Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis

Article *in* International Journal of Knowledge Discovery in Bioinformatics · March 2017

DOI: 10.4018/IJKDB.2017010103

4 authors, including:

Prayag Tiwari
Aalto University
76 PUBLICATIONS **1,218** CITATIONS

SEE PROFILE

Brojo Kishore Mishra
GIET University
164 PUBLICATIONS **534** CITATIONS

SEE PROFILE

Vivek Kumar
Università degli studi di Cagliari
33 PUBLICATIONS **226** CITATIONS

SEE PROFILE

International Journal of Knowledge Discovery in Bioinformatics
Volume 7 • Issue 1 • January-June 2017

Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis

Prayag Tiwari, National University of Science and Technology MISiS, Department of Computer Science and Engineering, Moscow, Russia

Brojo Kishore Kishore Mishra, C. V. Raman College of Engineering, Department of IT, Bhubaneswar, India

Sachin Kumar, Indian Institute of Technology Roorkee, Center for Transportation Systems, Roorkee, India

Vivek Kumar, National University of Science and Technology MISiS, Department of Computer Science and Engineering, Moscow, Russia

ABSTRACT

Sentiment Analysis intends to get the basic perspective of the content, which may be anything that holds a subjective supposition, for example, an online audit, Comments on Blog posts, film rating and so forth. These surveys and websites might be characterized into various extremity gatherings, for example, negative, positive, and unbiased keeping in mind the end goal to concentrate data from the info dataset. Supervised machine learning strategies group these reviews. In this paper, three distinctive machine learning calculations, for example, Support Vector Machine (SVM), Maximum Entropy (ME) and Naive Bayes (NB), have been considered for the arrangement of human conclusions. The exactness of various strategies is basically inspected keeping in mind the end goal to get to their execution on the premise of parameters, e.g. accuracy, review, f-measure, and precision.

KEYWORDS

ME, NB, n-gram, Rotten Tomatoes, Sentiment Analysis, SVM

INTRODUCTION

The Internet has altered the way people express their points of view. It is now done through the help of blog entries, online gatherings, item audit sites and so on. People rely on upon this client made dataset. When some person needs to buy a thing, they, as a rule, need to know its reviews through online before taking a decision. The measure of customer made dataset is excessively broad for a typical customer, making it impossible to examine. So, to computerize this, distinctive supposition analysis procedures are utilized. Sentiment analysis, otherwise called opinion mining, dissects individuals' opinion and additionally feelings towards datasets, for example, items, associations, and their related attributes. Machine learning proposal makes use of a planning set to add to a supposition classifier those gatherings suspicions. Sentiment analysis (Liu, 2012) is seen to be done in three distinct

DOI: 10.4018/IJKDB.2017010103

levels, for example, aspect level, document level and sentence level. Document level characterizes whether the record's opinion is negative, neutral or positive. Sentence level figures out if the sentence communicates any negative, positive or neutral opinion.

There is generally two types of machine learning techniques (Han et al., 2006) which has been used more often in sentiment analysis are unsupervised learning and supervised learning method. In supervised learning, we are provided a dataset and already having idea that what and how our output would look like and the idea that there is a relationship between the output and input. On the other hand, unsupervised learning (Kumar and Toshniwal, 2016) enables us to get problems with having little or do not have an idea that what and how our results supposed to look like. We can obtain structure from data where we don't know the effect of the variables.

The film reviews are generally in the text format and not structured in nature. Therefore, the stop words and other undesirable data are expelled from the reviews for further investigation. These frameworks are then offered a contribution to many machine learning methods for the arrangement of the surveys. Distinctive parameters are then used to assess the execution of the machine learning calculations.

The primary commitment of the paper can be expressed as takes after:

- There are many different kinds of machine learning techniques has been suggested to classify the film reviews of Rotten Tomatoes dataset using n-gram method viz., Bigram, Unigram, Trigram, an amalgamation of bigram and trigram, unigram and bigram and unigram and bigram and trigram.
- There are three machine learning techniques which SVM, NB and ME for purpose of classification by the help of n-gram proposal.
- The implementation of machine learning methods is estimated with the help of variables like recall, accuracy, precision and f-measure. The output acquired in this work demonstrates the better accuracy by comparing by other research works.

LITERATURE SURVEY

The literature review on the sentiment analysis shows the good research has been done by the various researchers based on sentiment analysis on document level.

