Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923146

POLARITY MULTI-VIEW TEXTUAL DATA FEATURE SELECTION USING PARAMETRIC FEATURE WEIGHT EQUIVALENCE BASED FEATURE SELECTION (PFWEFS)

[1]Dr.P.Logeswari , [2] J.Gokulapriya, [3]S.Sudha, [4]G.Banupriya
[1]Assistant Professor, School of Computer Applications,
Lovely Professional University
[2,3,4]Phd Research Scholar , [2,3,4]Department of Computer Science,
[2,3,4] Sri Krishna Arts & Science College,
[1]Phagwara,Punjab,[2,3,4] Coimbatore, Tamilnadu, India.
[1]tppselvalogu@gmail.com,[2]gokulapriyajaganathan1@gmail.com, [3]sudhasw89@gmail.com,
[4]banu.snmv7@gmail.com

Abstract - Feature selection has acquired importance in view of its commitments in saving classification costs as to time/computational loads. Searching for significant attributes, a feature search method is through decision trees. Features selection falls into 2 gatherings: filter as well as wrapper based techniques. Filters orders attributes through assessments models holding exclusively those features with values more noteworthy than a threshold. Wrappers search feature set for optimum sub-sets in a specific classifier. Execution measurements are connected to sub sets based on their presentation using a specific learning data set classifier. In this paper we proposed parametric feature weight equivalence based feature selection (PFWEFS) method for select a feature of polarity multi – view textual data. The proposed PFWEFS method provides the great result in experiment part.

Keywords: Feature selection, filter, wrapper, feature set, parametric feature weight equivalence.

1. Introduction

1.1 Feature Selection in Opinion Mining

The quick development of computer based high-throughput method has offered unrivaled open doors for people to expand limits in production, services, communications as well as research. Meanwhile, gigantic measures of high layered data are assembled to challenge astonishing data mining methods [2]. Features selection is a significant stage in data mining applications that can proficiently diminish data dimensionality through expulsion of non-important attributes. In the past couple of many years, researches have planned tremendous amounts of features selection protocols. The protocols are planned for filling different needs, of different models and have their own advantages as well as inadequacies. However there have been thorough endeavors in surveying currently present features selection protocols to the extent that is known, there is no devoted chronicle which assembles delegate features selection for working with the correlation as well as joint review. For filling this hole, Zhao et al., (2010) introduced a features selection chronicle that was formed for gathering the most popular protocols which have been formed in the features selection research for filling in as a stage to work with their application, examination as well as joint review [9]. The file additionally effectively helps research researchers in accomplishing more trustworthy assessments in the strategy of figuring out novel features selection protocols.

Features selection is utilized for diminishing the amount of features in a few applications wherein data has 100s or 1000s of attributes. Currently present features selection techniques basically center on finding significant attributes. Yu and Liu (2004) showed that feature significance alone isn't adequate for viable features selection of high-layered data [6]. Features overt repetitiveness was characterized as well as proposed for performing unequivocal overt repetitiveness examination in features selection. A novel framework was proposed which decouples pertinence as well as overt repetitiveness investigations. A connection based procedure was produced for pertinence as well as overt repetitiveness investigation, and led an observational investigation of its efficacy standing out it from other delegate techniques.

Classification of data crosses various areas has been broadly researched and is one of the fundamental methods for recognizing one from another, as need to realize which has a place with which bunch. It has the capacities to deduce the inconspicuous dataset with obscure class by dissecting its primary similitude to a given dataset with known classes [8]. Dependability on classification results is exceptionally pivotal issues. The higher the accuracy of created classification results, the better the classifier is. There are continually looking to expand the accuracy of classification, either through existing techniques or through advancement of new ones. Various cycles are applied to work on the accuracy of classification execution. While most existing methods tended to this assignment target further developing the classifier techniques, Omar et al., (2013) zeroed in on diminishing the quantity of features in dataset by choosing just the applicable features prior to giving the dataset to classifier. This spurred the requirement for adequate methods that fit for choosing the significant features with insignificant information misfortune. The point was to diminish the responsibility of classifier by utilizing feature selection methods [25]. With the attention on classification execution accuracy, the idea was featured, capacities and utilization of feature selection for different applications in classification issue. From the survey, classification with feature selection methods has shown noteworthy results with huge accuracy when contrasted with classification without feature selection.

Features selection research is filling in significance due to its commitment in saving classification costs concerning time as well as computational loads. In searching for essential attributes, a strategy is the searching of features through decision trees. Decision trees function as a middle person features space inducer for picking essential attributes. In decision tree-based feature selection, while others stay the decision tree, however involved pruning condition what functions as a threshold method for picking attributes [22].

