIS

Decision tree algorithms partitioned the training set into smaller subsets recursively as the tree is being built [9]. Decision tree consists of following nodes:
Decision node: this node indicates decision to be made.
Leaf node: shows final outcome of the decision path whether it is a spam or ham
Branch: each branch indicate possible outcome.
Each leaf of the tree is marked with a class and a supported choice tree is an outfit learning strategy in which the second tree revises for the mistakes of the main tree, the third tree rectifies for the blunders of the first and second trees, et cetera. Forecasts depend on the whole gathering of trees together that makes the expectation. For the most part, when legitimately arranged, supported choice trees are the simplest techniques with which to get top execution on a wide assortment of machine learning undertakings. In any case, they are likewise one of the more memory-concentrated students, and the present execution holds everything in memory This calculation doesn't simply build one tree, it develops the same number of as you need (100 for this situation).

Cost functions

cost function is used to find branches having group with similar function and is used for both classification and regression .by using this function we can sure about that a test data input will follow a certain path .

Stop splitting

complex tree arise the problem of over fitting because of large number of splits .so , we need to create a small tree and to know about how to stop splitting.

- We stop splitting by setting minimum number of training input on each leaf .for example we select minimum number 10 to reach a decision and ignore the other leaf that contain less than 10 values .
- Second way is to define the maximum depth .maximum depth define the path from root to leaf and this will define longest path

Pruning

To increase the performance of tree by removing the branches that have less importance we use pruning. this method will reduce the complexity of the tree ,reduce over fitting and it also increase the predictive power.

- Pruning start at leaves: this is a difficult method
- Pruning from root: this is very simple method and remove each node with most popular class in that leaf.

Two-Class Boosted Decision Tree makes a machine learning model that depends on the helped choice trees calculation. A helped choice tree is a troupe learning strategy in which the second tree redresses for the mistakes of the main tree, the third tree rectifies for the blunders of the first and second trees, et cetera. Expectations depend on the whole gathering of trees together that makes the forecast. Boosted choice trees are the most effortless techniques with which to get top execution on a wide assortment of machine learning undertakings.

Maximum number of leaves per tree, demonstrates the most extreme number of terminal hubs (leaves) that can be made in any tree. By expanding this esteem, we possibly increment the span of the tree and show signs of improvement exactness

Minimum number of tests per leaf hub, show the quantity of cases required to make any terminal hub (leaf) in a tree.

Learning rate, type a number somewhere in the range of 0 and 1 that characterizes the progression measure while learning. The learning rate decides how quick or moderates the student unites on the ideal arrangement. On the off chance that the progression measure is too enormous, we may overshoot the ideal arrangement. On the off chance that the progression measure is too little, preparing takes more time to join on the best arrangement.

quantity of trees developed, show the aggregate number of choice trees to make in the troupe. By making more choice trees, we can possibly show signs of improvement scope. This esteem additionally controls the quantity of trees showed while envisioning the prepared model.

Random number seed, alternatively type a non-negative whole number to use as the arbitrary seed esteem. Indicating a seed guarantees reproducibility crosswise over runs that have similar information and parameters. The irregular seed is set naturally to 0, which implies the underlying seed esteem is acquired from the framework clock. Progressive runs utilizing an arbitrary seed can have diverse outcomes.

Allow unknown categorical level obscure clear cut levels alternative to make a gathering for obscure qualities in the preparation and approval sets. On the off chance that we deselect this choice, the model can acknowledge just the qualities that are contained in the preparation information.

IV. RESULT AND ANALYSIS

A. Twitter Datasets

A tweet is not only a simple text message but it is a combination of text data and Meta data .These attributes are the features of tweets. They expresses the content of the tweet or what is that tweet about. The Metadata can be utilized to find out the domain of the tweet.

B. N-grams-

This will club N adjacent words in a sentence based upon N

If input is “wireless presenters for tv”

N=1 Unigram- Ouput- “wireless” , “presenters”, “for” , “tv”

N=2 Bigram- Ouput- “wireless presenters”, “presenters for” , “for tv”

N=3 Trigram - Output- “wireless presenters for” , “presenters for tv”

C. Performance Measures

Confusion matrix is a matrix representation used to show classification results. Table 1 below shows confusion matrix.

Table 1: Confusion Matrix

	Classified as sensitive	Classified as not sensitive
Actually sensitive	TP	FN
Actual not sensitive	FP	TN

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Precision= TP / (TP + FP)

International Journal of Electronics Engineering (ISSN: 0973-7383
Volume 11 • Issue 1 pp. 965-970 Jan 2019-June 2019 www.csjournals.com

Recall= TP / (TP + FN)

F Measure= 2(Precision* Recall)/(Precision+ Recall)

Accuracy is above 73% for Decision Tree using two gram , 76.81 for 3 gram and 76.81 for 4gram. Decision Tree model increases accuracy when we move from 2gram to 3gram but accuracy of 3gram and 4gram is same.
Precision and recall values also increases when we move from 2gram to 3gram but remain same when we move from 3gram to 4 gram.
F Measure values are Decision Tree, 2gram,3gram and 4gram are71.99%, 74.99% and 74.99% respectively. It reveals that there is very good agreement between actual and predicted class values.