In this paper was suggested diverse multi-mark order on sentiment analysis (Liu and Chen, 2015). They have utilized eleven multilevel characterization method seemed at on two smaller scale blog dataset furthermore eight distinctive assessment networks for examination. Aside from that, they have additionally utilized three distinctive sentiment lexicon for multi-level grouping. As indicated by the researcher, the multi-name arrangement handle plays out the undertaking basically in two stages i.e., issue change and calculation adjustment (Zhang and Zhou, 2007). In issue change stage, the issue is changed into different single-name issues. Amid preparing stage, the framework gains from these changed single mark information, and in the testing stage, the educated classifier makes expectation at a solitary name and after that makes an interpretation of it to several names. In calculation adaption, the information is changed according to the prerequisite of the calculation.

Another researcher told regarding on Overall Opinion Polarity (OvOp) idea utilizing machine learning calculations, for example, NB and Markov model for grouping (Salvetti et al., 2004). In this paper, the hypernym gave by the help of wordnet and Part Of Speech (POS) label goes about as lexical channel for order. Their trial demonstrates that the outcome got by WordNet channel is less exact in correlation with that of POS channel. In the field of OvOp, precision is given more significance in the examination with that of review. In their paper, the creators displayed a framework where they

rank audits taking into account capacity of likelihood. As per them, their approach indicates better result if there should be an occurrence of web information.

Another researcher gave the characterization of Chinese comments in light of word2vec and SVM (Zhang et al., 2015). Their approach depends on two sections. In initial segment, they have utilized word2vec apparatus to group comparable elements keeping in mind the end goal to catch the semantic components in chose space. At that point in second part, the dictionary based and POS based component determination approach is received to produce the preparation information. Word2vec device receives Continuous Bag-of-Words (CBOW) demonstrate and constant skip-gram model to take in the vector representation of words (Mikolov et al., 2013). SVM perf is a usage of SVM for multi-variate execution measures, which takes after an alternative auxiliary definition of SVM streamlining issue for double characterization (Joachims, 2006).

The relative investigation in light of results got utilizing ace postured way to deal with that of different writings utilizing Rotten Tomatoes dataset and n-gram methodologies appear in Table 3 have utilized machine learning calculation viz., SVM, NB, ME technique utilizing n-gram approach of unigram, bi-gram and blend of unigram and bigram (Salvetti et al., 2004). Furthermore, Beineke et. al., (2003) have actualized the NB strategy for order; yet just the unigram approach is utilized for grouping. Mullen and Collier, have proposed SVM strategy for order; with unigram approach as it were. Matsumoto et. al., (2013) executed the SVM for arrangements and utilized the unigram, bigram, and amalgamation of both unigram and bigram for characterization. In this presented paper, four distinct calculations viz., NB, ME technique, SVM utilizing n-gram approaches like uni-gram, trigram, bigram, bigram+trigram, unigram+bigram, and un-igram+bigram+trigram are completed. Result acquired in the present approach is seen to be superior to the outcome accessible in the writing where both Rotten Tomatoes dataset and n-gram approach are utilized.

MATERIALS AND METHODS

Order of sentiments might be sorted into two sorts, i.e. multi-class sentiment analysis and binary sentiment analysis (Tang et al., 2009). In multi-class sentiment analysis, every archive d_i is classified as a name in C^*, where C^*= {positive, strongly positive, neural, negative, solid negative}. In double classification sort, every record d_i in D, where D = $\{d_1, d_2, \ldots, d_n\}$ is classified as a name C, where C = {Positive, Negative} is a predefined class set. It is seen in the writing review, that a decent number of writers have connected twofold classification technique for sentiment analysis.

The film reviews gave by the commentators is for the most part in the content organization; however, for classification of the sentiment of the surveys utilizing the machine learning calculations, numerical lattices are required. Along these lines, the errand of change of content information in audits into numerical frameworks are completed utilizing diverse strategies, for example, Count Vectorizer, and TF-IDF have been connected to change the content report into a numerical vector, which is then considered as a contribution to regulated machine learning calculation.

IMPORTANCE OF MACHINE LEARNING METHOD

When supervised machine learning algorithms are considered for classification purpose, the info dataset is sought to be a marked one. In this study, distinctive directed learning strategies are connected for classification reason, for example, SVM, NB, ME, and n-gram method.