2. Existing methodology

2.1. Term Frequency - Inverse Document Frequency (TF-IDF)

F. Sebastiani [1] proposed TF-IDF is a common measurement utilized in message classification errands; however its utilization in sentiment examination has been less far and wide and shockingly it doesn't seem to have been utilized as a unigram feature weight. TF-IDF is made out of two scores, term frequency and inverse document frequency. Term frequency is found by basically counting the number of times that a given term has happened in a given document, and inverse document frequency is found by isolating the complete number of documents by the number of documents that a given word shows up in. At the point when these qualities are duplicated together we get a score that is most noteworthy for words that show up as often as possible in a couple of documents, and low for terms that show up regularly in each document, permitting us to find terms that are significant in a document.

2.2 Chi-Square (CHI)

Wang et al., (2010) proposed a successful method of semantic job labeling based on cross breed comparative examples for Chinese comparative sentences [26]. In the proposed method, the first mixture comparative examples were developed according to the syntactic designs of comparative sentences. Then they were generalized to work on the accuracy and coverage rate of the labeling results. A two-level algorithm was intended for comparative substances labeling and comparative features labeling separately. The results of experiments showed the efficacy of the proposed method.

2.3 Greedy feature Selection algorithm

I. Tsamardinos, et.al proposed a parallel FS algorithm, namely, the parallel forward-backward with pruning algorithm, for large datasets [31]. The experimental concentrate on laid out expanded scalability with running time. The creators proposed involving MI to decrease dimensionality and further develop accuracy for online streams. The proposed study zeroed in on introducing a methodology to address the computational cost, the stability of the generated results, and the size of the last subset of chosen features.

Abdulaziz Alarifi, Amr Tolba et.al proposed a big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks [32]. This paper commitment proposes a novel big data and machine learning technique that can be utilized to assess different sentiment analysis processes. As enormous datasets are helpful in investigating systems in a powerful way, data were gathered from a tremendous volume of datasets. The noise in the gathered data was killed with the assistance of pre-processing data mining ideas. The greedy methodology with the CSO-LSTMNN algorithm was really carried out. Chosen features were taken care of into the sentiment classifier, which classified them as per a rule-based system. The effectiveness of the system was examined utilizing exploratory results, which were then contrasted and the PSO algorithm. Exploratory results demonstrated that the CSO-LSTMNN algorithm accomplished preferred execution over any remaining PSO algorithms, and the classification technique was assessed to be sufficient. The datasets were gathered from the Amazon Web site. Nonetheless, the accuracy of the algorithm should be improved by limiting text noise through different heuristic methods during enhancement.

2.4 Document Frequency

Parlar, Tuba & Özel, Selma. (2016) et.al proposed a new feature selection method for sentiment analysis of Turkish reviews. Document frequency is the simplest and scalable to the size of the training set. Document frequency counts the number of documents in a training set in which a feature term happens [33]. In light of a given thresholds, a term will be disposed of since the term doesn't hold a lot of value in expanding classification accuracy of sentiment analysis. Document frequency is typically involved along for certain different methods as reinforcement. This is on the grounds that it doesn't gauge whether the term is useful not normal for other feature selection methods. This method is additionally viewed as less forceful in the selection of features particularly for uncommon features accepting that interesting features might have high unmistakable abilities towards classes of documents40. Nicholls and Song acquainted a variety with the essential Document Frequency considering displayed below

Score (Term)=(DocFreq(Pos)-DocFreq(Neg))/(#ofDocsInTrainingset) (1)

The formula assigns a score in light of the contrast between the event of the term in positive and negative documents arrived at the midpoint of out by the absolute number of documents in the training set. In light of this perception, assuming that the term exists in both positive and negative documents, the score is 0. In the event that the term exists in all positive documents however not in any negative document the score is a most extreme 1. A base value of - 1 is given in the event that the term exists in all negative documents yet not in any positive document.

2.5 Ant colony optimization

Azuraliza Abu Bakarb, Mohd RidzwanYaakub et.al proposed Ant colony optimization for text feature selection in sentiment analysis. The proposed ACO-KNN algorithm had directed the feature selection process and utilized KNN to assess the applicant subset of features. Extensive experiments were led to assess the performance of the proposed ACO-KNN in finding noticeable features in different datasets. In the proposed hybrid algorithm, the MSE value of classification and the feature subset length were considered as appropriate measures to assess the performance of the algorithm. In view of the results, this algorithm had the option to choose the ideal feature subset without earlier information on the features. The computational results have shown that the proposed ACO-KNN algorithm could accomplish great performance with fewer features. The hybrid ACO-KNN algorithm had shown promising performances in terms of precision, recall, and F-score. It had performed better compared to IG-GA and IG-RSAR, with the exception of the Apex dataset. Therefore, important to find boundary settings are more reasonable for the ACO part of the hybrid algorithm to direct the insects to find the best subset of features.

