Information Processing and Management 52 (2016) 885–910

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Helmholtz principle based supervised and unsupervised feature selection methods for text mining

Melike Tutkan[a], Murat Can Ganiz[b,*], Selim Akyokuş[a]

[a] Department of Computer Engineering, Doğuş University, Istanbul, Turkey
[b] Department of Computer Engineering, Marmara University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 5 December 2014
Revised 10 November 2015
Accepted 31 March 2016
Available online 5 May 2016

Keywords:
Feature selection
Attribute selection
Machine learning
Text mining
Text classification
Helmholtz principle

ABSTRACT

One of the important problems in text classification is the high dimensionality of the feature space. Feature selection methods are used to reduce the dimensionality of the feature space by selecting the most valuable features for classification. Apart from reducing the dimensionality, feature selection methods have potential to improve text classifiers' performance both in terms of accuracy and time. Furthermore, it helps to build simpler and as a result more comprehensible models. In this study we propose new methods for feature selection from textual data, called Meaning Based Feature Selection (MBFS) which is based on the Helmholtz principle from the Gestalt theory of human perception which is used in image processing. The proposed approaches are extensively evaluated by their effect on the classification performance of two well-known classifiers on several datasets and compared with several feature selection algorithms commonly used in text mining. Our results demonstrate the value of the MBFS methods in terms of classification accuracy and execution time.

1. Introduction

Automatic or semi-automatic processing of large amounts of texts with methods such as text classification and clustering gains importance as the textual content on Internet, social media and the companies increase exponentially. While the traditional data mining focused on structured data sources such as database or warehouse tables, text mining deals with semi-structured or completely unstructured data in the form of natural language. Hence it is very important to preprocess this unstructured data to convert it into a structured format. One of the important differences of the textual data lies in the number of attributes since terms or groups of terms (*n*-grams, phrases, etc.) are used to represent the documents. The common approach to represent documents is to use the frequencies of a bag of words that exists in the whole dataset which generally leads to tens of thousands of attributes. However, this constitutes a severe problem for several machine learning algorithms that are used for popular text mining tasks of text classification and text clustering. For instance, the high dimensionality of the feature space yields to severe sparsity which in turn negatively effects the estimation of the parameters. This is also known as the curse of dimensionality. Feature selection methods are used to reduce the dimensionality of the feature space by selecting the most valuable features for classification. Apart from reducing the dimensionality, feature

* Corresponding author.
E-mail addresses: mtutkan@dogus.edu.tr (M. Tutkan), murat.ganiz@marmara.edu.tr (M.C. Ganiz), sakyokus@dogus.edu.tr (S. Akyokuş).

http://dx.doi.org/10.1016/j.ipm.2016.03.007
0306-4573/

selection methods have potential to improve text classifiers' performance both in terms of accuracy and time. Furthermore, it helps to build simpler and as a result more comprehensible machine learning models.

In this study, we propose novel supervised and unsupervised Meaning Based Feature Selection (MBFS) which effectively reduce high dimensionality of feature space by identifying the most meaningful features (words) in a given context. The meaning measure is previously used in unusual behavior detection and information extraction from small documents (Dadachev, Balinsky, Balinsky, & Simske, 2012), for automatic text summarization (Balinsky, Balinsky & Simske, 2011c), defining relations between sentences using social network analysis and properties of small world phenomenon (Balinsky, Balinsky, & Simske, 2011a), rapid change detection in data streams and documents (Balinsky, Balinsky, & Simske, 2010), for keyword extraction and rapid change detection (Balinsky, Balinsky, & Simske, 2011b), to extractive text summarization by modeling texts and documents as a small world networks (Balinsky, Balinsky, & Simske, 2011d) and for automatic text and data stream segmentation (Dadachev, Balinsky, & Balinsky, 2014). It is based on Helmholtz principle (Balinsky et al., 2011b) and Gestalt theory of human perception (Desolneux, Moisan, & Morel, 2007). According to the Helmholtz principle from Gestalt Theory, an observed geometric structure is perceptually meaningful if it has a very low probability to appear in noise. This indicates that unusual and rapid changes will not happen by chance and they can be immediately perceived. These ideas in image processing suggest that meaningful features and interesting events can be detected by their large deviations from randomness. These ideas can be applied to the textual data i.e., documents by modelling a document by a set of meaningful words together with their level of meaning. A word, which is usually considered as a feature in text mining as mentioned above, is as locally meaningful or important if there is sharp rise in the frequency of a word inside some part of a text document. Meaning evaluates the importance of a term in a part of a document. These parts of the documents which are called container can be paragraphs or sentences or a group of consecutive words (Balinsky et al., 2011b). In our case we adapt the meaning calculations in the context of a class of documents for supervised feature selection purposes to select most meaningful words for each class. We assume that the most meaningful words are better representatives of the class and therefore more valuable for classification process. Additionally, an unsupervised feature selection algorithm is proposed by using meaning calculations in the context of each individual document to select and rank most meaningful words for each document in the whole dataset. The supervised approach can be used with labeled data as a preprocessing tool before text classification, while our unsupervised approach can additionally be used as a preprocessing tool for unsupervised text mining approaches such as text clustering.