Support Vector Machine (SVM)

SVMs are binary linear classifiers. In preparing, they make a hyperplane between two classes of information, where the hyperplane amplifies the edge between the two classes. The fundamental supposition that we have made about our information is that it is straightly divisible. Since our

information is content based, one illustration that could bring about non-straight distinctness is mockery: assume an analyst uses a "positive" word wryly.

For a following training data set with having labeled pair (x_i, y_i), $i = 1, 2, \ldots$where $x_i \in R^n$ and y $\in \{1, -1\}^l$, the SVM method need to resolve the given optimization problem, which can be explained as:

$$\min \ \underline{1}2W_TW + c\sum\nolimits_{i=1}^{1}\xi \qquad i\, w, b, \xi$$

$$\textit{subject to } \ y_i\left(w^T\varphi\left(X_i\right)+b\right) \geq 1-\xi_i, \ ^{(5)}\ \xi_i \geq 0$$

where "W" is the weight parameter allowed to factors, ξ is the slack or mistake rectification included and "C" is the regularization figure. Since the goal of the issue is to minimize "$W_TW + c\sum_{i=1}^{1}\xi$" where estimation of "$y_i\left(w^T\varphi\left(X_i\right)+b\right)$" supposed to be greater than "$1-\xi_i$". Also, the estimation of "ξ" is thought to be little i.e., almost equivalent to 0. Here preparing vector "x_i" is mapped to upper dimensional space by "φ". Since SVM requires a contribution to the type of a vector of numbers, the audits of content record for classification should be changed over to numeric esteem. After the content record is changed over to numeric vector, it might experience a scaling procedure, which deals with the vectors and keeps them in the scope of [1, 0].

Naive Bayes (NB)

This strategy is utilized for both classification and in addition preparing purposes. This is a probabilistic classifier strategy taking into account Bayes' hypothesis. In this paper, multinomial NB classification procedure is utilized. The multinomial model considers word recurrence data in the archive for analysis, where a report is thought to be a requested arrangement of words acquired from vocabulary 'V'. In this manner, NB strategies are an arrangement of administered learning calculations in light of applying Bayes' hypothesis with the "gullible" suspicion of autonomy between each combine of components. Given a class variable y and a reliant element vector x_1 through x_n, Bayes' hypothesis expresses the accompanying relationship:

$$P\left(y \mid x_1,\ldots..x_n\right) = \frac{P\left(y\right)P\left(x_1,\ldots\ldots\ldots\ldots..x_n \mid y\right)}{P\left(x_1 \ldots\ldots\ldots\ldots..x_n\right)}$$

By using the NB assumption that:

$$P\left(x_i \mid y, x_1 \ldots\ldots\ldots, x_{i-1}, \ldots\ldots\ldots, x_n\right) = P\left(x_i \mid y\right)$$

for all i, this relationship is streamlined to:

$$P\left(y \mid x_1,\ldots..x_n\right) = \frac{P\left(y\right)\prod_{i=1}^{n}\mathrm{P}\left(x_i \mid y\right)}{P\left(x_1 \ldots\ldots\ldots\ldots..x_n\right)}$$

Since $P\left(x_1, - - - - -, x_n\right)$ is steady given the information, we can utilize the accompanying classification run the show:

$$P\left(y \mid x_1,.....x_n\right) \propto P\left(y\right)\prod_{i=1}^{n}\mathrm{P}\left(x_i|y\right)$$

$$\hat{y}=arg\,max= P\left(y\right)\prod_{i=1}^{n}\mathrm{P}\left(x_i \mid y\right)$$

What's more, we can utilize Maximum A Posteriori (MAP) estimation to gauge P(y) and P (x_i | y); the previous is then the relative recurrence of class y in the preparation set. Regardless of their clearly over-improved suppositions, credulous Bayes classifiers have worked great in some genuine circumstances, broadly record classification and spam separating. They require a little measure of preparing information to gauge the fundamental parameters.

