Polarity Detection of Sentiment Scoring Methodusing Dempstershafer Theory

M. Edison, A. Aloysius

Abstract*: Sentiment Analysis (SA) is a big task to measure the people opinion. The aim of the SA is to obtain the essential viewpoint of text, which could be opinions, blogs, reviews, online rating comments etc. Nowadays, most of the peoples are familiar for using internet to express their opinions. However, the opinions are classified in a different way like positive, negative and neutral and assigned score to the sentimentword like +1, -1, and 0 respectively. Nevertheless, the sentiment score have been assigned formally in the existing works. Some of the works the sentiment scores are assigned based on the threshold value 0.5.In that case, several existing works were applied this value for polarity detection.The limitation of the existing works given more attention to compute the score of a sentiment. Therefore, the new algorithm proposed namely Senti_Demp_Score, which performs to measure the sentiment score based on the DempsterShaper Theory (DST). The DST perform to calculatethe sentiments in a sentence and sum of all the sentiments with a category, then they converted into percentage. Concurrently, the percentage values are summed with a category and subtractedthe percentage score which gives the accuracy of the Senti_Demp_Score is 0.7756 like in percentage 77.56%. Mainly, the algorithm Senti_Demp_Score has given better solution to detect the polarity of the sentiment and measure the sentiment score in a right way.

Keywords: Sentiment analysis, Polarity, Dempster Shafer Theory, Sentiment Score.

.

I. INTRODUCTION

Sentiment analysis is a domain to study people's opinion, attitude and appraisals. Currently, microblogging is very popular messageconversation in the world. Especially, Twitter is a very familiar microblogging Social Media Network (SMN). Twitter allows only 140 characters to post comments by the users. Therefore, the users are communicating their feelings with a short communication like acronyms, emoticons, emoji's etc. However, the user can post their reviews with the different opinions. The opinions like sentiments are big task to classify the sentiments. Because the sentiments are being ambiguous data, therefore, different way to concentrate and classified data like positive, negative and neutral. In this part, DST mathematical theory has taken to measure the sentiment score. In past research, the researchers have been assigned numerical value for the sentiments, but not concentrated to assign the sentiment values in percentage. In this proposed work focussed on to determine the sentiment values in percentage. Hence, applied DST theory to calculate the sentiment scores in percentage and predict the result, which was helped to enhance the result of the sentiment scores.

II. LITERATURE REVIEW

Quan Zou et. al [1] proposed finding the proper prediction possibility threshold of a testing set. Experimental assessment was accomplished through using an established benchmark, and the effects showed that the proposed approach can efficaciously enhance prediction overall performance over extra commonly hired techniques. In widespread, the classification threshold is in reality set to 0.5, which is typically incorrect for an imbalanced classification. The drawbacks of the usage of ROC as the sole degree of imbalance in statistics classification problems. Deepak Singh et. al[2] proposed Boolean method contextual polarity of a sentence, and find out the correct contextual polarity of textual content. The contextual polarity rules are successfully placed in a sentence by the use of sentiwordnet lexicons and evaluated it. By the use of part of speech tagger to tag words and search simplest those words with polarity. However, the polarity detection method very helpful for calculating the score of a particular word like sentiment in a sentence. EsraAkbas [3] has proposed sentiment strength model to construct opinion word list using feature extraction method. Applied feature selection by using the use of sentiment lexicons to decrease the complexity and enhance the accuracy of the end result.According to Mohammed et. al [4] proposed scoring aggregation method for the prediction of the sentiment accuracy. Concurrently, aggregate methods are used for compute sentiment scores in sentence level into document rating. The proposed approach used based on the DST of evidence, which detects the polarity of the each and every sentiments. Then it predicts the overall sentiment score. The DST model has detected an individual sentence within a review.

III. BACKGROUND STUDY

The DST is a mathematical theory which helps to detect the polarity of the sentiment.

DST is calculate the each individual sentiment and compute the overall weight of a sentence.

