# A New Approach For Selecting Informative Features For Text Classification

Zinnar Ghasem[1], Ingo Frommholz[1], and Carsten Maple[2]

[1] University of Bedfordshire, UK
[2] University of Warwick, UK
{zinnar.ghasem,ingo.frommholz}@beds.ac.uk
carsten.maple@warwick.ac.uk

**Abstract.** Selecting useful and informative features to classify text is not only important to decrease the size of the feature space, but as well for the overall performance and precision of machine learning. In this study we propose a new feature selection method called Informative Feature Selector (IFS). Different machine learning algorithms and datasets have been utilised to examine the effectiveness of IFS, and it is compared to well-established methods, namely Information Gain, Odd Ratio, Chi Square, Mutual Information and Class Discriminative Measure. Our experiments show that IFS is able to outperform aforementioned methods and to produce effective and efficient results.

**Keywords:** Feature selection, text classification, text preprocessing

## 1 Introduction

Automatic text classification is increasingly imperative for managing a substantial amount of data and information on the Web, for instance for search, spam filtering, malware classification, etc. It has been well studied over the past half century [1, 2], and a number of machine learning and statistical algorithms have extensively been used in various types of classification. In text classification, each document needs to be represented as a vector of features, for which various methods have been employed – Bag-of-Words is one of the major methods, where all unique features (in this case terms) of documents are utilised. This approach results in a high dimensional vector space of features, particularly in the case of a big dataset or where we find a large volume of document content. This large set of features is one of main issues of text classification, therefore it is highly desirable to reduce the size of the vector space by selecting useful features.

For this purpose, we propose a probabilistic *Informative Feature Selector (IFS)*. The IFS is compared using ten folds cross validation to some of the

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

well-known methods, namely Chi Square *(chi)*, Odd Ratio*(OR)*, Information Gain *(IG)*, Class Discriminative Measures *(CDM)* and Mutual Information *(MI)*. The results of our experiment show that IFS is comparable to some and superior to others in term classification accuracy. The rest of this paper is organised as follows: Section 2 lists all feature selection methods which have been compared with IFS. Section 3 introduces IFS method. Section 4 outlines the experiments; datasets and performance measures, while results are analysed in section 5 and conclusion is the last section.

## 2 Related Work

As aforementioned, there exist a number of feature selection methods for text classification; among them IG, OR, CDM, and Chi have been claimed to perform well in experimental studies [3–6, 11, 8]. Thus, specifically those methods have been selected to be evaluated against IFS. In the following subsections, a brief description of each method is provided.

### 2.1 Information Gain (IG)

Information gain evaluates the usefulness of a feature for classification based on absence and presence of that feature in documents [9]. Information Gain defines the relationship between class $c_j$ and feature $t_i$ in the following formula.

$$IG(t_i, c_j) = -\sum_{j=1}^{m} P(c_j)\log(P(c_j)) + P(t_i)\sum_{j=1}^{m} P(c_j|t_i)\log(P(c_j|t_i)) + P(\bar{t_i})\sum_{j=1}^{m} P(c_j|\bar{t_i})\log(P(c_j|\bar{t_i})), \quad (1)$$

where $P(c_j)$ is the probability of $c_j$, $P(t_i)$, and $P(\bar{t_i})$ is the probability of $t_i$ occurring and not occurring in $c_j$ correspondingly , while $P(c_j|t_i)$ is the probability of $c_j$ given the probability of $t_i$ , and $P(c_j|\bar{t_i})$ is the probability of $c_j$ given the probability of $\bar{t_i}$ and $m$ is number of classes in the training set.