Maximum Entropy (ME)

In this strategy, the preparation information is utilized to set a limitation on restrictive dispersion Nigam, Lafferty, and McCallum (1999). Every imperative is utilized to express attributes of preparing information. Maximum entropy (ME) esteem as far as exponential capacity can be communicated as:

$$P_{ME}\left(c \mid d\right) = 1/z\left(d\right)\left\{\exp\left(\Sigma\lambda_{i,c}f_{i,c}\left(d,c\right)\right)\right\}$$

where P_{ME} (c | d) alludes to likelihood of document "d" having a place with class 'c', $f_{i,c}$ (d, c) is the element/class work for attribute f_i and class c, $\lambda_{i,c}$ is the parameter to be evaluated and Z(d) is the normalizing variable. Keeping in mind the end goal to utilize ME, an arrangement of elements is should have been chosen. For content arrangement reason, word considers are considered characterized. Attribute/class capacity can be instantiated as takes after:

$$f_{i,c}\left(d,c\right) = \begin{cases} 0 & if\ \ c \neq c` \\ \dfrac{N\left(d,i\right)}{N\left(d\right)} & otherwise \end{cases}$$

where $f_{i,c}$ (d, c) alludes to characterize in word-class blend in class "c" and archive 'd', N (d, i) speaks to the event of characterizing "i" in the record "d" and "N(d)" number of words in'. According to the expression, if a word happens oftentimes in a class, the heaviness of word-class match gets to be higher in contrast with different sets. These most astounding recurrence word-class sets are considered for a grouping reason.

Model of n-gram

It is a strategy for checking "n" persistent words or sounds from a given grouping of content or discourse. This model predicts the following thing in a grouping. In conclusion examination, the n-gram display examines the opinion of the content or archive. Unigram alludes to n-gram of size one, Bigram alludes to n-gram of size two, Trigram alludes to n-gram of size three. Higher n-gram alludes to four-gram, five-gram, six-gram and so on. The n-gram strategy can be clarified utilizing taking after the case: An ordinary case of a sentence might be considered as "This is not a very good film".

- This is unigram "'This', 'is', 'not', 'a', 'good', 'film'" in which single word is taken into consideration.

- This is diagram "'This is', 'not a', 'good film'" in which double word is taken into consideration.
- This is trigram "'This is not', 'a good film'" in which three words are taken into consideration and this process further gets proceed for more than three.

Performance Assessment Parameter

The parameters accommodating to assess execution of supervised machine learning calculation depends on the component from a lattice known as perplexity network or possibility Table 1. It is utilized as a part of regulated machine learning calculation to help in surveying execution of any calculation. From arrangement perspective, terms, for example, "True Positive (TP)", "False Positive (FP)", "True Negative (TN)", "False Negative (FN)" are utilized to think about the mark of classes in this lattice as appeared in Table 2 (Mouthami et al., 2013). True Positive speaks to the quantity of audits those are sure furthermore delegated positive by the classifier, whereas False Positive demonstrates positive surveys, however, classifier does not group it as positive. Additionally, True Negative speaks to the surveys which are negative likewise delegated negative by the classifier, whereas False Negative will be negative audits yet classifier does not characterize it as negative. In view of the qualities got from perplexity grid, different parameters, for example, "exactness", "review", "f-measure", and "precision" are discovered for assessing execution of any classifier.

- **Exactness:** It quantifies the precision of the classifier output. It is the proportion of a number of illustrations accurately named as positive to an aggregate number of emphatically characterized case:

$$Exactness = T\,P/(TP + F\text{-}P)$$

- **Revoke:** It quantifies the culmination of the classifier result. It is the proportion of an aggregate number of the decidedly named case to aggregate illustrations which are genuinely positive:

$$Revoke = T\,P/(T\,P + F\,N)$$

Table 1. Confusion matrix

	Correct Labels	
	Negative	**Positive**
Negative	TN (True negative)	FN (False negative)
Positive	FP (False positive)	TP (True positive)

Table 2. Matrix generated under CountVectorizer scheme

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sentence 1	1	1	1	0	0
Sentence 2	1	1	0	1	0
Sentence 3	1	1	0	0	1

Table 3. Results

Method		Salvetti et.al.	Beineke et.al.	Mullen & Collier	Matsumoto et.al.	Pang et.al.	Proposed approach
SVM	Unigram	×	×	86.0	83.7	72.9	87.53
	Bigram	×	×	×	80.4	77.1	85.60
	Trigram	×	×	×	×	×	81.42
	Fourgram	×	×	×	×	×	72.13
NB	Unigram	79.5	65.9	×	×	81.0	85.71
	Bigram	×	×	×	×	77.3	84.06
	Trigram	×	×	×	×	×	80.39
	Fourgram	×	×	×	×	×	69.07
	Unigram + Bigram	×	×	×	×	80.6	87.08
	Bigram + Trigram	×	×	×	×	×	84.68
	Unigram + Bigram+ Trigram	×	×	×	×	×	86.05
Unigram+Bigram	Bigram + Trigram	×	×	×	×	×	84.6
	Unigram + Bigram+ Trigram	×	×	×	×	×	89.65
ME	Unigram	×	×	×	×	80.4	89.63
	Bigram	×	×	×	×	77.4	84.01
	Trigram	×	×	×	×	×	79.45
	Fourgram	×	×	×	×	×	68.02
	Unigram + Bigram	×	×	×	×	80.8	89.64
	Bigram + Trigram	×	×	×	×	×	83.09
	Unigram + Bigram+ Trigram	×	×	×	×	×	84.65

× means this is not taken by the author.