It is natural to map sentiment strength prediction task to a binary classification.

DST describes the method to use the score of aggregation hassle. Dempster-Shafer is a theory of uncertainty that allows to quantify the diploma to which a few supply of

Revised Manuscript Received on December 22, 2018.

M. Edison, Assistant Professor Apollo Arts and Science College, Chennai – 602105.
Dr.A. Aloysius, Assistant Professor St. Joseph's College (Autonomous), Tiruchirappalli – 620 002.
aloysius1972@gmail.com

.

evidence supports a selected proposition. In truth, it's miles an opportunity to probability theory, allowing the explicit representation of lack of information and combination of proof. This theory has become firstly evolved by the means of Dempster and then prolonged via Shafer in his 1976 e-book, A Mathematical theory of evidence [5].

Dempster–Shafer theory is a generalization of the Bayesian idea of subjective theory. Perception features base ranges of belief (or self-belief, or accept as true with) for one question at the possibilities for a related query. The levels of belief itself may also or may not have the mathematical homes of probabilities; how much they fluctuate relies upon on how carefully the two questions are associated. Placed any other manner, its miles a way of representing epistemic plausibility. However, it is able to yield answers that contradict those arrived at the use of probability theory.

Dempster–Shafer theory is based on two ideas: obtaining degrees of notion for one query from subjective possibilities for an associated query, and Dempster's rule for combining such tiers of notion while they are primarily based on impartial objects of evidence. In essence, the degree of notion in a proposition depends commonly upon the wide variety of solutions (to the related questions) containing the proposition, and the subjective probability of each solution. Additionally contributing are the guidelines of aggregate that reflect widespread assumptions about the information [6].

IV. DATA PREPARATION

A. Data Collection and Pre-Processing

Data acquiring is a process of data collection, the data acquiring process concentrates on the performance analysis and the data set have been collected from the Twitter using Twitter Application Programming Interface (TAPI). A total of, 1048555 tweets were stored, in the form of comma separate value (CSV) file format. Then, the collected tweets were taken for the assessment and the total number of sentiments, acronyms and emoticons were extracted from the collected data [7].

Among the structured data, unstructured data and semi-structured data, varieties of unstructured data have been collected. The collected data have been reflected as unstructured data. Therefore, unwanted data is removed from

Emoticons	Dictionary Lexicon
:)	Happy
(:	Sad
:-)	Joy
(-:	sorrow

the data set because, the meaningless data are useless in nature [8, 9]. The pre-processing algorithmand methods are taken from the paper and applied as same procedure and pre-processed the data [10].

V. PROPOSED WORK

In this section, wefirst represent a brief description of Dempster Combinational Theory (DCT)theory after which describe the manner in which we use it on therating aggregation trouble. DCT is a concept ofuncertainty that allows to quantify the diploma to which somesource of proof supports a selected proposition. In truth,it's far an opportunity to conventional probability idea, permittingthe explicit representation of lack of knowledge and aggregate ofevidence. This concept turned into firstly developed via Dempsterand then prolonged by means of Shafer in his 1976 e book, AMathematical principle of evidence. In addition the acronyms and emoticons are collected from various sources and they have briefly discussed in the section of acronyms and emoticons with an examples.

A. Acronyms

The acronyms are collected from no-slang web site. There are 28539 words each and every word has a different distinct meaning. In this paper, the acronyms dictionary has been built manually which isfurther divided into two different dictionaries, the first one is a positive acronym dictionary and the second one is negative acronym dictionary. If the positive dictionaryhas 9580 acronyms and the negative dictionary has 7360 acronyms, then the rest of the acronyms are neutral.

The acronym dictionary is very helpful to expand the tweets and improve the overall sentiments score [11, 12]. The acronyms have ambiguous characters and different abbreviations. The example translation table is illustrated in table. 1.