### 2.2 Mutual Information (MI)

Mutual information evaluates the association between a feature and a class as

$$MI(t_i, c_j) = \log\left(\frac{P(t_i \wedge c_j)}{P(t_i)P(c_j)}\right) = \log\left(\frac{P(t_i|c_j)}{P(t_i)}\right). \quad (2)$$

However, using the the two ways contingency table shown in Table 1, Eq. 2 can be represented in terms of the number of documents in a class as shown in Eq. 3 where $\alpha$ and $\lambda$ represent the number of documents containing and not containing term $t_i$ in class $c_j$ respectively, similarly $\beta$ and $\delta$ are the number of documents

**Table 1.** Two ways contingency table

| | $C_j$ | $\neg C_j$ |
|---|---|---|
| term $t_i$ | $\alpha$ | $\beta$ |
| term $\bar{t_i}$ | $\lambda$ | $\delta$ |

containing and not containing $t_i$ in $\neg C_j$. If $\eta = \alpha + \beta + \lambda + \delta$ is total number of documents in all classes, MI can be estimated as

$$MI(t_i, c_j) = \log\left(\frac{\alpha \times \eta}{(\alpha + \lambda)(\alpha + \beta)}\right) \tag{3}$$

### 2.3 Chi Square ($\chi^2$)

$\chi^2$ or *chi* has been reported as one of top feature selection methods and it measures the correlation between feature $t_i$ and class $c_j$ [3]. Using Table 1, $\chi^2$ is defined as

$$\chi^2(t_i, c_j) = \frac{\eta \times (\alpha\delta - \beta\lambda)^2}{(\alpha + \lambda)(\beta + \delta)(\alpha + \beta)(\lambda + \delta)} \tag{4}$$

The goodness of $t_i$ decreases as the independence between $t_i$ and $c_j$ increases to the point where the value is zero, that is, when the number of documents containing term $t_i$ in classes is equal.

### 2.4 Odd Ratio (OR)

Odd Ratio was initially developed for a binary classification; it has been reported by [6] that this value has additional advantages to other classification methods. Odd Ratio can be computed as follow:

$$OR(t_i) = \log\left(\frac{odd(t_i|pos)}{odd(t_i|neg)}\right) = \log\left(\frac{P(t_i|pos)(P(1 - P(t_i|neg)))}{P(t_i|neg)(P(1 - P(t_i|Pos)))}\right) \tag{5}$$

where both "pos" and "neg" represent a positive and negative class, respectively.

### 2.5 Class Discrimination Measure (CDM)

The CDM is an improved version of Multi-class Odd Ratio (MOR), where MOR is originally based on Odd Ratio. CDM is introduced in [10] and has reportedly outperformed IG.

$$CMD(t_i, c_j) = \sum_{j=1}^{m} \left| log \frac{P(t_i|c_j)}{P(t_i|\bar{c_j})} \right| \tag{6}$$

where $P(t_i|\bar{c_j})$ is the probability that $t_i$ occurs in another class than $c_j$.

## 3 Informative Feature Selector (IFS)

The principal aim of a feature selection method is to differentiate between informative and uninformative features by giving highest values to most informative and lowest values to least informative features, which subsequently facilitates the process of reducing the size of the feature vector space. However, the following issues must be taken into account by a feature selection method in the process of weighting the features for text classification. A feature that appears in all documents of a class and does not appear in any documents of other class(es) is a good discriminator and should be given the highest value. Consequently, a feature which equally appears in all classes should be given a lower value. The value assigned to a feature should reflect its degree of discrimination and its correlation between the classes.[3].

Based on these consideration, the IFS method is formulated as follows:

$$IFS(t_i, c_j) = \log(|(P(t_i|c_j)P(\bar{t_i}|\bar{c_j}) - P(t_i|\bar{c_j})P(\bar{t_i}|c_j))| \frac{|P(t_i|c_j)-P(t_i|\bar{c_j})|}{\min(P(t_i|c_j),P(t_i|\bar{c_j}))+1} + 1) \quad (7)$$

where $P(t_i|c_j)$ is the probability of $t_i$ appearing in class $c_j$ and $P(\bar{t_i}|c_j)$ is the probability of $t_i$ not appearing in class $c_j$. Similarly $P(t_i|\bar{c_j})$ is the probability of $t_i$ appearing in another class $\bar{c_j}$ and $P(\bar{t_i}|\bar{c_j})$ is probability of $t_i$ not occurring in $\bar{c_j}$.

