G. R. Brindha, P. Swaminathan, B. Santhi, 2016. Performance analysis of new word weighting procedures for opinion mining. *Frontiers of Information Technology & Electronic Engineering*, **17**(11):1186-1198. http://dx.doi.org/10.1631/FITEE.1500283

Performance analysis of new word weighting procedures for opinion mining

Key words: Inferred word weight, Opinion mining, Supervised classification, Support vector machine (SVM), Machine learning

Corresponding author: G. R. Brindha
E-mail: brindha.gr@ict.sastra.edu
ORCID: http://orcid.org/0000-0001-5911- 8327

Motivation

- The proliferation of forums and blogs leads to challenges and opportunities for processing large amounts of information.
- The information shared on various topics often contains opinionated words which are qualitative in nature.
- These qualitative words need statistical computations to be converted into useful quantitative data.
- Interesting studies concerning text analysis using different approaches are the basic motivation for text mining (Church and Hanks, 1989; Geng and Hamilton, 2006; Armstrong *et al*., 2009).

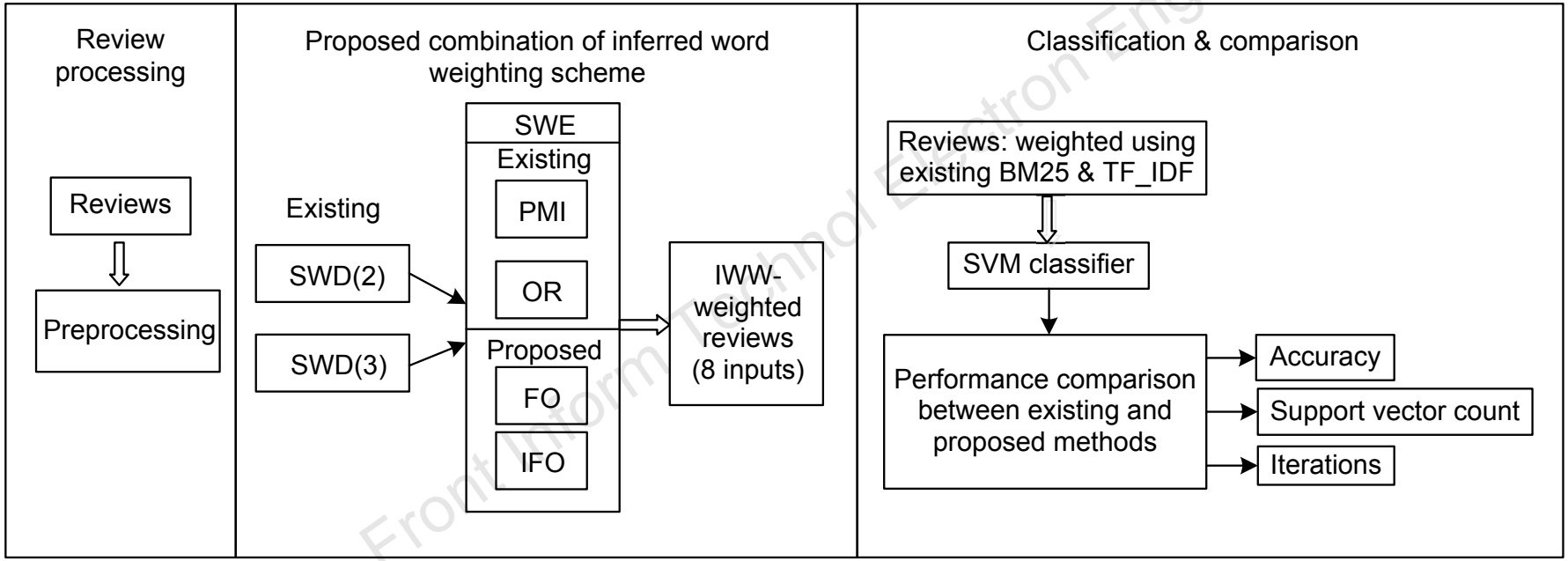
Motivation

- Classification is the significant process among all review processing.
- Recent works focused on the procedure of term selection, since each term has its own value in conveying its opinions (Das and Chen, 2001; Debole and Sebastiani, 2003).
- The value of a term depends on its contribution to the review document and the significance of the meaning it conveys.
- Hence, inbetween the two stages of term selection and classification, the term ‘weighting scheme’ should be included (Das and Chen, 2001).

Core concepts

- To process the linguistic meaning of words into data and enhance opinion mining analysis, we propose a novel weighting scheme, referred to as ‘inferred word weighting’ (IWW).
- IWW is computed based on the significance of the word in the document (SWD) and the significance of the word in the expression (SWE) to enhance their performance (IWW=SWD*SWE).
- In addition to the new weighting methods, further enhancement is done to improve the performance of text classification by including stop-words (Generally, stop-words are removed in text processing).