Table. 1. Example Translation of Acronyms

Acronyms	Dictionary Lexicon
Asap	as soon as possible
gr8	Great
@mazing	Amazing
aprece8	Appreciate
Phab	fabulous

B. Emoticons

185 emoticons (smilies) have been collected from the web, which are mostly used by the users regularly. In this work, the emoticon dictionary has been made manually, which are divided into two different dictionaries, the first one is positive emoticon dictionary and the second one is negative emoticon dictionary. The positive dictionary has 85 emoticons and the negative dictionary has 70emoticons then the rest of the emoticons are neutral.

The emoticon dictionary is very helpful to expand the tweets and improve the overall sentiments score. The emoticons have adifferent combination of symbols as different abbreviations [13, 14, 15]. The example translation of emoticons is shown in table. 2.

Table. 2. Example Translation of Emoticons

The data extraction has explained in the previous chapter. The collected tweets are pre-processed and stored in the desired format (.txt and .csv). The pre-processing step includes stopwords removal, url removal, audio and video removals, user name, negation and stemming. The processes are briefly explained in the following subsections.

The overall process of the pre-processing and proposed approach has been divided into two phases. Figure 1 exhibits the methodology diagram of the three phases.

They are explained below.

Phase 1:Senti_Demp_Score approach has been proposed for sentiment, acronyms and emoticon handling. A mathematical equation has been framed for measuring the sentiment score to classify the tweets as positive, negative and neutral.

Phase 2: The positive, negative and neutral tweets are classified and discussed in the results section. Mainly, apply DempsterShaper Theory and measure the sentiment score based on it. The Framework for Senti_Demp_score is shown in figure 1.

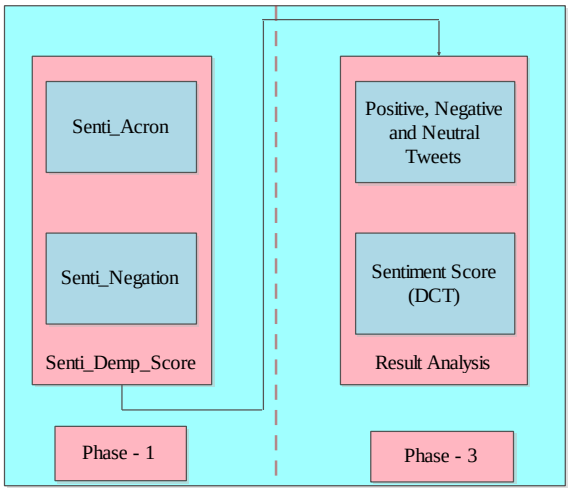

Fig 1. Framework for Senti_Demp_Score

The concept of Senti_Demp_Score approach is used to evaluate the scoring of the sentiments. The proposed scoring method is to measure the scores of the sentiment based on the value between 0 and 1. In the existing work, sentiment scores are calculated using semantic score method like positive, negative and neutral to detect the polarity and assigning the values +1, -1, and 0 respectively. The proposed work of Senti_Demp_Score uses Dempster Combinational Theory for measuring the sentiment score and detect the polarity of the sentiment in a sentence.

In the proposed measuring model (Senti_Demp_Score), computes the sentiments in a sentence and tokenizes the sentiments. The sentiments are tokenized in different category such as: positive, negative and neutral. The tokenized sentiments are summed within a sentence which are converted into percentage. Then the percentage value will be assigned for the sentiment with a category. The sum of all tokenized sentiments of the sentences are summarized using Dempster Combinational method. The overall polarity is classified into positive, negative and neutral. The Senti_Demp_Score scoring method has produced better result in classification of sentiment by detecting polarity in an efficient way. The Senti_Demp_Score algorithm is illustrated in figure