- **F-measure:** It is the consonant mean of exactness and review. It is required to streamline the framework towards either accuracy or review, which have more impact on the definite result:

F-measure= (2*Exactness*revoke) / (Exactness + revoke)

- **Precision:** It is the most widely recognized measure of arrangement process. It can be ascertained as the proportion of effectively grouped case to an aggregate number of cases:

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

Dataset Used

Rotten Tomatoes is quite a prominent website dedicated to news reviews, film reviews and details; it is mostly known as a film review aggregator. Rotten Tomatoes has assembled an express path for seeing if a motion picture sucks or will change your life. Individuals adore Rotten Tomatoes since it's anything but difficult to utilize. The staff determines for every review whether it is negative or positive. Where positive stands for recent, marked by a minor icon of a red tomato) and Negative is signified by rotten, marked by a minor icon of a green splattered tomato. It comprises of 8000 emphatically marked test surveys, and 8000 decidedly named prepare audits. Thus, there are 8000 negative marked test surveys, 8000 negative named prepare audits. Aside from marked supervised dataset, an unsupervised dataset is likewise present with 32000 unlabeled surveys.

SUGGESTED APPROACH

The survey of Rotten Tomatoes dataset is handled to expel the stop words and undesirable data from the dataset. The printed information is then changed to a network of number utilizing vectorization methods. Assist, preparing of the dataset is done utilizing machine learning calculation. Ventures of the approach are talked about

Level 1

The Rotten Tomatoes dataset comprising of 8000 positive and 8000 negative surveys for preparing furthermore 8000 positive and 8000 negative audits for testing Rotten Tomatoes is into consideration.

Level 2

The text audits once in a while comprise of crazy information, which should be evacuated before considered for the arrangement. The typically recognized ridiculous information is as follows:

- **Stop Words:** They do not have any important duty in sentiment analysis;
- **Special and Numeric Character:** In the text audits, it is frequently watched that there are distinctive numeric (1, 2, …, 5, 6, 7 and so forth.) and exceptional characters (#, @, %, $, and so on.) present, which don't have any impact on the investigation. In any case, they frequently make perplexity amid transformation of content document to a numeric vector.

Level 3

After the pre-handling of content surveys, they should be changed over to a network of numeric vectors. The accompanying strategies are considered for change of content record to numeric vectors:

- **CounterVectorizer:** It changes over the content surveys into a framework of token checks. It executes both tokenizations also, event tallying. The yield framework acquired after this procedure is a scanty network. A case is considered to clarify the means of figuring components of the framework which helps in enhancing the comprehend capacity. Assume, three unique reports containing taking after sentences are taken for investigation:

Sentence 1st: "Film is good"
Sentence 2nd: "Film is terrible"
Sentence 3rd: "Film is ok"

A lattice might be framed with various qualities for its components estimate 4 ∗6, as there exist 3 archives and 5 particular elements. In the framework given in Table 1, the

components are doled out with estimation of '1', if the element is available or else if there should arise an occurrence of the nonappearance of any element, the component is doled out with esteem '0'.

- **TF-IDF:** It recommends the significance of the word to the record and entire corpus. Term recurrence educates about the recurrence of a word in a record and IDF illuminates about the recurrence of the specific word in the entire corpus. A case might be considered to enhance understandability. In the event that a motion picture survey contains 500 words wherein "Wonderful" shows up 5 times. The term recurrence (i.e., TF) esteem for "Marvelous" might be found as 5/500 = 0.01. Once more, assume there are 1 million audits in the corpus and "Wonderful" shows up 500 times in entire corpus. At that point, the backward record recurrence (i.e., IDF) esteem is computed as log(250000/500) = 2.7. In this way, the TF-IDF esteem is ascertained as $0.01 * 2.7 = 0.027$.

