Probabilitic Feature Selection

(Filter method)

Ashutosh Upreti . Siddharth Mundada

Major Components

- ➔ Introduction
- ➔ Previous Work
- ➔ Our Approach
- ➔ Experimental Setup
- ➔ Conclusion and Future Work

Feature Selection

INTRODUCTION

- Feature Selection and its importance in text classification
- Types of Feature Selection methods
  a. Filter Methods
  b. Wrapper Methods
  c. Embedded Methods

PREVIOUS WORK

- Distinguishing Feature Selector (DFS)
- Improved Gini-Index Algorithm
- Chi-square with K-Means

DISTINGUISHING FEATURE SELECTON (DFS)

$$\mathrm{DFS}(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\overline{C_i}) + 1}$$

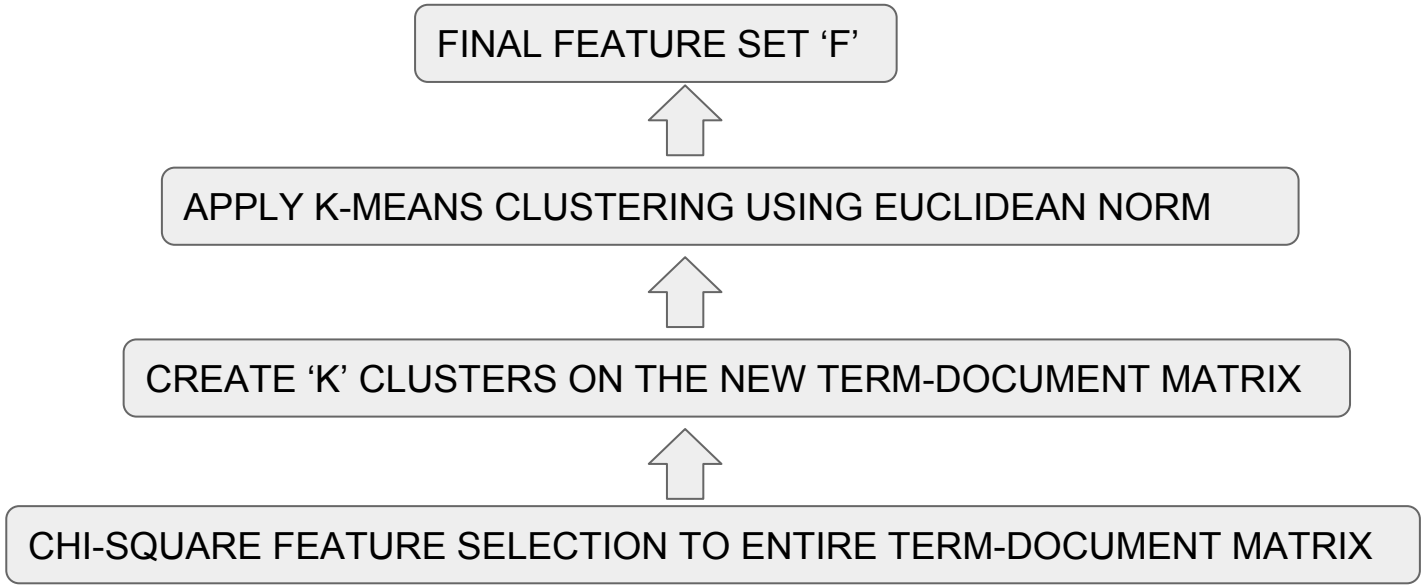
IMPROVED GINI-INDEX ALGORITHM

$$\text{GiniTxt}(w) = \sum^{m} P(w|C_i)P(C_i|w)$$

$$\text{TF}(w|C_i) = \sum_{d \in C_i}^{|C_i|} TF(w|d)$$

$$\text{Gini-TF}(w) = \sum_{i=1}^{m} \sum_{d \in C_i}^{|C_i|} P(C_i|w)TF(w|d)$$

CHI-SQUARE WITH K-MEANS



OUR APPROACH

- Gini-DFS
- Improved Gini-DFS
- Applying K-Means before Improved Gini-DFS

Gini-DFS

$$\text{Gini-DFS}(w) = \sum_{i=1}^{m} \sum_{d \in C_i}^{|C_i|} \text{DFS}_i(w)\text{TF}(w|d)$$

$$\text{DFS}_i = \sum_{i=1}^{m} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\overline{C_i}) + 1}$$