3. Proposed methodology

This part details the methodology that adapted to identify the weights of each dimension of the features, record level, and corpus level. Every badge of feature with respect to a dimension like a term, slang, emoji, and emojis reflects their effect on conclude the given record is positive or negative to sentiment polarity [3]. The weight of the multitude of badge of striking dimensions of the features can further be utilized to signify the effect of the record to fall in one of the two labels of the sentiment polarity. To such an extent that the effect of the records with respect to a specific label is utilized further to demonstrates the weight threshold of the corpus of the relating label. The methods of evaluating the weight of the tokens connected with distinctive dimensions of the features, the weight of the records fall under a label of the given two labels of the given training corpus, and the effect threshold of the corpus relating to a label.

3.1 Feature level Weight

List every unique term, unique Slang tokens, and unique Emojis of the positive label as a set $Term^{+}$, set $Slang^{+}$ and set $Emoj^{+}$in individual request which is as follows:

To find the term weight of each term exists in the set$Term^+$, which is the coverage frequency of the comparing term in records of the positive label.

$$\bigvee_{i=1}^{|Term_+|} \{r_i^+ \exists t_i \in Term^+\}$$

Begin

$$\boldsymbol{isc(t_i)} = \frac{\sum_{j=1}^{|R^+|}\{1 \exists t_i \in r_j \in R^+\}}{|R^+|} \quad \textbf{(2)}$$

End

To find the slang weight of every symbolic exists in the set $Slang^+$ as follows, which is the coverage frequency of the comparing token in records of the positive label.

$$\bigvee_{i=1}^{|Slang^+|} \{S_i \exists s_i \in Slang^+\}$$

Begin

$$\boldsymbol{isc(Slang_i)} = \frac{\sum_{j=1}^{|R^+|}\{1 \exists Ss_i \in r_j \in R^+\}}{|R^+|} \quad \textbf{(3)}$$

End

To find the slang weight of every symbolic exists in the set $Emoj^+$ as follows, which is the coverage frequency of the comparing token in records of the positive label.

$$\bigvee_{i=1}^{|Emoj^+|} \{ej_i \exists ej_i \in Emoj^+\}$$

Begin

$$\boldsymbol{isc(emoj_i)} = \frac{\sum_{j=1}^{|R^+|}\{1 \exists ej_i \in r_j \in R^+\}}{|R^+|} \quad \textbf{(4)}$$

End

3.2 Impact Threshold of the feature dimensions

The impact threshold $ist^+_{Term}, ist^+_{Slang}, orist^+_{Emoj}$, of each dimension in particular request of term, slang, and emoji, and of the features is assessed further, which is the outright difference of the normal of the weight of all tokens of the relating dimension of the features, and their root mean square distance.

Impact threshold of the feature dimension terms is

$$\boldsymbol{\langle Term^+ \rangle} = \frac{\sum_{i=1}^{|Term^+|}\{isc(t_i) \exists t_i \epsilon Term^+\}}{|\,Term^+|} \quad \textbf{(5)}$$

$$eTerm^+ = \frac{\sum_{i=1}^{|Term^+|}\left\{\sqrt{\left(\langle Term^+\rangle - isc(t_i)\right)^2} \exists t_i \epsilon Term^+\right\}}{|\,Term^+|}$$ // assessing root mean square error

$$\boldsymbol{ist(Term^+)} = \sqrt{(\langle Term^+\rangle - eTerm^+)^2} \quad \textbf{(6)}$$

Impact threshold of the feature dimension slang is

$$\boldsymbol{\langle Slang^+ \rangle} = \frac{\sum_{i=1}^{|Slang^+|}\{isc(s_i) \exists s_i \epsilon Slang^+\}}{|Slang^+|} \quad \textbf{(7)}$$

$$eSlang^{+} = \frac{\sum_{i=1}^{|Sl^{+}|}\left\{\sqrt{\left(\langle Sl^{+}\rangle - isc(S_i)\right)^2} \exists t_i \epsilon Sl^{+}\right\}}{|Sl^{+}|}$$ // assessing root mean square error

$$\boldsymbol{ist(Slang^{+}) = \sqrt{(\langle Slang^{+}\rangle - eSlang^{+})^2}} \quad \textbf{(8)}$$