IFS has been formulated in a way so that the overall value given to a feature is sensitive to changes in the number of features which are absent, shared or present in classes. In order for IFS to assign a value which reflects the usefulness of a feature for classification, and to adhere to above considerations, IFS makes use of the difference between the probability of a feature that appears in both classes, then dividing it by the smallest value between $P(t_i|c_j)$ and $P(t_i|\bar{c_j})$; 1 is added in case the smallest value is zero (this is the 2nd part of the equation). This makes sure the feature is assigned an appropriate value not only according to the differences between $P(t_i|c_j)$ and $P(t_i|\bar{c_j})$ but also according to the probability of $t_i$ occurring in the intersection between $c_j$ and $\bar{c_j}$. However, the calculation so far does not consider the number of documents which do not contain features in all classes; therefore, both the absence and presence of features in classes $P(t_i|c_j)$, $P(\bar{t_i}|c_j)$, $P(t_i|\bar{c_j})$ and $P(\bar{t_i}|\bar{c_j})$ are used in the calculation to reflect the probability of a feature being absent or present in classes. The whole formula makes sure the feature is assigned a value which reflects its usefulness for classification. This is not always the case in some other approaches such as OR, in which the intersection or number of shared features between classes is not taken into account. The maximum and minimum values are between zero and one, and a feature which equally appears in both classes is assigned a minimum value,

[3] We find similar considerations in classical probabilistic information retrieval models (where often the probability that a term occurs in the relevant/non-relevant documents is used) and also in the well-known concept of the inverse document frequency (terms appearing in less documents are deemed good discriminators).

which is zero in case of balanced binary classes. The method adheres to all above mentioned considerations.

## 4 Evaluation

A 10-fold cross-validation of text classification on both unbalanced and balanced datasets of spam emails and SMS have been carried out to test the performance and sensitivity of the IFS method to the number of documents in classes and the distribution ratio of documents between classes. In this experiment we followed [11] and used binary classification, because the results of binary classification reflect the effectiveness of the used measure more directly than that of multi-class classification [11]. Moreover, resolving the binary text classification also means resolving multi-classes [2]. For this experiment, two supervised algorithms namely Support Vector Machines (SVM) and Neural Network (NN) have been employed, as both algorithms are performing well in text classification [1, 12].

### 4.1 Dataset and Performance Measure

Two sets of data have been used in our experiments, namely the enron1 email collection [13], in total 5172 emails, in which 1500 are spam email and 3672 are legitimate emails, and an SMS collection, which was introduced in[14, 15] and consists of 5574 SMS messages, where 4827 are legitimate SMS and 747 spam SMS. From these two datasets we have created 3 test datasets, to enable us to test the sensitivity of the methods with respect to allocation and distribution of documents in classes. The first (*balanced*) test dataset consist of 3000 emails with 1500 spam and randomly chosen 1500 from legitimate emails. The 2nd set (*unbalanced*) consists of 4500 emails in which 1500 are spam emails and randomly chosen 3000 legitimate emails. Finally third test dataset (*unbalanced*) is based on the SMS dataset, which consists of 747 spam and randomly chosen 2576 legitimate ones from a total of 3323 SMS messages. The top $n$ subset of features with highest weight were used in experiment, and $n$ was set to 10, 15, 25, 50, 100, 200, 300, 400 and 500.

The performance of all methods was assessed using the F-measure ($F_1$) based on precision and recall as defined in [2] using micro-averaging.

## 5 Result Analysis

### 5.1 Unbalanced datasets

The F-measures of the 10-fold cross validation of the email and SMS datasets based on SVM and NN classifiers are shown in Figures 1, 2, 3, and 4 respectively. Based on the results, IFS is performing well with both classifiers. Across all used feature sets, IFS is either superior or comparable to others, while in some instances it is shown as second best.