IMPROVED GINI-DFS

$$\mathrm{GiniImpTF}(t) = \sum_{C_i} \mathrm{DFS}_i(t) * \mathrm{ImpTF}_i(t)$$

$$\mathrm{ImpTF}_i(t) = \frac{\mathrm{TF}_i(t) * \mathrm{ATF}_i(t)}{\mathrm{M}_i}$$

$$\mathrm{ATF}_i(t) = \frac{\sum_{d=1}^{k} tf_{t,d}}{N_i}$$

$$\mathrm{M}_i = \frac{\sum_{d \in C_i} \mathrm{termCount}(d)}{D_i}$$

K-MEANS WITH IMPROVED-GINI-DFS

- Applying K-Means Clustering to preprocess the data
- Using a threshold, we select the top-N features from every cluster
- Feature set constructed is passed to Improved-Gini-DFS

Experimental setup

- Dataset Information
- Classification Algorithms
- Performance measures
- Experimental setting
- Visualizations

DATASET INFORMATION

- WebKB
- Reuters-8
- Newsgroup20

CLASSIFICATION ALGORITHMS

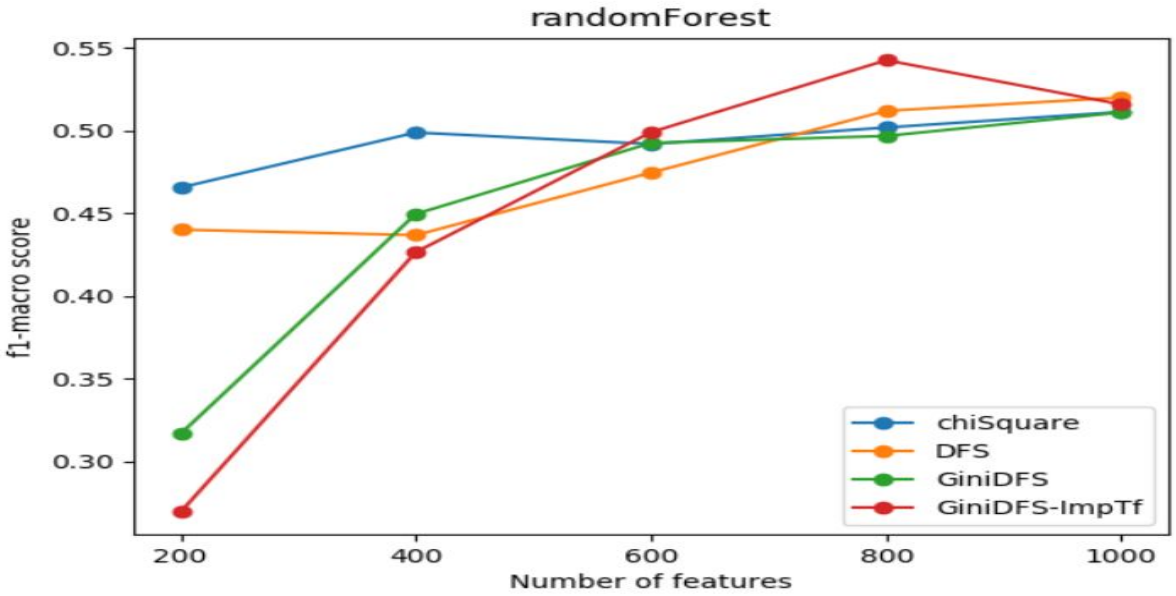
- Support Vector Machine (SVM)
- Multinomial Bayes Classifier
- Random Forest Classifier