Impact threshold of the feature dimension Emojis is

$$\boldsymbol{\langle Emoj^{+}\rangle = \frac{\sum_{i=1}^{|Emoj^{+}|}\{isc(ej_i) \exists s_i \epsilon Emoj^{+}\}}{|Emoj^{+}|}} \quad \textbf{(9)}$$

$$\boldsymbol{eEmoj^{+} = \frac{\sum_{i=1}^{|Emoj^{+}|}\left\{\sqrt{\left(\langle Emoj^{+}\rangle - isc(ej_i)\right)^2} \exists ej_i \epsilon Emoj^{+}\right\}}{|Emoj^{+}|}}$$ // assessing root mean square error

$$\boldsymbol{ist(Emoj^{+}) = \sqrt{(\langle Emoj^{+}\rangle - eEmoj^{+})^2}} \quad \textbf{(10)}$$

Essentially, the further cycle finds the unique tokens of each feature dimension in different $Term^{-}, Slang^{-}, and\ Emoj^{-}$in particular request of feature dimensions terms, slang, and emojis. Then, the weight of every badge of each feature dimensions of the corpus having records labeled as negative.

Later the process finds the weight thresholds $ist(Term^{-}), ist(Slang^{-}), and\ ist(Emoj^{-})$of each dimension of the features in regard to the records labeled as negative.

3.3 Record Level Weight

The record level weight of each dimension of the features of the label positive is assessed, which is as follows:

Record level weight of the relating dimension called as terms, which is the normal of weight noticed for all features of the dimension term that existing the comparing recordr_i^{+}.

$$\bigvee_{i=1}^{|R^{+}|}\{r_i \exists r_i \in R^{+}\}$$

Begin

$$\boldsymbol{isc_{+}(r_i^t) = \frac{\sum_{j=1}^{|Term^{+}|}\{isc(t_j) \exists t_i \in Term^{+} \Delta t_j \in r_i\}}{|Term^{+}|}} \quad \textbf{(10)}$$

End

Record level weight of the relating dimension called as slang, which is the normal of weight noticed for all features of the dimension term that existing the comparing recordr_i^{+}.

$$\bigvee_{i=1}^{|R^{+}|}\{r_i \exists r_i \in R^{+}\}$$

Begin

$$\boldsymbol{isc_{+}(r_i^s) = \frac{\sum_{j=1}^{|Slang^{+}|}\{isc(s_j) \exists s_i \in Slang^{+} \Delta s_j \in r_i\}}{|Slang^{+}|}} \quad \textbf{(11)}$$

End

Record level weight of the relating dimension called as emojis, which is the normal of weight noticed for all features of the dimension term that existing the comparing record r_i^{+}.

$$\bigvee_{i=1}^{|R^{+}|}\{r_i \exists r_i \in R^{+}\}$$

Begin

$$\boldsymbol{isc}_{+}\left(\boldsymbol{r}_{\boldsymbol{i}}^{\boldsymbol{ej}}\right)=\frac{\sum_{j=1}^{|\boldsymbol{Emoj}^{+}|}\{\boldsymbol{isc}(\boldsymbol{ej}_{\boldsymbol{j}})\exists \boldsymbol{t}_{\boldsymbol{i}}\in \boldsymbol{Emoj}^{+}\Delta \boldsymbol{ej}_{\boldsymbol{j}}\in \boldsymbol{r}_{\boldsymbol{i}}\}}{|\boldsymbol{Emoj}^{+}|} \quad \textbf{(12)}$$

End

The transformation of the cycle portrayed above empowers to find the record level weights different dimensions of the features of the records labeled as negative, which are further signified as $isc\_(r_i^t), isc\_(r_i^s),\ isc\_(r_i^{ej})$ representing for each record ir of the label negative.

3.2.2 Corpus Level Impact Thresholds

This part defines the corpus level effect thresholds of each feature dimension concerning each label.

Corpus level Impact threshold of the terms of the records labeled as positive

$$\langle \boldsymbol{R}_{\boldsymbol{t}}^{+}\rangle = \sum_{\boldsymbol{i=1}}^{|\boldsymbol{R}^{+}|}\{isc_{+}(r_i^t)\exists r_i \in R^{+}\}/|\boldsymbol{R}^{+}| \quad \textbf{(13)}$$

//It is finding the average of the term level weight of the records labeled as positive

Corpus level Impact threshold of the slang of the records labeled as positive

$$\langle \boldsymbol{R}_{\boldsymbol{s}}^{+}\rangle = \sum_{\boldsymbol{i=1}}^{|\boldsymbol{R}^{+}|}\{isc_{+}(r_i^s)\exists r_i \in R^{+}\}/|\boldsymbol{R}^{+}| \quad \textbf{(14)}$$

//finding the average of the slang level weight of the records labeled as positive

Corpus level Impact threshold of the emojis of the records labeled as positive

$$\langle \boldsymbol{R}_{\boldsymbol{ej}}^{+}\rangle = \sum_{\boldsymbol{i=1}}^{|\boldsymbol{R}^{+}|}\left\{isc_{+}(r_i^{ej})\exists r_i \in R^{+}\right\}/|\boldsymbol{R}^{+}| \quad \textbf{(15)}$$

The comparative interaction on negative labeled records signifies thecorpus level effect thresholds $ist(R_t^-), ist(R_s^-), ist\left(R_{ej}^-\right), and\ ist(R_{et}^-)$, concerning feature dimensions terms, slang, emojis and emojis of the records addressing the negative sentiment polarity [28].