The Senti_Demp_Score algorithm is working efficiently for the classification and measure the polarity of sentiments. In the process every tweet is taken for the process of classification. Then the Senti_Demp_Score performs and focuses on isolation method for the sentiments, acronyms and emoticons. The pre-processed data has been taken for the process into Senti_Demp_Score algorithm, each and every tweet is stored into T'. Then the feature selection is applied as unigram which splits the words separately with the identification of white space. T' (T'words) $\leftarrow \sum_{i=1}^{n} unigram\,(T_i')$ the unigram word is considered as T_i' especially i which indicates unigram, the summation is calculated from word 1 to n and the sum of words are assigned into (T'words). Formerly, it checks the condition T'wordif found in a dictionary then it checks the word either acronym or emoticon, suppose the word is found then the word replaced acronym or emoticon into equivalent semantic word like “gr8” into great, “gud” into good, “5n” into fine, “:-) into happy, :-(into sad”. Else it identifies a word as acronym or emoticon but if the word is not in a dictionary then has to be inserted into the dictionary with the equivalent meaning.

The equation 1 PS is denotes the measure of the positive scores. It computes all the negative sentiments and neutral sentiments, then sum negative and neutral sentiments. After that, the summed sentiments divided into total number of sentiments like T'

$$PS \leftarrow PScore = \frac{\sum_{i=1}^{n}(Ne_i + N_i)}{T_i} \quad \text{.....}$$

Equation 1.

The equation 2 NS is denotes the measure of the negative scores. It computes all the positive sentiments and neutral sentiments, then sum positive and neutral sentiments. After that, the summed sentiments divided into total number of sentiments like T'

$$NS \leftarrow NScore = \frac{\sum_{i=1}^{n}(P_i + Ne_i)}{T_i} \quad \text{.....}$$

Equation 2.

In DST, ∅ is set be a finite set of mutually exclusive hypotheses, known as frame of discernment. $t'\sum_{j=1}^{n} Se$ is a mass function used to represent the strength of the sentence supporting a subset Se ⊆∅ based on a given sentiment score. It is a Basic Probability Assignment (BPA) to all subsets t' of ∅. The value of the function is a real number range from 0 to 1 with the following properties and it denotes in equation 3.

$$S(\varnothing) = 0\sum_{i=1}^{n}\sum_{j=1}^{n} Se = 1 \quad \text{.....}$$

Equation 3.

where $t'\sum_{j=1}^{n} Se$ is interpreted by Se. It is based on the available computation score of Se. A subset Se of ∅ is referred as focal element of a mass function S over ∅.

The fundamental operation of DS theory of evidence is a rule for the pooling of evidence from a variety of sources, known as Dempster's rule of combination. Specifically, this rule has been proposed for aggregating of sentiment score over a common sentence to obtain the sentiment of the sentence, which is calculated for the entire sentence. The total number of sentence is denoted as n as given Equation 4 and 5.

$$Se_{1,2}(S) = \frac{\sum X \cap Y = S \; Se_1(X) Se_2(Y)}{1 - \sum X \cap Y = \varnothing Se_1(X) Se_2(Y)} \quad \ldots\ldots$$

Equation 4.

$$Se_{1,2,\ldots,n}(S) = \frac{\sum \bigcap_{i=1}^{n} X_i = S \left(\prod_{j=1}^{n} Se_j(X_i) \right)}{1 - \sum \bigcap_{i=1}^{n} Se = \varnothing \left(\prod_{i=1}^{n} Se_j(X_i) \right)} \quad \ldots\ldots$$

Equation 5.

This equation PS>NS then T'class(i) ← Positive checks the condition to see if PSis greater than NS negative then it stores the class as positive which is assigned into the T'class(i) then NS> PS then T'class(i) ← Negative the equation checks the condition NSto see if it is greater than PSpositive which is assigned into T'class(i) then stored class as negative and the rest of the class is neutral. Finally, T'PS> T'NSthe condition checks to see if T'PSis greater than the negative, then it stores the polarity class into T'PSand it reflects a positive impact on the T'. If the condition T'NS> T'PS, T'NSis greater than T'PSthen it stores the polarity class into T'NSand it is reflected as negative impact, rest of the condition reflects a neutral impact.The Senti_Demp_Score has played an effective role in checking whether the sentiments are at sentence level.