Performance measures

$$\text{Micro-F1} = \frac{2 * p * r}{p + r}$$

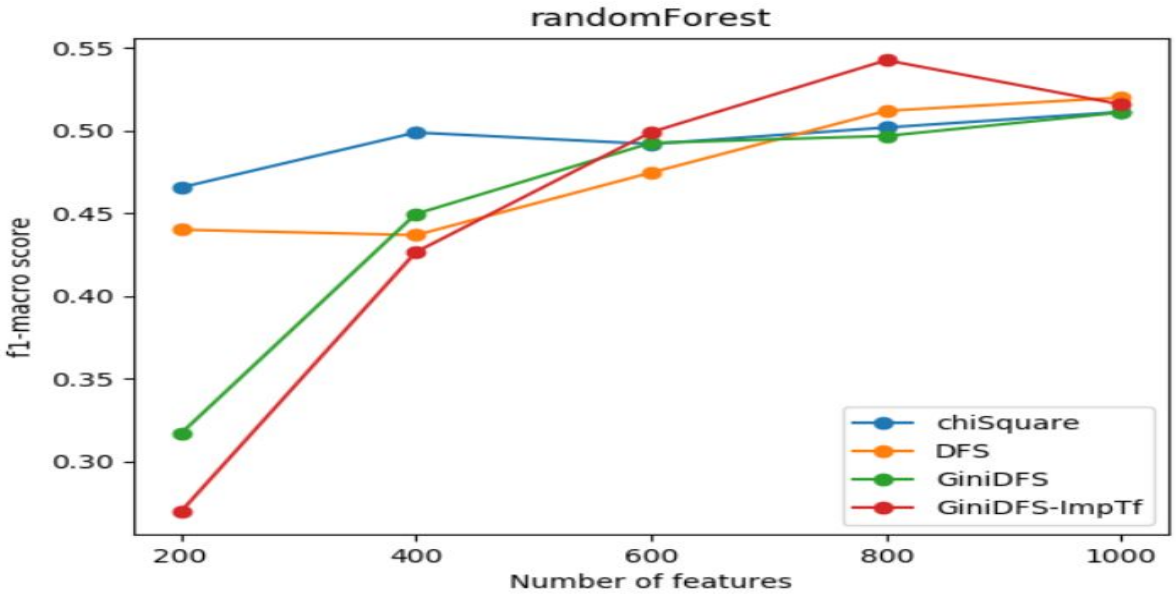
$$\text{Macro-F1} = \frac{\sum_{k=1}^{C} F_k}{C}, \quad F_k = \frac{2 * p_k * r_k}{p_k + r_k}$$