Algorithm for Parametric Feature Weight Equivalence Based Feature Selection (PFWEFS)

Step 1: Start the process.

Step 2: Let the sets $Term^{+}, Slang^{+}\ and\ Emoj^{+}$which are empty at their empty state.

Step 3: $\bigvee_{i=1}^{|R_+|}\{r_i^+\exists r_i^+ \in R_+\}$

Step 4: $Term^{+} = Term^{+} \cup r_i v_t^+$

Step 5: $Slang^{+} = Slang^{+} \cup r_i v_s^+$

Step 6: $Emoj^{+} = Emoj^{+} \cup r_i v_{ej}^+$

Step 7: Find the term weight of each term exists in the set$Term^{+}$with the use of equ 2.

Step 8: Find the slang weight of each token exists in the set $Slang^{+}$ with the use of equ 3.

Step 9: Find the emoji weight of each token exists in set $Emoj^{+}$ with the use of equ 4.

Step 10: Find the threshold of the feature dimension of term by equ 5

Step 11: Find the threshold of the feature dimension of slang by equ 7

Step 12: Find the threshold of the feature dimension of emoji by equ 9

Step 13: Find the Corpus level Impact threshold of the terms of the records 13

Step 14: Find the Corpus level Impact threshold of the slangs of the records 14

Step 15: Find the Corpus level Impact threshold of the emojis of the records 15

Step 16: Stop the process.

4. Experimental Result

4.1 Precision

Dataset	CHI	TF-IDF	Proposed PFWEFS
1	0.67	0.83	0.92
2	0.63	0.73	0.85
3	0.57	0.66	0.83
4	0.52	0.63	0.81
5	0.48	0.56	0.72

Table 1.Comparison table of Precision

The Comparison table 1 of Precision Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. While comparing the Existing algorithm (CHI, TF-IDF) and proposed PFWEFS method provides the better results. The proposed method provides the great results.

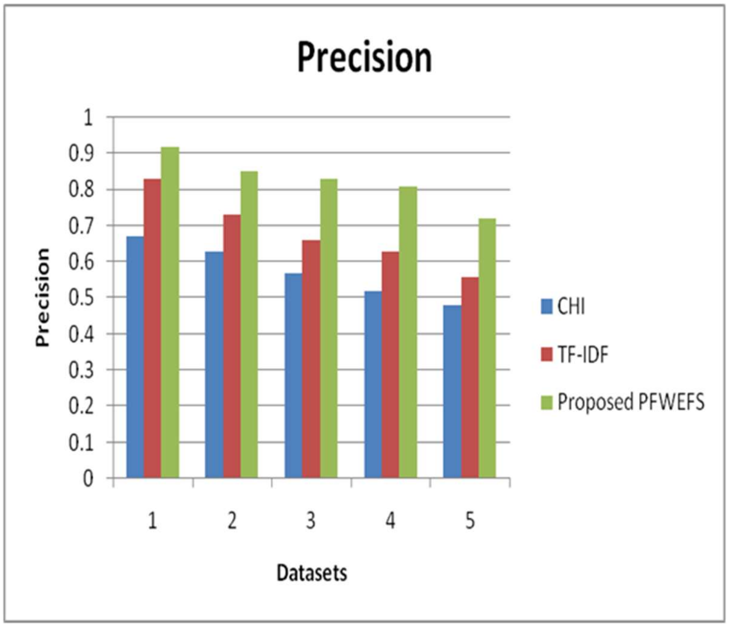

Figure 1.Comparison chart of Precision

The Figure 1 comparison chart of Precision Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. X axis denote the Datasets and Y axis denotes the Precision in percentage. The proposed method provides the great results. The existing algorithm values start from 0.67 to 0.48, 0.83 to 0.56 and proposed PFWEFS method values start from 0.92 to 0.72. The proposed method provides the great results.