The efficiency of our proposed approach is extensively evaluated by observing the effect of the attribute subset selection on the accuracy of two well-known text classifiers on several benchmark textual datasets. We use Multinomial Naive Bayes (MNB) classifier since it is more sensitive to feature selection. Additionally, it is simple, efficient, speed and popular classifier (Rennie, Shih, Teevan, & Karger, 2003). The second classifier we use is the Support Vector Machine (SVM) classifier (Joachims, 1998) which is SMO (Sequential Minimal Optimization) with linear kernel (Platt, 1999). The effect of feature selection on classifier performance using our approaches are compared with several commonly used feature selection methods including chi-square (χ^2) (Yang & Pedersen, 1997), information gain (IG) (Quinlan, 1986), and several Naïve Bayes inspired approaches; MOR, WOR, EOR & CDM (Chen, Huang, Tian, & Qu, 2009), MC_OR (Zhou, Zhao, & Hu, 2004), and much simpler weighting methods of TF-ICF (Ko & Seo, 2000) and TF-IDF (Jones, 1972). The results of our extensive experiments show that MBFS outperforms several others in all datasets in terms of classification accuracy and execution time in many cases.

This paper is organized as follows: Section 2 gives related works and existing feature selection methods and classical meaning measure, Section 3 describes how we use meaning measure for MBFS and some applications in order to increase performance of MBFS, Section 4 describes performance measure and introduces data sets, Section 5 presents the experimental results and analysis and the last section, we give the conclusion.

2. Related work and preliminaries

In this section we briefly review related work about feature selection methods and classification algorithms used in this study. Then, we introduce Helmholtz principle from Gestalt theory and its applications to text mining. The new feature selection methods proposed in this paper is based on meaning measure derived from Helmholtz principle of the Gestalt theory.

2.1. Related work

The most commonly used algorithms in text classification are Naïve Bayes and Support Vector Machines. In our experiments, we use Multinomial Naïve Bayes (MNB) (Rennie et al., 2003) and Sequential Minimal Optimization (SMO) (Platt, 1999) version of Support Vector Machines (SVM) classifiers (Joachims, 1998) that are implemented in WEKA (Hall et al., 2009) machine learning toolkit. We use these classifiers in order to measure the effect of feature subset selection on the classification accuracy.

There are many studies on Naïve Bayes (NB) (Lewis, 1998) because understanding and implementing NB is easier than other classifiers moreover it has higher speed. MNB is firstly proposed by (McCallum & Nigam, 1998) and then discussed, analyzed, improved by (Rennie et al., 2003). Documents are represented by number of word occurrences from each

document. In this model, the order of words are not important. This model is also known as unigram model (McCallum & Nigam, 1998).

SVM is a more complex classifier than MNB. SVM is a discriminative and binary classifier. Between two classes, SVM finds optimal hyper plane by maximizing the margin among the closest points of classes. SMO is one of the approaches that are used in the learning phase of the SVM. Although other SVM learning algorithms use numerical quadratic programming (QP) as an inner loop, SMO uses an analytic QP step. Because of this, SMO is simple, easy to implement, is often faster and has better scaling properties than standard algorithms which include analytic QP step (Platt, 1999). In general the linear kernel works well for the text classification domain probably due to the large number of features.

The high dimensionality of feature space (Joachims, 1998) and feature redundancy (Joachims, 2001) are two important problems in text classification. Not all the features are relevant or beneficial for text classification. Some of these features may include noise and therefore reduce the classification accuracy. Moreover, the high dimensionality of the feature space can slow down the classification process. Therefore, it is desirable to select most relevant features and eliminate the noisy ones from this high dimensional feature space. Feature selection can improve the scalability, efficiency and accuracy of a text classifier (Chen et al., 2009). It is common to use attribute selection methods in the preprocessing phase of text mining. There are several approaches for reducing the dimensionality of the feature space. There are three main categories of feature selection methods; filter, wrapper and embedded models. These methods can be applied in supervised, unsupervised or semi-supervised settings (Liu, Motoda, Setiono, & Zhao, 2010). Feature selection in text mining is an important and active research area with several recent studies (Baccianella, Esuli, & Sebastiani, 2013) (Uysal & Gunal, 2012) (Yang, Liu, Liu, Zhu, & Zhang, 2011) (Yang, Liu, Zhu, Liu, & Zhang, 2012) (Shang, Li, Feng, Jiang, & Fan, 2013) (Zhou, Hu, & Guo, 2014). There are also studies that use LDA (Latent Dirichlet Allocation) for feature selection. LDA is a method that allows the construction of a model of topics that exist in a document ranked by term relevance (Blei, Ng, & Jordan, 2003) (Tasci, Gungor 2009).

One of the most popular feature selection methods in text classification is the Information Gain (IG) (Yang & Pedersen, 1997). The formulation of IG is given in (1) where w represents the feature which can be a word or a term, c_i represents the ith class, $P(c_i)$ represents probability of class c_i, $P(c_i|w)$ represents conditional probability of class c_i for presence of given feature w, $P(c_i|\bar{w})$ represents conditional probability of class c_i for absence of given feature w, $P(w)$ represents probability of presence of w and $P(\bar{w})$ represents probability of absence. IG is proposed by (Quinlan, 1986) which is based on information theory by (Shannon & Weaver, 1949) who is study on information content of messages.

$$IG(w) = -\sum_i P(c_i)\log_2 P(c_i) + P(w)\sum_i P(c_i|w)\log_2 P(c_i|w) + P(\overline{w})\sum_i P(c_i|\overline{w})\log_2 P(c_i|\overline{w}) \quad (1)$$

The IG inspired several other feature selection methods due to its highest performance and simplicity. One of these methods is is Gain Ratio (GR) (Quinlan, 1993) which is a normalized extension of IG. In another study (Lee & Lee, 2006), authors proposed new feature selection method based on IG and divergence-based feature selection called maximal marginal relevance (MMR) approach. Their method selects each feature according to a combined criterion of IG and novelty on information. The latter one measures the degree of dissimilarity between feature being considered and the previously selected features. The idea behind the MMR based feature selection is to reduce redundancy between features without reducing the IG in the process of selecting features for text classification (Lee & Lee, 2006).