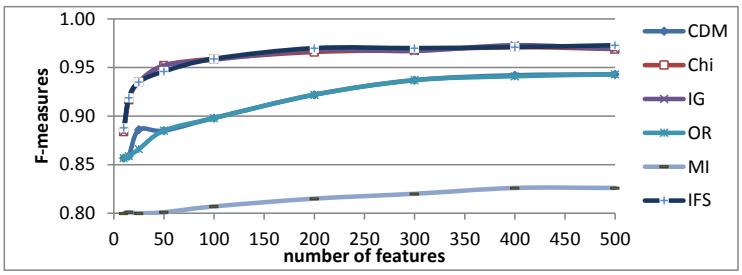


**Fig. 1.** F-measures of SVM with unbalanced emails dataset

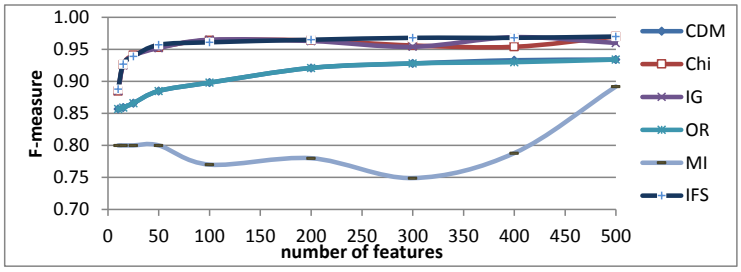


**Fig. 2.** F-measures of NN with unbalanced emails dataset

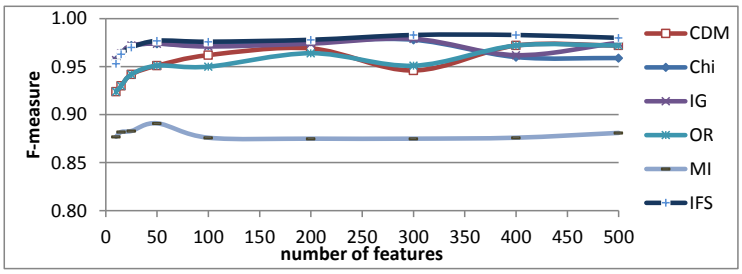


**Fig. 3.** F-measures of NN with unbalanced SMS collection

### 5.2 Balanced datasets

The outcome of F-measures of the 10-fold cross validation for above mentioned methods using SVM and the NN classifier with datasets, 3000 emails with ratio of 1:1 respectively, is shown in Figures 5 and 6. Considering F-measure, IFS yet again is performing well in balanced datasets, and it could be noticed from the results that IFS is superior to others and only in some cases slightly scoring behind. In general, we can conclude that IFS is robust in performing well in both SVM and NN classifiers, compared to other methods.