4.2 Recall

Dataset	CHI	TF-IDF	Proposed PFWEFS
2	0.60	0.65	0.75
4	0.321	0.453	0.643

6	0.65	0.71	0.83
8	0.28	0.41	0.61
10	0.704	0.76	0.89

Table 2.Comparison table of Recall

The Comparison table 2 of Recall Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. While comparing the Existing algorithm (CHI, TF-IDF) and proposed PFWEFS method provides the better results. The proposed method provides the great results.

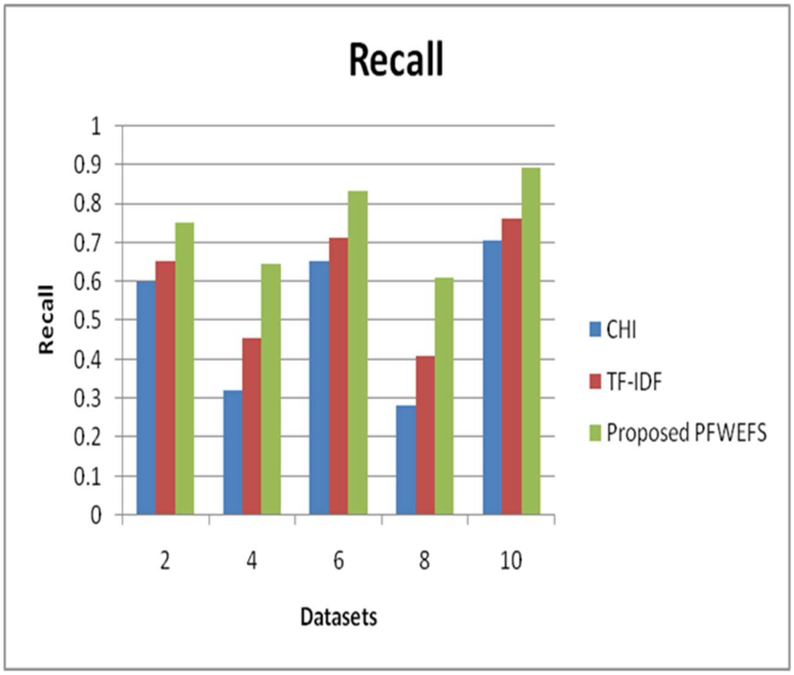

Figure 2.Comparison chart of Recall

The Figure 2 comparison chart of Recall Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. X axis denote the Datasets and Y axis denotes the Recall in percentage. The proposed method provides the great results. The existing algorithm values start from 0.60 to 0.704, 0.65 to 0.76 and proposed PFWEFS method values start from 0.75 to 0.89. The proposed method provides the great results.

4.3 F-Measure

Dataset	CHI	TF-IDF	Proposed PFWEFS
2	0.58	0.62	0.81
4	0.32	0.25	0.64
6	0.65	0.71	0.84
8	0.52	0.41	0.70
10	0.71	0.78	0.95

Table 3.Comparison table of F-Measure

The Comparison table 3 of F-Measure Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. While comparing the Existing algorithm (CHI, TF-IDF) and proposed PFWEFS method provides the better results. The proposed method provides the great results.

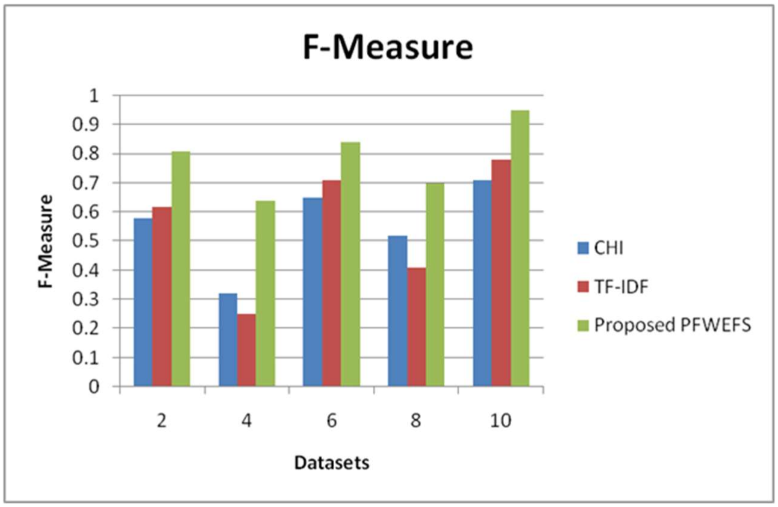

Figure 3.Comparison chart of F-Measure

The Figure 3 comparison chart of F-Measure Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. X axis denote the Datasets and Y axis denotes the F-Measure in percentage. The proposed method provides the great results. The existing algorithm values start from 0.58 to 0.71, 0.62 to 0.78 and proposed PFWEFS method values start from 0.81 to 0.95. The proposed method provides the great results.