VISUALIZATIONS



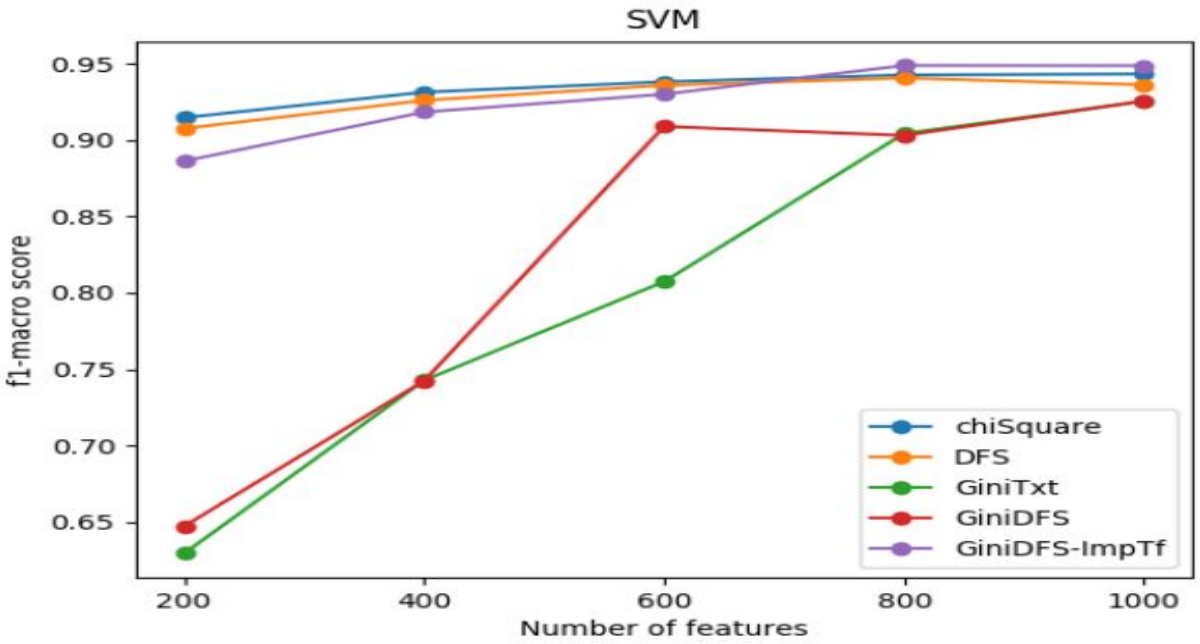
(a) Newsgroup-20 dataset

VISUALIZATIONS



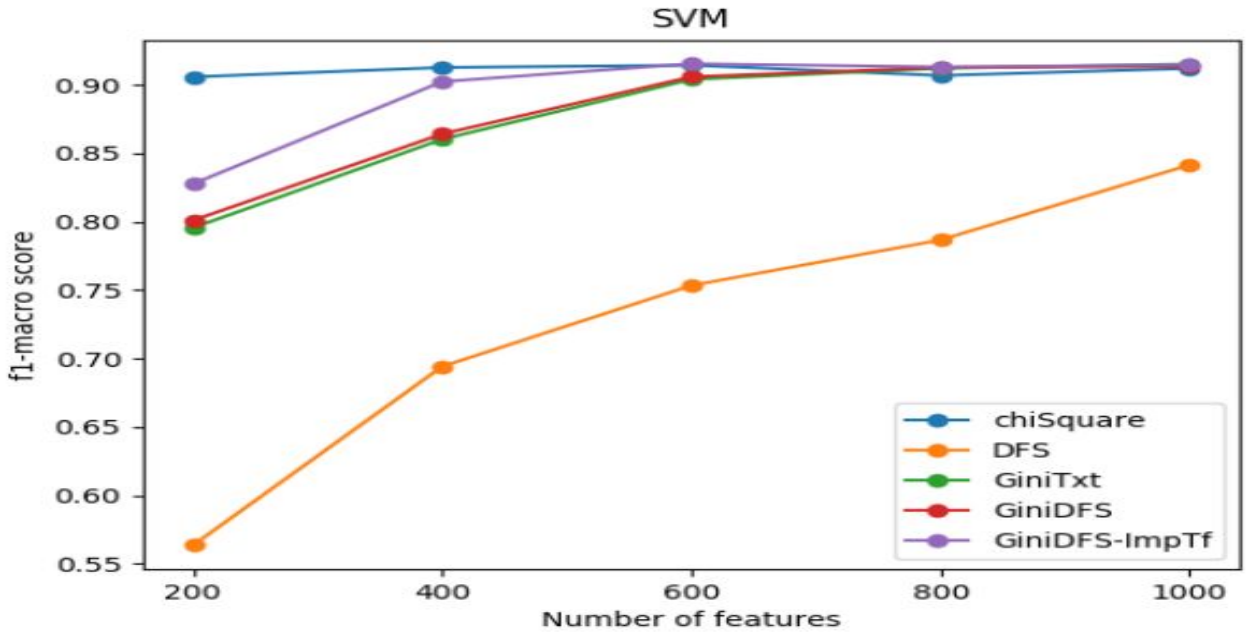
(a) Newsgroup-20 dataset

VISUALIZATIONS



(b) Reuters-8 dataset

VISUALIZATIONS



(a) WebKB dataset comparison

CONCLUSION AND FUTURE WORK

- Dataset can be skewed in nature. In such cases, inclusion of TF (term-frequency) in the scoring function proves to be useful.
- Filter methods sometimes become vulnerable to high frequency terms.
- Performance of feature selection techniques rely on the classifier used.
- In our future work, we can use model probabilities with different probability distributions. We can also use a weighted scoring function.