6. Result and Discussion

Performance Evaluation (Evaluation Matrix)

Performance evaluation metrics named as confusion matrix is evaluated after the classification results. Plenty ofEvaluation Matrix's (EM) exist to measure the accuracy of the SA. The most commonly used EM'sarePrecision, Recall, F-measure and Accuracy of the proposed approaches. These common terminologies are measured based on the valueslike true positive (tp), true negative (tn), false positive (fp) and false negative (fn). The predicted positive and negative instance is illustrated in Table. 3.

	Predicted Positives	Predicted Negatives
Actual Positive Instances	Number of True Positive	Number of False Negative
Actual Negative Instances	Number of False Positive	Number of True Negative

True Positive (TP) – Correctly Identified
True Negative (TN) – Correctly Rejected
False Positive (FP) – Incorrectly Identified
False Negative (FN) – Incorrectly Rejected

Table. 3. Confusion Matrix Terminology.

The Table. 4 shows the confusion matrix result of the Senti_Demp_Score approach. This result is analyzed and interpreted in the following Figure. 3.

Table 4. Confusion Matrix Result of the Senti_Demp_Score

	Recall	Precision	F-Measure	Accuracy
Senti_Demp_Score	0.9038	0.9425	0.9346	0.7756

The Senti_Demp_Score approach confusion matrix result has been detected and visualized in the Figure. 3. The Figure. 3 is represented as bar chart with a different colours left to right. The red colour bar indicates recall metrics result, the green colour bar indicates precision metrics result. The blue colour bar indicates F-measure metrics result, and then finally yellow colour bar indicates accuracy metrics result.

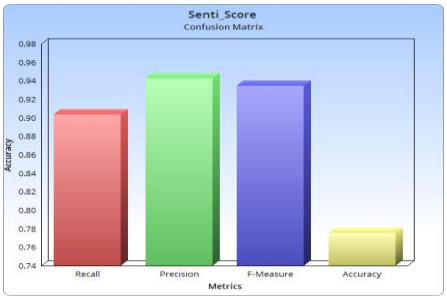

Figure 3. Confusion Matrix Result of the Senti_Demp_Score

The Table. 5 is illustrated comparison result of the proposed work with existing work. This result is analyzed and interpreted in the following Figure. 4.

Table.5. Comparison Result of the Proposed Work with Existing Work

S. NO.	Authors	Year	Accuracy
1	Edison et al. (Senti_Acron)	2017	68.75%
2	Esra et al.	2017	70.94%
3	Mohammad et al.	2014	72%

4	Edison et al. (Senit_Negation)	2017	73.27%
5	Edison et al. (Proposed Work) Senti_Demp_Score	2018	77.56%

The comparison result has been visualized in the Figure 4. The Figure.4 indicates a bar chart with different colours from left to right. The red colour bar indicates Edison et al. (Senti_Acron) projected result, the green colour bar indicatesEsra et al. projected result. The blue colour bar is represented as Mohammad et al.projected result. The yellow colour bar indicates Edison et al. (Senti_Negation) projected result and then finally violet colour bar indicates proposed work (Senti_Demp_Score) result. Compare with existing works the proposed work result has been enhanced 4%.