4.4 Accuracy

Dataset	CHI	TF-IDF	Proposed PFWEFS
1	25	35	55
2	35	45	60
3	45	55	70
4	55	65	80
5	65	75	95

Table 4.Comparison table of Accuracy

The Comparison table 4 of Accuracy Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. While comparing the Existing algorithm (CHI, TF-IDF) and proposed PFWEFS method provides the better results. The proposed method provides the great results.

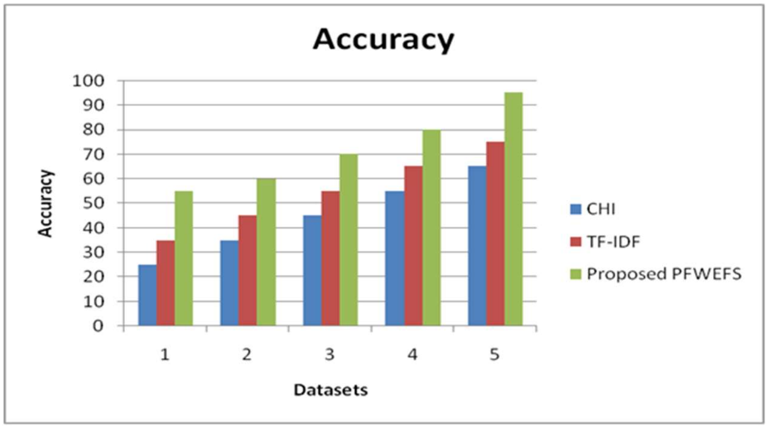

Figure 4.Comparison chart of Accuracy

The Figure 4 comparison chart of Accuracy Values explains the different values of existing algorithms (CHI, TF-IDF) and proposed PFWEFS method. X axis denote the Datasets and Y axis denotes the Accuracy in percentage. The proposed method provides the great results. The existing algorithm values start from 0.25 to 0.65, 0.35 to 0.75 and proposed PFWEFS method values start from 0.55 to 0.95. The proposed method provides the great results.

5. Conclusion

In this paper, the proposed algorithm Parametric Feature Weight Equivalence Based Feature Selection (PFWEFS) identifies the weights of each dimension of the features, record level, and corpus level. Every badge of feature with respect to a dimension like a term, slang, emoji, and emojis reflects their effect on conclude the given record is positive or negative to sentiment polarity. Feature Selection can eliminate insignificant or excess features, and in this way decline the number of features to work on the accuracy of the model utilizing proposed PFWEFS. After training and selecting the best models, the accuracy of each best model was obtained by testing on holdout data set. This accuracy was considered for further elimination of models which were lesser than 90% accurate.

References

- Esuli and F. Sebastiani. SentiWordNet: a publicly available lexical resource for opinion mining. In Proc. of LREC 2006 - 5th Conf. on Language Resources and Evaluation, Volume 6, 2006.
- Abbasi, HC Chen and a Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems, Volume 26, Number 3, 2008.
- Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. of the ACL, pages 271–278. ACL, 2004.
- Y. Ong, S. W. Goh and C. Xu, "Sparsity adjusted information gain for feature selection in sentiment analysis," 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 2122-2128, doi: 10.1109/BigData.2015.7363995.
- Pong-Inwong and K. Kaewmak, "Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 1222-1225, doi: 10.1109/CompComm.2016.7924899.

- Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. 2016, 66, 245–260. [Google Scholar] [CrossRef]
- A. Kristiyanti and M. Wahyudi, "Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review," 2017 5th International Conference on Cyber and IT Service Management (CITSM), 2017, pp. 1-6, doi: 10.1109/CITSM.2017.8089278.
- Debole, F.; Sebastiani, F. Supervised term weighting for automated text categorization. In Proceedings of the Text Mining and Its Applications, 1st ed.; Janusz, K., Ed.; Springer: Berlin, Germany, 2004; Volume 138, pp. 81–97. [Google Scholar]
- S. Usop, R. R. Isnanto and R. Kusumaningrum, "Part of speech features for sentiment classification based on Latent Dirichlet Allocation," 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2017, pp. 31-34, doi: 10.1109/ICITACEE.2017.8257670.
- Akbarian and F. Z. Boroujeni, "An Improved Feature Selection Method for Sentiments Analysis in Social Networks," 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 2020, pp. 181-186, doi: 10.1109/ICCKE50421.2020.9303710.
- F. R. Saputra Rangkuti, M. A. Fauzi, Y. A. Sari and E. D. L. Sari, "Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection," 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018, pp. 88-91, doi: 10.1109/SIET.2018.8693211.
- F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., Volume 34, Number 1, pages 1–47, 2002.
- Ighazran, L. Alaoui and T. Boujiha, "Metaheuristic and Evolutionary Methods for Feature Selection in Sentiment Analysis (a Comparative Study)," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2018, pp. 1-6, doi: 10.1109/ISAECT.2018.8618799.
- Utama, "Sentiment Analysis in Airline Tweets Using Mutual Information for Feature Selection," 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2019, pp. 295-300, doi: 10.1109/ICITISEE48480.2019.9003903.
- Kurniawati and H. F. Pardede, "Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis," 2018 International Conference on Information Technology Systems and Innovation (ICITSI), 2018, pp. 1-5, doi: 10.1109/ICITSI.2018.8695953.
- Kun and Z. Lei, "Sentiment Feature Selection Algorithm for Chinese Micro-blog," 2014 International Conference on Management of e-Commerce and e-Government, 2014, pp. 114-118, doi: 10.1109/ICMeCG.2014.32.
- J. Liang, X. Zhou, P. Liu and L. Guo, "Latent sentiment representation for sentiment feature selection," 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2015, pp. 71-72, doi: 10.1109/INFCOMW.2015.7179348.