A similar feature selection method is Gini Index (GI). It is one of the many methods used in Decision Tree algorithm as a feature splitting criteria along with IG and GR. An improved version of Gini Index is proposed as a feature selection method on text classification domain and it is being reported to be a promising method. They used SVM and K-Nearest Neighbor (k-NN) algorithms the measure the performance of GI and compare with other feature selection methods (Shang et al., 2007).

Another popular feature selection method in text classification is Chi-square (χ^2) (Yang & Pedersen, 1997). The formulation of χ^2 is given in (2) where w represents feature, c represents class, A denotes observed frequency of each state feature w and class c, E denotes expected frequency of each state feature w and class c. Basically, χ^2 statistic measures the lack of independence between term w and class c. χ^2 is used for feature selection with the formula (3) where $P(c_i)$ represents probability of class c_i and $\chi^2(w,c_i)$ represents class specific χ^2 score of feature w (Chen & Chen, 2011).

$$\chi^2(w,c) = \sum_w \sum_c \frac{(A_{wc} - E_{wc})^2}{E_{wc}} \quad (2)$$

$$\chi^2(w) = \sum_i P(c_i)\chi^2(w,c_i) \quad (3)$$

Odds Ratio (OR) method and its variations are also widely used for feature selection. Traditional Odds Ratio (OR) (Mladenic & Grobelnik, 1999) is extended for multi-class domains and is named as Extended Odds Ratio (EOR) which is described in (4), Weighted Odds Ratio(WOR) which is described in (5), Multi-class Odds Ratio (MOR) which is described in (6), Class Discriminating Measure (CDM) which is described in (7) (Chen et al., 2009). A similar method called Multi Class Ratio (MC_OR) is proposed by Zhou (Zhou et al., 2004) and described in (8). MOR is quite similar to MC_OR. The only difference between MC_OR and MOR is that MC_OR weights each term with class distribution and give more emphasis to features

appearing in large classes. Since the large classes are likely to contribute much more valuable attributes to the classification. In all the formulas below ((4) to (8)), $P(w|c_j)$ is the probability that word w occurs in class j, which can be calculated by dividing the total occurrence probability of term w in the documents of the class to the total occurrence probability of all terms in the class. $P(w|\bar{c}_j)$ is the probability that word w does not occurs in class j, also occurs in all classes. $P(c_j)$ is the class prior probability calculated by dividing the number of documents in that class to the total number of documents in all classes.

$$EOR(w) = \sum_j \frac{\log P(w|c_j)(1 - P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1 - P(w|c_j))} \tag{4}$$

$$WOR(w) = \sum_j P(c_j) \frac{\log P(w|c_j)(1 - P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1 - P(w|c_j))} \tag{5}$$

$$MOR(w) = \sum_j \left| \frac{\log P(w|c_j)(1 - P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1 - P(w|c_j))} \right| \tag{6}$$

$$CDM(w) = \sum_j \left| \frac{\log P(w|c_j)}{\log P(w|\overline{c}_j)} \right| \tag{7}$$

$$MC\_OR(w) = \sum_j P(c_j) \left| \frac{\log P(w|c_j)(1 - P(w|\overline{c}_j))}{\log P(w|\overline{c}_j)(1 - P(w|c_j))} \right| \tag{8}$$

The common term weighting approach of TF-IDF (Term Frequency – Inverse Document Frequency) can also be used as an unsupervised means of feature selection whose formula defined in (10). tf_w represents the term frequency of the term w in the document and IDF is inverse of the document frequency of the term in the dataset (IDF) whose formula defines in (9) where $|D|$ denotes number of documents; df_w denotes number of documents which contain term w. TF denotes that word w occurs in document d_i. This well-known approach to term weighting was proposed in (Jones, 1972). TF-IDF has proved extraordinarily robust and difficult to beat, even by much more carefully worked out models and theories (Robertson, 2004). In order to use TF-IDF as an unsupervised feature selection method, we calculate TF-IDF scores for each term on each document and take the average of these values with the approach is introduced in Section 3.2. Then TF-IDF scores of features are sorted and selected the top R features from sorted set in order to use in classification. Since it is a very fundamental approach we use it in our experiments.

$$IDF(w) = \frac{|D|}{df_w} \tag{9}$$

$$TF - IDF(w, d_i) = tf_w \cdot \log(IDF(w)) \tag{10}$$

A similar but supervised version of the TF-IDF is called TF-ICF (Term Frequency – Inverse Class Frequency). TF-ICF, whose formula given in (12), is a supervised approach which uses class information (Ko&Seo, 2000). In (11), $|C|$ denotes number of classes and cf_w denotes number of classes which contain term w. It is simply calculated by dividing the total number of classes to the number of classes that this term w occurs in classes and as in TF-IDF, tf_{wj} denotes that word w occurs in class c_j (Lertnattee & Theeramunkong, 2004). In order to use TF-ICF as a supervised feature selection method, we calculate TF-ICF scores for each term on each class. We apply rank approach which is introduced in Section 3.1. Then TF-ICF scores of features are sorted and selected the top R features from sorted set in order to use in classification.