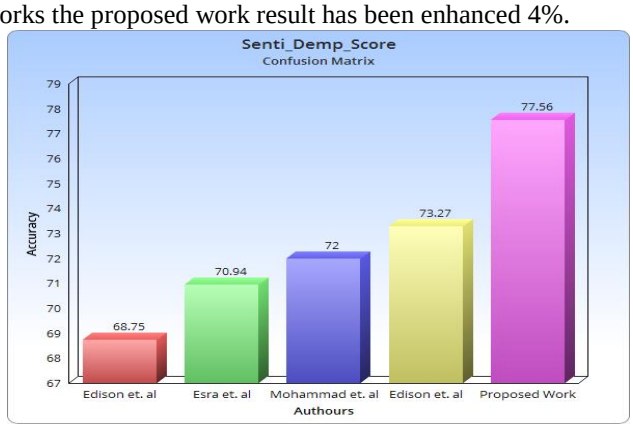

Figure4Comparison Result of the Proposed Work with Existing Work

VI. CONCLUSION AND FUTURE DIRECTION

From the literature study, least amount of studies were performed to calculate the sentiment score in the percentage level. The methodological diagram mentioned on this chapter, which wascompared with existing approaches. Anew approach Senti_Demp_Score proposed, which offers better accuracy than the existing work. Computation of the scoring method Like Senti_Demp_Score has produced better result, 4% of results have increased than the existing approach. In future the sentiment scoring method to be applied and calculated with a statistical analysis in different aspects. The proposed approaches can be proposed for measuring the accuracy in Machine Learning Approach.

REFERENCES

1. Quan Zou, SifaXie, Ziyu Lin, Meihong Wu and Ying Ju, “Finding the Best Classification Threshold in Imbalanced Classification”, Big Data Research, Elsevier, Vol. 5, 2016, pp. 2-8.
2. Deepak Singh Tomar, and Pankaj Sharma, “A Text Polarity Analysis Using Sentiwordnet Based an Algorithm”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7, Issue 1, 2016, pp: 190-193.
3. EsraAkbas, “Opinion Mining on Non-English Short Text”, International Symposium on Methodologies for Intelligent Systems, Springer, 2017.
4. Mohammad Ehsan Basiri, Ahmad Reza Naghsh-Nilchi, and Nasser Ghasem-Aghaee, “Sentiment Prediction Based on Dempster-Shafer Theory of Evidence”, Hindawi Publishing Corporation Mathematical Problems in Engineering, 2014, http://dx.doi.org/10.1155/2014/361201.
5. G. Shafer, “A Mathematical Theory of Evidence”, PrincetonUniversity Press, Princeton, NJ, USA, Vol. 1, 1976.
6. https://en.wikipedia.org/wiki/Dempster%E2%80%93Shafer_theory.
7. M. Edison and A. Aloysius, “Lexicon based Acronyms and Emoticons Classification of Sentiment Analysis on Big Data”, International Journal of Database Theory and Application (IJDTA), Vol. 10, Issue 7, 2017, pp. 41-54
8. E. Haddi, X. Liu and Y. Shi “The Role of Text Pre-Processing in Sentiment Analysis”, SciVerseScienceDirect ELSEVIER, 2013, pp. 26-32.
9. S. Roy, S. Dhar, S. Bhattacharjee and A. Das “A Lexicon based Algorithm for Noisy TextNormalization as Pre-Processing for Sentiment Analysis”, International Journal of Research inEngineering and Technology (IJRET), 2013, pp. 67-70.
10. H. Hamdan, P. Bellot, and F. Bechet. "lsislif: Feature extraction and label weighting for sentimentanalysis in twitter", Proceedings of the 9th International Workshop on Semantic Evaluation, 2015, pp.568-573.
11. Fuji Ren, and Kazuyuki Matsumoto. "Semi-automatic creation of youth slang corpus and its application to affective computing" IEEE Transactions on Affective Computing, 2016, pp. 176-189.
12. LU Xing, LI Yuan, WANG Qinglin and LIU Yu “An Approach to Sentiment Analysis of Short Chinese Text Based on SVMs”, 34th Chinese Control Conference (CCC), IEEE, 2015, pp. 9115-9120.
13. F. M. Kundi, S. Ahmed, A. Khan and M. Z. Asghar “Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet”, Life Science Journal, 2014, pp. 66-72.
14. S. Huang, W. Han, X. Que and W. Wang “Polarity Identification of Sentiment Words based onEmoticons”, 9th Conference on Computational Intelligence and Security, (2013), pp. 134-138.
15. G. G. Dayalani “Emoticon based unsupervised sentiment classifier for polarity analysis in tweets”,International Journal of Engineering Research and General Science, vol. 2, 2014, pp. 438-445.
16. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
17. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
18. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
19. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
20. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.