Level 4

After the text audits are changed over to framework of numbers, these lattices are considered as contribution for the accompanying four diverse directed machine learning calculations for:

- **SVM:** Data are broken down and choice limits are characterized by having hyper planes. In two class case, the hyper plane isolates the record vector of one class from different classes, where the detachment is kept up to be huge as could be expected under the circumstances;
- **NB:** Using probabilistic classifier and example taking in, the arrangement of archives is characterized;
- **ME:** The preparation information is utilized to set requirement on contingent appropriation. Each imperative is utilized to express attributes of preparing the information. These requirements then are utilized for testing the information.

Level 5

As said in level 1, the motion picture surveys of Rotten Tomatoes dataset is considered for examination, utilizing the machine learning calculations talked about in level 4. At that point, diverse variety of the n-gram strategies i.e., unigram, bigram, trigram, unigram + bigram, unigram + trigram, and unigram + bigram + trigram are connected to acquire the outcome.

Level 6

The outcomes acquired from this investigation are contrasted and the outcomes accessible in different literature.

IMPLEMENTATION

- **The Importance of ME Method:** As ME calculation in view of contingent dispersion and word-class match group the audit, unigram technique which considers single word for examination, gives the best result in correlation with different strategies. In both bigram and trigram techniques, the negative or positive extremity word seem more than once; along these lines, influencing the grouping result. The bi-gram and trigram techniques when joined with unigram and between themselves, the precision estimations of different blends are seen to below;
- **The Importance of SVM Method:** As SVM technique is a non-probabilistic direct classifier and trains model to discover hyper-plane keeping in mind the end goal to isolate the dataset, the unigram show which breaks down single words for investigation gives a better result. In bigram

and trigram, there exist different word blends, which, when plotted in a specific hyperplane, befuddles the classifier and along these lines, it gives a less exact result in the examination with the esteem acquired utilizing unigram. Accordingly, the less precise bigram and trigram, when consolidated with unigram and with each other likewise, give a less exact result;

- **The Importance of NB Method:** NB technique is a probabilistic strategy, where the elements are autonomous of each other. Consequently, when the investigation is completed utilizing "single word (unigram)"and "twofold word (bigram)", the exactness esteem got is relatively superior to that got utilizing trigram. Be that as it may, when 'triple word (trigram)' is being considered for analysis of components, words are rehashed various times; therefore, it influences the likelihood of the archive. Accordingly, the exactness of arrangement abatements. Once more, when the trigram model is consolidated with unigram or bigram or unigram + bigram, the effect of trigram makes the precision esteem similarly low.

RESULTS AND DISCUSSION

This paper makes an endeavor to group movie surveys utilizing differently directed machine learning calculations, for example, SVM, NB, and ME. These calculations are further connected utilizing n-gram approach on Rotten Tomatoes dataset. It is observed that as the estimation of "n" in n-gram builds the order precision diminished i.e., for unigram, bigram and trigram the outcome acquired utilizing the calculation is strikingly better; however, when four-gram, five-gram grouping are completed, the estimation of accuracy reduces.

Rather than utilizing unigram and POS tag, the utilization of unigram, bigram, trigram, and their blend have demonstrated a superior result. Once more, utilization of TF-IDF and CountVectorizer methods as a blend for changing the content into the framework of numbers additionally get the estimation of precision in an enhanced way, when machine learning systems are utilized.

The present study has likewise a few constraints as said underneath:

- So as to give weight to a word, it is watched that some per-children regularly rehash the last character of the word various times, for example, "fantasticcc, goodddd". These words don't have a legitimate importance; however, they might be viewed as and further prepared to distinguish notion. In any case, this viewpoint is likewise not considered in this paper;
- The Twitter remarks are generally little in size. In this way, the professional postured approach may have a few issues while considering these surveys.

It might likewise happen that the precision esteem may enhance if a portion of the half and half machine learning systems are considered for an order of the estimation. All of the previously mentioned confinements might be considered for the future work, keeping in mind the end goal to enhance the nature of notion arrangement.

CONCLUSION

This study endeavored to classify dataset of movie reviews by utilizing different supervised machine learning techniques such as Maximum entropy (ME), Naïve bayes (NB) and Support Vector Machine (SVM) and we utilized these techniques because they are providing better results. These techniques are implemented on dataset of Rotten Tomatoes by utilizing n-gram approach. It is noticed that accuracy of classification diminished by increasing the value of 'n' in n-gram i.e., it is noted that result is better for unigram, bigram and tri-gram but accuracy decrease when observed for four-gram, five-gram, six-gram and further on.