$$ICF(w) = \frac{|C|}{cf_w} \tag{11}$$

$$TF - ICF(w, c_j) = \sum_{d \in c_j} tf_{wj} \cdot \log(ICF(w)) \tag{12}$$

There are several recent studies related to the feature selection methods for text classification which are proposed in years from 2011 to 2014. In 2011, a novel study on feature selection is called Bi-Test which is based on binomial distributions is proposed in (Yang et al., 2011). Bi-Test uses binomial hypothesis testing to estimate whether probability of a feature belonging to spam or ham messages by satisfying a given threshold. They evaluate their method on six different spam corpora using NB and SVM classifiers. In 2012, a new feature selection algorithm, Distinctive Feature Selector (DFS) (Uysal & Gunal, 2012) is proposed. DFS is a filter based probabilistic method for feature selection. Basically, DFS selects distinctive features while removing uninformative ones considering certain requirements on term characteristics. Another feature selection method proposed this year is called Comprehensively Measure Feature Selection (CMFS) (Yang et al., 2012). CMFS calculates significance of terms from both inter-class and intra-class. They use NB and SVM classifiers and three text datasets for evaluation. In 2013, a new method on feature selection which is called Maximizing Global Information Gain (MGIG) is proposed

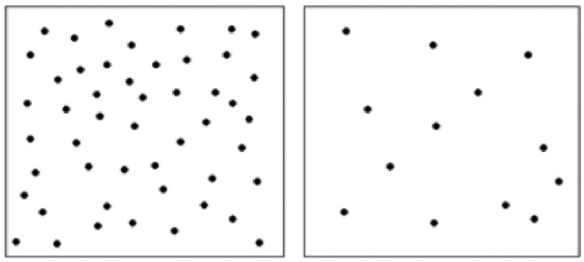

Fig. 1. The Helmholtz principle in human perception (adopted from (Balinsky et al., 2011b)).

in (Shang et al., 2013). First, they proposed Global Information Gain (GIG) metric, then use it for feature selection for text classification. The feature selection method MGIG is based on GIG. GIG is higher-order feature selection metric in addition, it can avoid redundancy naturally. GIG properties are firstly informative which means that selected feature should have more information with class label, secondly representative which means that tried to guarantee selection of informative features and exclusion of outliers and the lastly distinctive which means that tried to select features that should maintain diversity. They use six dataset and two classifiers, SVM and NB. In 2014, k-means clustering algorithm based method is used for feature selection in text classification (Zhou et al., 2014).

2.2. On Helmholtz principle from Gestalt theory and its applications to text mining

According to Helmholtz principle from Gestalt theory in image processing; "observed geometric structure is perceptually meaningful if it has a very low probability to appear in noise" (Balinsky et al., 2011b). This means that events have large deviation from randomness or in other words noise can be noticed easily by humans. This can be illustrated in Fig. 1. In the left hand side of Fig. 1, there is a group of five aligned dots but it is not easy to notice it due to the high noise. Because of the high noise, i.e. large number of randomly placed dots, the alignment probability of five dots increases. On the other hand, if we remove the number of randomly placed dots considerably, we can immediately perceive the alignment pattern in the right hand side image since it is very unlikely to happen by chance. This phenomenon means that unusual and rapid changes will not happen by chance and they can be immediately perceived.

As an example, assume you have unbiased coin and it is tossed 100 times. Any 100-sequence of heads and tails can be generated with probability of $(½)^{100}$ and following Fig. 2 is generated where 1 represents heads and 0 represents tails (Balinsky et al., 2010).

First sequence, s_1 is expectable for unbiased coin but second output, s_2 is highly unexpected. This can be explained by statistical physics where we observe macro parameters but we don't know the particular configuration. We can use the expectation calculations (Balinsky et al., 2010).

A third example is known as birthday paradox in literature. There are 30 students in a class and we would like to calculate the probability of two students having the same birthday and how likely or interesting this is. Firstly, we assume that birthdays are independent and uniformly distributed over the 365 days of a year. Probability P_1 of all students having different birthday in the class is calculated in (13) (Desolneux et al., 2007).

$$P_1 = \frac{365x364x\ldots x336}{365^{30}} \approx 0.294 \tag{13}$$

The probability P_2 of at least two students born on same day is calculated in (14) .This means that approximately 70% of the students can have the same birthday with another student in the class of 30 students.

$$P_2 = 1 - 0.294 = 0.706 \tag{14}$$

When probability calculations are difficult to compute, we compute expectations. The expectation of number of 2-tuples of students in a class of 30 is calculated as in (15). This means that on the average, 1.192 pairs of students have the same birthday in the class of 30 students and therefore it is not unexpected. However the expectation values for 3 and 4 students having the same birthday, $E(C_3)\approx 0.03047$ and $E(C_4) \approx 0.00056$, which are much smaller than one, indicates that these events will be unexpected (Desolneux et al., 2007).

$$E(C_2) = \frac{1}{365^{2-1}}\begin{pmatrix} 30 \\ 2 \end{pmatrix} = \frac{1}{365}\frac{30!}{(30-2)!2!} = \frac{30x29}{2x365} \approx 1.192 \tag{15}$$

$$s_1 = 10101\ 11010\ 01001\ \ldots\ 00111\ 01000\ 10010$$
$$s_2 = \underbrace{111111111\ldots111111}_{50\ \text{times}}\ \underbrace{000000000\ldots000000}_{50\ \text{times}}$$

Fig. 2. The Helmholtz principle in human perception (adopted from (Balinsky et al., 2010)).

In summary, the above mentioned principles indicate that meaningful features and interesting events appears in large deviations from randomness. Meaningfulness calculations basically correspond to the expectation calculations and they are justifiable by standard mathematical and statistical physics approaches (Balinsky et al., 2011b).