Workflow



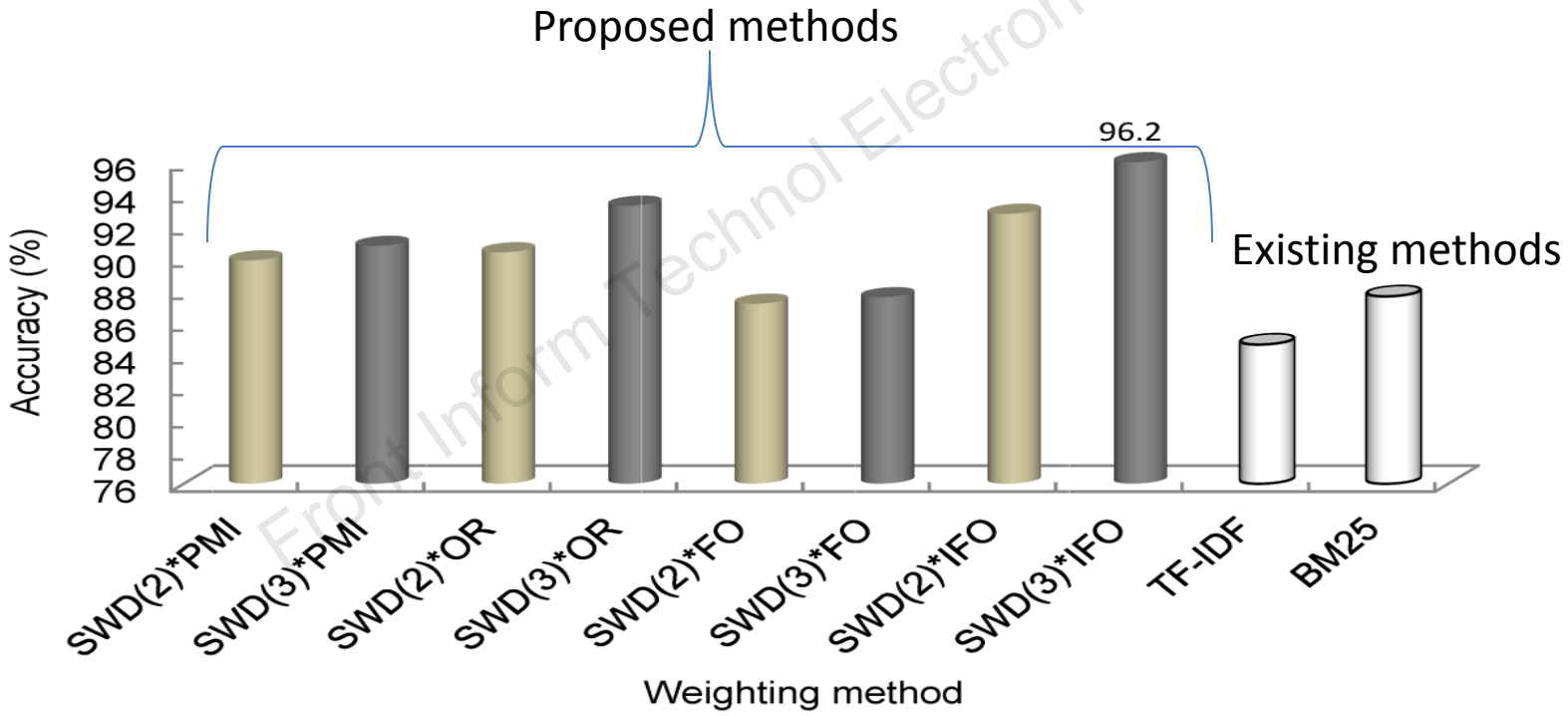
Workflow of the proposed weighting and classification

Method

1. Novelty in this proposal is to introduce an integrated weighting technique:
   - Significance of a word in a document (SWD based on frequency):
     1. Term frequency
     2. Normalized term frequency
   - Significance of a word in expression (SWE based on information gain):
     1. Pointwise mutual information
     2. Odds ratio
     3. Frequency and odds
     4. Improved frequency and odds
2. Establish the strength of the proposed method through the optimum classifier——support vector machine.
3. Corpus: benchmark data sets are used, including Cornell movie reviews, Amazon product reviews, and Stanford movie reviews.

Major results

- Our proposed methods enhanced the classification accuracy by IWW



Weight function performance comparison for Cornell movie reviews

Major results (Cont'd)

- Our proposed methods enhanced the classification accuracy further by including stop-words

Corpus	SWE(t_j)	Accuracy (%)					
		Proposed methods				Existing methods	
		PMI	OR	FO	IFO	TF-IDF	BM-25
Cornell movie reviews	TF	90.2	91.3	88.1	93.4	89.1	90.7
	NTF	91.7	93.9	88.2	97.4		
Amazon product reviews	TF	89.5	90.3	88.8	92.5	87.3	90.9
	NTF	90.9	92.5	87.7	96.3		
Stanford movie reviews	TF	90.8	90.2	86.8	89.8	84.6	91.8
	NTF	94.5	92.9	88.1	94.5		

Conclusions

- Inspired by the observations in a recent survey, this paper presents a novel integrated statistical method for weighting words in the reviews.
- We verified our method using benchmark data sets through the performance of the classifier.
- The corpus was weighted by including stop-words:

  Two facts were observed:

  (1) Classification performance was enhanced;

  (2) The outcome difference between inclusion and exclusion of stop-words was smaller in the proposed methods, and larger in existing methods.

Hence, for frequency based weighting methods, stop-word removal is not necessary.