REFERENCES

Beineke, P., Hastie, T., & Vaithyanathan, S. (2004). The sentimental factor: improving review classification via human-provided information.*Proceedings of the 42nd annual meeting on association for computational linguistics* (p. 263). Association for Computational Linguistics. doi:10.3115/1218955.1218989

Beineke, P., Trevor, H., Christopher, D. M., & Shivakumar, V. (2003), Exploring sentiment summarization. In Q. Yan, J.G. Shanahan, & J. Wiebe (Eds.), Exploring Attitude and Affect in Text: Theories and Applications. *Papers from the AAAI Spring Symposium* (pp. 4–7).

Han, J., Kamber, M., & Pei, J. (2006). Data Mining: Concepts and Techniques (2nd ed.). San Francisco, CA, USA: Morgan Kaufmann.

Joachims, T. (2006). Training linear svms in linear time.*Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 217–226). ACM. doi:10.1145/1150402.1150429

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data, 3*(13), 1–11.

Liu, B. (2012). *Sentiment analysis and opinion mining. Synthesis lectures on human language technologies*. Morgan & Claypool Publishers.

Liu, S. M., & Chen, J. H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, *42*(3), 1083–1093. doi:10.1016/j.eswa.2014.08.036

Matsumoto, D., & Hwang, H. C. (2013). Assessing cross-cultural competence: A review of available tests. *Journal of Cross-Cultural Psychology*, *44*(6), 849–873. doi:10.1177/0022022113492891

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Mouthami, K., Devi, K. N., & Bhaskaran, V. M. (2013). Sentiment analysis and classification based on textual reviews. Proceedings of the2013 international conference onInformation communication and embedded systems (ICICES) (pp. 271–276). IEEE doi:10.1109/ICICES.2013.6508366

Salvetti, F., Lewis, S., & Reichenbach, C. (2004). Automatic opinion polarity classification of movie. *Colorado research in linguistics, 17*(2).

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*(7), 10760–10773. doi:10.1016/j.eswa.2009.02.063

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). Association for Computational Linguistics.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and svm perf. *Expert Systems with Applications*, *42*(4), 1857–1863. doi:10.1016/j.eswa.2014.09.011

Zhang, M. L., & Zhou, Z. H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. doi:10.1016/j.patcog.2006.12.019

Prayag Tiwari is currently Masters student in computer science and engineering department from National University of Science and Technology MISiS Moscow, Russia and his research interest are machine learning/data mining, big data, database system

Dr. Brojo Kishore Mishra is an Associate Professor in the Department of IT and Institutional IQAC Coordinator, C. V. Raman College of Engineering (Autonomous), Bhubaneswar, Odisha, India. He is the Regional Student Coordinator (2016-17), CSI Region – IV, India. Also, he is the IEEE Day 2016 Ambassador for IEEE Kolkata Section. He has received his Ph. D. (Computer Science) from Berhampur University in 2012 and has supervised more than 08 M. Tech. thesis and currently guiding 04 Ph.D research scholars in the area of Data Mining, Opinion Mining, Soft Computing and Security. Dr. Mishra has published more than 25 research papers in international journals and conference proceedings and 3 invited book chapters. He serves as Guest Editor in IJKDB, IJSE, IJACR, and IJRSDA special issue journals and an editorial board member of many international journals. He is associated with a CSI funded research project as a Principal Investigator. He was the Regional Convener of CSI YITP 2015-16, CSI State Student Coordinator (2015-16), Jury Coordination Committee Member of All IEEE Young Engineers' Humanitarian Challenge (AIYEHUM 2015) project competition, organized by IEEE Region 10 (Asia pacific) and IEEE Day 2015 Ambassador for IEEE Kolkata section.

Sachin Kumar Agnihotri is working as a PhD candidate in Indian Institute of Technology Roorkee. His area of research is data mining/machine learning for traffic safety. He is B Tech, M Tech in Computer Science & Engineering. His area of interest are big data analytics, semantic analysis, text analytics, sentiment analysis, traffic safety etc.

Vivek Kumar is currently doing masters in computer science and engineering department from National University of Science and Technology MISiS Moscow, Russia and his research interest are image processing, data mining. for contact - vivekkumar0416@gmail.com