- Mihret and M. Atinaf, "Sentiment Analysis Model for Opinionated Awngi Text," 2019 IEEE AFRICON, 2019, pp. 1-6, doi: 10.1109/AFRICON46755.2019.9134016
- S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," in IEEE Access, vol. 9, pp. 52177-52192, 2021, doi: 10.1109/ACCESS.2021.3069001.
  - Nurhayati, A. E. Putra, L. K. Wardhani and Busman, "Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document," 2019 7th International Conference on Cyber and IT Service Management (CITSM), 2019, pp. 1-7, doi: 10.1109/CITSM47753.2019.8965332.
  - R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. Journal of Informetrics, 2009.
  - S. Fong, E. Gao and R. Wong, "Optimized Swarm Search-Based Feature Selection for Text Mining in Sentiment Analysis," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1153-1162, doi: 10.1109/ICDMW.2015.231.
  - S. R. Ahmad, A. A. Bakar and M. R. Yaakub, "Metaheuristic algorithms for feature selection in sentiment analysis," 2015 Science and Information Conference (SAI), 2015, pp. 222-226, doi: 10.1109/SAI.2015.7237148.
  - T. Chen, P. Su, C. Shang, R. Hill, H. Zhang and Q. Shen, "Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2019.8858916.
  - T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In Proc. of the Conf. on Empirical Methods in Natural Language Processing EMNLP, pages 412–418, 2004.
  - Wang, Y.; Sun, L.; Wang, J.; Zheng, Y.; Youn, H.Y. A Novel Feature-Based Text Classification Improving the Accuracy of Twitter Sentiment Analysis. In Proceedings of the Advances in Computer Science and Ubiquitous Computing, Singapore, 18 December 2017. [Google Scholar]
  - X. Chen, J. Ma and Y. Lu, "Feature selection for Chinese online reviews sentiment classification," 2013 International Conference on Computational Problem-Solving (ICCP), 2013, pp. 79-82, doi: 10.1109/ICCPS.2013.6893490.
  - Yang, A.; Jun, Z.; Lei, P.; Yang, X. Enhanced twitter sentiment analysis by using feature selection and combination. In Proceedings of the Security and Privacy in Social Networks and Big Data, Hangzhou, China, 16–18 November 2015. [Google Scholar]
  - Zheng, L.; Wang, H.; Gao, S. Sentimental feature selection for sentiment analysis of Chinese online reviews. Int. J. Mach. Learn. Cybern. 2018, 9, 75–84. [Google Scholar]
  - Zhou, H.; Guo, J.; Wang, Y.; Zhao, M. A feature selection approach based on interclass and intraclass relative contributions of terms. Comput. Intell. Neurosci. 2016.
  - Tsamardinos, Ioannis&Borboudakis, Giorgos&Katsogridakis, Pavlos&Pratikakis, Polyvios&Christophides, Vassilis. (2017). Massively-Parallel Feature Selection for Big Data.

- Alarifi, Abdulaziz &Tolba, Amr & Al-Makhadmeh, Zafer & Said, Wael. (2020). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. The Journal of Supercomputing. 76. 10.1007/s11227-018-2398-2.
- Parlar, Tuba & Özel, Selma. (2016). A new feature selection method for sentiment analysis of Turkish reviews. 1-6. 10.1109/INISTA.2016.7571833.
- J.Gokulapriya&Dr.P.Logeswari (2021). Pre-Processing and Feature Extraction of Polarity Multi-View Textual Data using Text Data Mining, ISSN: 0011-9342 |.