In the context of text mining, the textual data consist of natural structures in the form of sentences, paragraphs, documents, and topics. In (Balinsky et al., 2011b), the authors attempt to define meaningfulness of these natural structures using the human perceptual model of Helmholtz principle from Gestalt Theory. Modelling the meaningfulness of these structures is established by assigning a meaning score to each word or term. Their new approach to meaningful keyword extraction is based on two principles. The first one state that these keywords which are representative of topics in a data stream or corpus of documents should be defined not only in the document context but also the context of other documents. This is similar to the TF-IDF approach. The second one states that topics are signaled by "unusual activity", a new topic can be detected by a sharp rise in the frequencies of certain terms or words. They state that sharp increase in frequencies can be used in rapid change detection. In order to detect the change of a topic or occurrence of new topics in a stream of documents, we can look for bursts on the frequencies of words. A burst can be defined as a period of increased and unusual activities or rapid changes in an event. A formal approach to model "bursts" in document streams is presented in (Kleinberg, 2003). The main intuition in this work is that the appearance of a new topic in a document stream is signaled by a "burst of activity" with certain features rising sharply in frequency as the new topic appears.

Based on the theories given above, new methods are developed for several related application areas including unusual behavior detection and information extraction from small documents (Dadachev et al., 2012), for text summarization (Balinsky, Balinsky & Simske, 2011c), defining relations between sentences using social network analysis and properties of small world phenomenon (Balinsky, Balinsky & Simske, 2011a) and rapid change detection in data streams and documents (Balinsky et al., 2010), for keyword extraction and rapid change detection (Balinsky et al., 2011b), to extractive text summarization by modeling texts and documents as a small world networks (Balinsky et al., 2011d) and for automatic text and data stream segmentation (Dadachev et al., 2014). These approaches make use of the fact that meaningful features and interesting events come into view if their deviations from randomness are very large.

The motivating question in these studies is "if the word w appears m times in some documents is this an expected or unexpected event?" (Balinsky et al., 2011b). They assume that S_w is the set of all words in N documents and a particular word w appears K times in these documents. They use random variable C_m to count m-tuple of the elements of S_w appears in the same document. Following this they calculate expected value of C_m under the assumption that the words are independently distributed to the documents. C_m is calculated by using random variable $X_{i1,i2\ldots im}$ which indicates if words $w_{i1,\ldots,}w_{im}$ co-occurs in the same document or not based on this the expected value $E(C_m)$ can be calculated as in (17) by summing the expected values of all these random variables for all the words in the corpus.

$$C_m = \sum_{1\le i1<\ldots<im\le K} X_{i1,\ldots,im} \tag{16}$$

$$E(C_m) = \sum_{1\le i1<\ldots<im\le K} E(X_{i1,\ldots,im}) \tag{17}$$

The random variable $X_{i1,i2\ldots im}$ can only take two values which are one and zero. As a result the expectation of this random variable which shows if these m words co-occurs in the same document can be calculated in formula (18). In this formula N is the total number of documents. "If in some documents the word w appears m times and $E(C_m)<1$ then it is an unexpected event" (Balinsky et al., 2011b).

$$E(X_{i1,\ldots,im}) = \frac{1}{N^{m-1}} \tag{18}$$

As a result $E(C_m)$ can simply be expressed as in formula (19) and this expectation actually corresponds to number of false alarms (NFA) of m-tuple of word w which is given in formula (20). This corresponds to the number of times m-tuple of the word w occurs by chance (Balinsky et al., 2011b). Based on this, in order to calculate the meaning of a word w which occurs m times in a context (document, paragraph, sentence), we can look its NFA value. If the NFA (expected number) is less than one than occurrence of m times can be considered as a meaningful event because it is not expected by our calculations but it is already happened. Therefore, word w can be considered meaningful or important word in the given context.

$$E(C_m) = \binom{K}{m} \frac{1}{N^{m-1}} \tag{19}$$

Based on the NFA, the meaning score of words are calculated using (20) and (21) in (Balinsky, Balinsky & Simske, 2011a).

$$NFA(w,P,D) = \binom{K}{m} \frac{1}{N^{m-1}} \tag{20}$$

$$Meaning(w,P,D) = -\frac{1}{m}\log NFA(w,P,D) \tag{21}$$

In these formulas, NFA means Number of False Alarms, w represents a word, P represents a part of document such as a sentence or paragraph, and D represents whole document. The word w appears m times in P and K times in D. $N=L\ /\ B$

where L is length of *D* and *B* is length of *P* in words (Balinsky, Balinsky & Simske, 2011a). In Meaning formula log of NFA is used based on the observation that NFA values can be exponentially large or small (Balinsky, Balinsky & Simske, 2011a).

In (Altınel, Ganiz, & Diri, 2015), authors used meaning calculations that are formulated above in a different setting; to build a semantic kernel for SVM for text classification. In this study, similar to ours, they calculated the meaning values of terms in the context of classes to form a class by term matrix. This matrix is, in turn, multiplicated by its inverse to obtain a term by term semantic matrix which shows the semantic relatedness between terms. The semantic matrix is used in semantic kernel calculations. Experimental results show significant increase in the performance of SVM when used with the above mentioned semantic kernel compare to the traditional linear kernel.