Table.2 Classification methods based on accuracy, precision, recall, and F-measure

Classification Method	**Accuracy**	**Precision**	**recall**	**F-score**
Decision Tree 2 gram	73.24	71.22	72.71	71.99
Decision Tree 3 gram	76.81	74.57	75.89	74.99
Decision Tree 4 gram	76.81	74.57	75.99	74.99

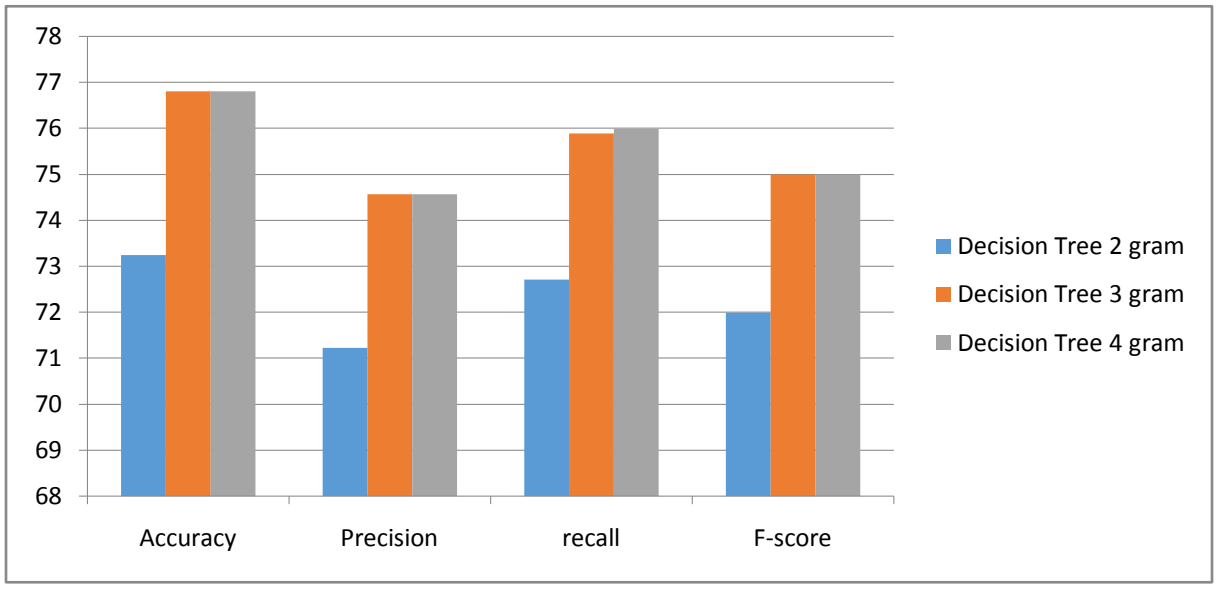

Fig. 2: Comparison of various performance measures

V. CONCLUSION

Decision tree is used for Sentiment analysis in this paper. 2 gram, 3 gram and 4 gram are used to analyze Twitter dataset. Our study illustrated that Decision tree technique using 3 gram features comes out to be most excellent classifier among others for Sentiment analysis.

REFERENCES

1 Heredia, B., Khoshgoftaar, T. M., Prusa, J., & Crawford, M. (2016, November). Integrating multiple data sources to enhance sentiment prediction. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)* (pp. 285-291). IEEE.

2 Bharti, O., & Malhotra, M. M. (2016). SENTIMENT ANALYSIS ON TWITTER DATA.

3 Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "An examination of machine learning procedures for phishing identification." *In Proceedings of the counter phishing working gatherings second yearly eCrime analysts summit*, pp. 60-69 , 2007.

4 Kolari, Pranam, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. "Distinguishing spam writes: A machine learning approach." *In AAAI*, vol. 6, pp. 1351-1356. 2006.

5 Crawford,Michael, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. "*Overview of audit spam location utilizing machine learning systems." Journal of Big Data 2*, no. 1: 23,2015

6 Wang, Alex Hai. "Identifying spam bots in online long range interpersonal communication locales: a machine learning approach." *In IFIP Annual Conference on Data and Applications Security and Privacy,. Springer, Berlin, Heidelberg*, pp. 335-342, 2010.

7 Viktorov, Oleg. "Distinguishing Phishing Emails Using Machine Learning Techniques." *PhD diss., Middle East University*, 2017.

8 P.Rohini , K.Ramya " Phishing Email Filtering Techniques A Survey", *International Journal of Computer Trends and Technology (IJCTT)* – volume 17 number 1 – Nov 2014.

9 Jagatic, Tom N., Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. "Social phishing." *Communications of the ACM 50*, vol no. 10 pp.94-100,2007.