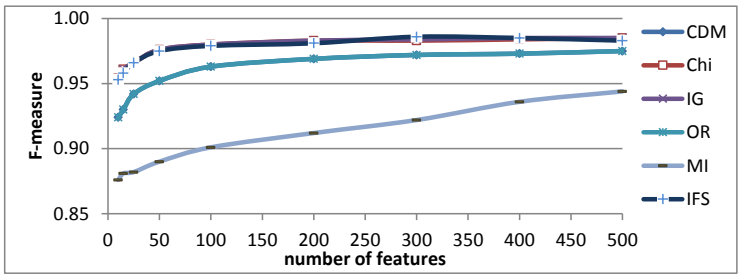


**Fig. 4.** F-measures of SVM with unbalanced SMS collection

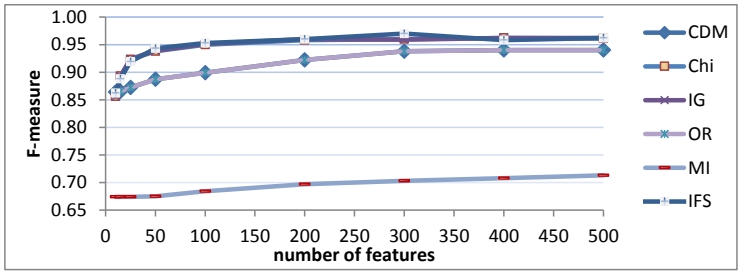


**Fig. 5.** F-measures of SVM with balanced emails dataset

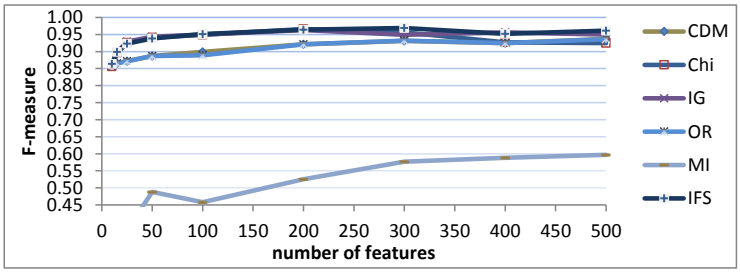


**Fig. 6.** F-measures of NN with balanced emails dataset

## 6 Conclusion

This paper has introduced a new method called IFS to select informative features for classification. The method has been evaluated against some well established feature space reduction methods. Throughout the F-measure values of the 10-fold cross validations result, and processing time, IFS has shown to be robust and often superior, at least competitive compared to other methods. In future work, we aim to test IFS on multi classes datasets. We also aim to improve our method by considering and calculating the frequency of a feature in document and add its weight to the overall usefulness of the feature.

## References

1. Alan S. Abrahams, Eloise Coupey, Eva X. Zhong, Reza Barkhi, and Pete S. Manasantivongs.: Audience targeting by B-to-B advertisement classification: A neural network approach. Expert Systems with Applications, Elsevier, 40(8): 2777-2791, (2013)
2. Fabrizio Sebastiani.: Machine Learning in Automated Text Categorization. ACM computing surveys (CSUR), 34(1):1-47, (2002)
3. Yiming Yang and Jan O Pedersen.: A Comparative Study of Features selection in text categorization. The Fourteenth International Conference on Machine Learning (ICML), 97:412-420 (1997)
4. Jingnian Chen, Houkuan Huang, Shengfeng Tian, and Youli Qu.: Expert Systems with Applications Feature selection for text classification with Naive Bayes. Expert Systems With Applications, Elsevier, 36(3):5432-5435 (2009)
5. George Foreman.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research, 3:1289-1305 (2003)
6. Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying.: In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM, New York, USA, ACM Press, 1-6 (2012)
7. Hiroshi Ogura, Hiromi Amano, and Masato Kondo.: Feature selection with a measure of deviations from Poisson in text categorization. Expert Systems with Applications, Elsevier, 36(3):6826-6832 (2009)
8. Dunja Mladenic and Marko Grobelnik.: Feature selection for unbalanced class distribution and Naive Bayes. In Proceedings of the Sixteenth International Conference on Machine Learning, 258-267 (1999)
9. Yan Xu. A Comparative Study on Feature Selection in Unbalance Text Classification. In Information Science and Engineering (ISISE), 2012 International Symposium, IEEE-CPS, 44-47 (2012)
10. Matthew Chang and Chung Keung Poon.: Using phrases as features in email classification.The Journal of Systems and Software, Elsevier, 82(6): 1036-1045 (2009)
11. Hiroshi Ogura, Hiromi Amano, and Masato Kondo.: Feature selection with a measure of deviations from Poisson in text categorization. Expert Systems with Applications, Elsevier, 36(3):6826-6832 (2009)
12. Yang, Y. and Liu, X., A re-examination of text categorization methods. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99, pp.4249 (1999)
13. Vangelis Metsis, I Androutsopoulos, and G Paliouras.: Spam filtering with naive bayes - which naive bayes?. CEAS, the third conference on Email and Anti-Spam, 27-28 (2006)
14. Gordon V. Cormack, Jose Mara Gomez Hidalgo, and Enrique Puertas Sanz.: Feature engineering for mobile (SMS) spam filtering. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 871-872 (2007)
15. Tiago A Almeida, Jose Maria G Hidalgo, and Akebo Yamakami.: Contributions to the study of SMS spam filtering: new collection and results. In Proceedings of the 11th ACM symposium on Document engineering, 259-262 (2011)