3. Approach

In this study we propose a new method for feature selection called Meaning Based Feature Selection (MBFS) for text mining based on meaning measure which is previously used for rapid change detection, keyword extraction, text summarization (Balinsky et al., 2010;2011a;2011b;2011c;2011d) (Dadachev et al., 2012; 2014). Meaning measure is based on Helmholtz principle from the Gestalt theory. In these studies, a text document is modelled by a set of meaningful words together with their meaning scores. A word is accepted as meaningful or important if the occurrence term frequency of a word in a document is unexpected by considering the term frequencies of this word in all the documents in our corpus. The method can be applied on a single document or on a collection of documents to find meaningful words inside each part or context (paragraphs, pages, sections or sentences) of a document or a document inside of a collection of documents (Balinsky, Balinsky & Simske, 2011a). The calculations and formulas are described in detail in the previous Section 2.1. We adapt the meaning measure idea and use it for supervised and unsupervised feature selection in order to calculate the importance of words in a given collection of documents. For supervised feature selection, we use class based meaning scores for feature selection before the classification process. In our case, the context is a class of documents. Therefore, our approach finds the meaning scores of words for each class. Then these class based word lists are ordered by meaning scores to select most meaningful words for each class. These class based meaning ordered lists are than combined into a single term list by using three different approaches: Rank, Max, and Average.

For given a dataset *S*, we find the meaning of a feature *w* inside a class *c* or a single document *d*. We apply two different approaches. First one is a supervised approach since it makes use of the class information. We call it Supervised MBFS. Second one is an unsupervised MBFS since it doesn't make use of the class information. We call it Unsupervised Meaning Feature Selection.

3.1. Supervised approach

In supervised approach, we use a class of documents as our basic unit or context in order to calculate meaning scores for words. In this approach meaning measure basically shows how expected a particular words' frequency is in a class of documents compare to the other classes of documents. If it is unexpected then meaning measure results in high meaning scores. In this aspect it is similar to the Multinomial Naïve Bayes in which the all the documents in a class is merged into a single document and then the probabilities are estimated from this one large class document. It also bears similarities to TF-ICF approach in which the term frequencies are normalized using the class frequencies.

In supervised meaning measure which is given in formulas (22) through (27), parameter c_j represents documents which belong to class *j* and *S* represents the complete training set. Assume that a feature *w* appears *k* times in the dataset *S*, and *m* times in the documents of class c_j. The length of dataset (i.e. training set) *S* and class c_j measured by the total term frequencies is *L* (22) and *B* (23) respectively. *N* is the rate of the length of the dataset and the class which calculate in (24). Based on these the number of false alarms (*NFA*) is defined in (25)

$$L = \sum_{d \in S} \sum_{w \in d} tf_w \tag{22}$$

$$B = \sum_{d \in c_j} \sum_{w \in d} tf_w \tag{23}$$

$$N = \frac{L}{B} \tag{24}$$

$$NFA(w, c_j, S) = \begin{pmatrix} k \\ m \end{pmatrix} \cdot \frac{1}{N^{m-1}} \tag{25}$$

Based on NFA, the meaning score of the word *w* in a class c_j is defined as:

$$meaning(w, c_j) = -\frac{1}{m} \log\ NFA(w, c_j, S) \tag{26}$$

In order to simplify the calculations meaning formula can be re-written as:

$$meaning(w, c_j) = -\frac{1}{m} \log \binom{k}{m} - [(m-1)\log N] \tag{27}$$

The larger the meaning score of a word w in a class c_j means that the given word w is more meaningful, significant or informative for that class. Strictly speaking, the words with larger meaning scores correspond to more meaningful, significant or informative words for that class. However, for feature selection we need a way to combine these class-based scores into one and select top R features. In order to do this we employ three different approaches: Rank, Average, and Maximum.

- Rank: In this approach, we sort the features by using their meaning scores for each class. For instance, the rank of the first element on each sorted list will be 1 and the last element will be the dictionary size. We use rank of the features in each class instead of their meaning scores. When combining these class based lists into a single feature list, for each feature we select the highest rank among all classes as in (28). This approach is called Supervised Meaning Rank (SMR).

$$score(w) = \max_{c_j \in C} \left(Rank(w, c_j) \right) \tag{28}$$

- Average: We take the average of class based meaning scores of a given feature w (29). $|C|$ denotes that number of class. This approach is called Supervised Meaning Average (SMA).

$$score(w) = \left(\left(\sum_{i=1}^{|C|} meaning(w, c_j) \right) / |C| \right) \tag{29}$$

- Max: We take the maximum of class based meaning scores of a given feature w as in (30). This approach is called Supervised Meaning Maximum (SMM).

$$score(w) = \max_{c_j \in C} \left(meaning(w, c_j) \right) \tag{30}$$

After applying one of these methods, we sort the scores of features and select the top R features from sorted set in order to use in classification.

3.2. Unsupervised approach

In unsupervised approach, we use a single document as our basic unit or context in order to calculate meaning scores for words (features). In this approach meaning calculations basically shows how expected a particular words' frequency is in a document compare to the other documents in the dataset. If it is unexpected then meaning measure results in high meaning scores. This is similar to the approach in (Balinsky et al., 2011b). In this approach, d_j is the jth document in dataset S. Assume that a word w appears k times in dataset S and m times in document d_j. The length of dataset S and document d_j measured by the sum of term frequencies is L (31) and B (32) respectively. N is the rate of the length of the dataset and a single document which calculated in (33). The number of false alarms (NFA) in this setting is defined in (34)

$$L = \sum_{d \in S} \sum_{w \in d} tf_w \tag{31}$$

$$B = \sum_{w \in d} tf_w \tag{32}$$

$$N = \frac{L}{B} \tag{33}$$

$$NFA(w, d_j, S) = \binom{k}{m} \cdot \frac{1}{N^{m-1}} \tag{34}$$

A measure of meaning of the word w inside a document d_j is defined as:

$$meaning(w, d_j) = -\frac{1}{m} \log\, NFA(w, d_j, S) \tag{35}$$

In order to simplify the calculations meaning formula can be re-written as:

$$meaning(w, d_j) = -\frac{1}{m} \log \binom{k}{m} - [(m-1)\log N] \tag{36}$$

In this setting, we have a meaning score for word w for all the documents in the dataset. Similar to the supervised approach above, we can combine these scores using Rank, Average, and Max which are called Unsupervised Meaning Rank (UMR), Unsupervised Meaning Average (UMA), and Unsupervised Meaning Max (UMM), respectively. After applying either

one of these methods, we sort the scores of the features and select the top R features from sorted set in order to use in classification. Please note that this approach does not make use of class information and therefore it can be used as a preprocessing method for unsupervised text mining algorithms such as text clustering.

The larger the meaning score of a word w in a single document d_j the more meaningful, significant or informative that word is for that document. However, for feature selection we need a way to combine these document-based scores into one and select top R features. In order to do this we employ three different approaches: Rank, Average, and Maximum.

- Rank: In this approach, we sort the features by using their meaning scores for each document. For instance, the rank of the first element on each sorted list will be 1 and the last element will be the dictionary size. We use rank of the features in each document instead of their meaning scores. When combining these document based lists into a single feature list, for each feature we select the highest rank among all documents as in (37). This approach is called Unsupervised Meaning Rank (UMR).

$$score(w) = \max_{d_j \in S} \left(Rank(w, d_j) \right) \tag{37}$$

- Average: We take the average of document based meaning scores of a given feature w (38). $|S|$ denotes that the number of documents in training set or corpus. This approach is called Unsupervised Meaning Average (UMA).

$$score(w) = \left(\sum_{i=1}^{|S|} meaning(w, d_j)\right)/|S| \tag{38}$$

- Max: We take the maximum of document based meaning scores of a given feature w as in (39). This approach is called Unsupervised Meaning Maximum (UMM).

$$score(w) = \max_{d_j \in S} \left(meaning(w, d_j) \right) \tag{39}$$

After applying one of these methods, we sort the scores of features and select the top R features from sorted set in order to use in classification.

3.3. Performance improvements for meaning based feature selection

When the dataset size is large, i.e. including large number of documents with many words, calculating combinations or factorials becomes problematic due to the scale of the resulting numbers which grows exponentially. This causes overflow and memory problems in the software implementations. In order to avoid this, we use log gamma function instead of factorials in calculating the m combinations of k (40). Please remember that a word w appears k times in dataset S, and m times in document d_j in the unsupervised setting or m times in class c_j in the supervised setting. Especially in supervised setting, m can be quite large. The idea of using log gamma function comes from Binomial Coefficient theory. The binomial coefficient is the number of ways of picking k unordered outcomes from m possibilities, which is also known as a combination or combinatorial number. The combination can be calculated with gamma function ($\Gamma(n)$) (41) (Press, Flannery, Teukolsky, & Vetterling, 1992).

$$\binom{k}{m} = \frac{k!}{m!(k-m)!} \tag{40}$$

$$\binom{k}{m} = \exp(\ln \Gamma(k+1) - \ln \Gamma(m+1) - \ln \Gamma(k-m+1)) \tag{41}$$

However, in supervised setting in which the m corresponds to the sum of term frequencies of word w for documents in the class, we can have overflow and performance problems although we use the log gamma function. In order to avoid this, k and m are normalized into [0150] space. This range of 150 is selected empirically, based on the observation that a factorial of a number larger than 150 converges to infinity in our implementation.

4. Experiment setup

The performance of the proposed MBFS methods is evaluated by observing their effect on the performance of the text classifiers. This approach is commonly used in feature selection studies (Schneider, 2004)(Lee & Lee, 2006). Therefore we employed several benchmark datasets and state-of-the-art text classifiers that are commonly used in text classification studies in our experiments. We assess the effectiveness of our MBFS methods by observing the performance of MNB and SVM classifiers at different feature sizes. We compare our approach with several other feature selection methods which are commonly used in text classification. We employ 10-fold cross-validation approach in evaluating the classifiers. In order to measure the performance of the classifiers we use accuracy (42) and macro-F1 (46) which is the arithmetic mean of F1

Table 1
Datasets.

Dataset	|C|	|S|	|F|
tweet65k	2	64,204	9905
ohscal	10	11,162	11,466
new3s	44	9558	26,833
la1s	6	3204	13,169
wap	20	1560	8461
1150haber	5	1150	6656

measures (45) (Han, 2006). F1 is basically the harmonic mean of Precision and Recall measures and these are in turn calculated by using true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn). The tp and tn shows the correct classifications of positive and negative class respectively while fp and fn shows the misclassifications.

$$Accuracy = (tp + tn)/(tp + fp + tn + fn) \tag{42}$$

$$P = tp/(tp + fp) \tag{43}$$

$$R = tp/(tp + fn) \tag{44}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{45}$$

$$macro - F_1 = \frac{\sum_{c=1}^{|C|} F_{1c}}{|C|} \tag{46}$$

4.1. Datasets

In our experiments, we use the datasets given in Table 1. In Table 1, |C| is the number of classes, |S| is the number of documents and |F| gives the number of features. The “tweet65k” dataset is a modified subset of “Sentiment140” dataset (Go, Bhayani, & Huang, 2009) containing a total of 64,204 tweets of which 34,233 of them are negative and 29,971 of them are positive. The “Ohscal” dataset is a part of the OHSUMED (Hersh, Buckley, Leone, & Hickam, 1994) collection which contains 11,162 documents in ten categories under the topics: Antibodies, Carcinoma, DNA, In-Vitro, Molecular-Sequence-Data, Pregnancy, Prognosis, Receptors, Risk-Factors, Tomography (Han & Karypis,2000). The “wap” dataset includes a subset of Yahoo web pages from WebACE (Han et al., 1998) project which contains 1560 documents in twenty categories: Art, Business, Cable, Culture, Entertainment, Film, Health, Industry, Media, Multimedia, Music, Online, People, Politics, Review, Sports, Stage, Technology, Television, and Variety (Han & Karypis,2000). The “new3s” dataset, on the other hand, is composed of news articles from San Jose Mercury newspaper which contains 9558 documents in 44 classes (Han & Karypis, 2000). The “la1s” dataset is collected from Los Angeles Times news articles and it contains 3204 documents in six categories under the topic of Entertainment, Financial, Foreign, Metro, National and Sports (Han & Karypis, 2000). Our last dataset “1150haber” is in Turkish and is composed of 1150 newspaper articles in five categories, collected from one of the mainstream Turkish newspapers (Amasyalı & Beken, 2009). This is also commonly used in text classification studies (Ganiz, George, & Pottenger, 2011) (Poyraz, Kilimci, & Ganiz, 2014). Before applying feature selection algorithms, we apply stemming and stopword filtering.

5. Experiment results and discussion

In this section, we present and discuss experimental results of MBFS methods. Before going into feature selection evaluation, we present the most meaningful terms (actually the stems of the words since we apply stemming) for each class obtained by sorting by meaning scores in descending order in Section 5.1. In Section 5.2, we compare MBFS methods with existing well-known feature selection methods with selected feature sizes ranging from 500 to 10,000 depending on the data set. Section 5.3 compares new SMR feature selection method with IG on smaller feature sizes ranging from 10 to 500. Section 5.4 gives a comparison of new supervised and unsupervised MBFS methods.

5.1. Most meaningful terms

To get a feeling and practical understanding of MBFS method, Tables 2–5 gives a listing of the most meaningful words (keywords or topics) in selected five categories in “ohscal”, “la1s”, “wap” and “1150haber” datasets. For each dataset we only select five classes due to space constraints. Experts or people studying in the fields related with these datasets can easily

Table 2
The most meaningful 10 words for each class on "ohscal" dataset.

DNA	Pregnancy	Prognosis	Risk-factors	Tomography
mtdna	preeclampsia	psa	cvd	Positron
tem	amniocentesi	prism	radon	Spect
hprt	ritodrin	transscler	fontan	Pseudocyst
tetraploid	chorioamnion	iol	vietnam	Hrct
hybridis	oligohydramnio	esotropia	urolithiasi	Collim
Norfloxacin	parturi	barthel	refuge	Ultrafast
transconjug	tocolysi	dlcl	player	Petrou
hyperprolifer	eclampsia	tdt	ivdu	Rcbf
polyploid	polyhydramnio	vitreoretin	driver	Ptsm
meca	partum	subretin	hyperlipidemia	discography

Table 3
The most meaningful 10 words for each class on "la1s" dataset.

Entertainment	Financial	Foreign	National	Sports
aleen	jefferi	nato	teamster	ncaa
macmin	milken	settler	panetta	playof
fugard	fuji	tripoli	alexandria	clipper
mcguin	icahn	mig	counterfeit	lendl
quartet	ast	shevardnadz	lackner	newswir
roseann	gaf	warsaw	wiretap	titan
bogosian	shamrock	galicia	brownsvill	oiler
rehears	banfill	imhausen	dominicci	knick
ensembl	opec	walesa	riba	scorer
mozart	xidex	rabta	darman	socker

Table 4
The most meaningful 10 words for each class on "wap" dataset.

Health	Entertainment	Film	Industry	Art
vitamin	casino	Affleck	paxson	Sculptor
protein	mirag	Northam	wga	necklace
vaccin	trump	Unspool	corp	Gogh
calcium	legion	Hofler	westinghous	Gardner
hormon	hilton	Zeitgeist	murdoch	Galleri
obes	farm	Gross	cineplex	Exhibit
mutat	resort	Kull	digest	Stolen
antibiot	atlant	Beaver	benkoe	Michelangelo
estrogen	miami	Regina	warmer	Rembrandt
intak	airlin	Conqueror	fairfax	Hitler

Table 5
The most meaningful 10 words for each class on "1150haber" dataset.

Ekonomi	Magazin	Sağlik	Siyasi	Spor
cari	pekin	hasta	anaya	takım
borsa	hande	tümör	annan	futbo
açığı	pekka	ultra	kerkü	maçta
döviz	sosye	ışınl	dgm	luces
varil	ataiz	cildi	aihm	sahad
unakı	madon	lazer	mhp	orteg
tahvi	laila	kanam	mgk	stadı
mevdu	ajda	enfek	laikl	dk
ötv	dizid	menop	bayar	tribü
venez	çapkı	cilt	şaron	defan

recognize that the selected words are really most common, significant and informative words in categories of each given dataset.

For instance, in Table 4, the most meaningful words of the class "Health" include words: vitamin, protein, calcium, and obesity which are important concepts in the health domain. We can see the similar results for classes "Entertainment", "Film", "Industry" and "Art" where important domain concepts for each class are highly ranked and quite distinguishable from the terms of other classes. The case is also valid for even "ohscal" dataset whose classes are quite close to each other

Table 6
Comparison of different feature selection methods using the accuracy of MNB on "tweet65k" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	74.45	73.97	75.11	**75.12**	57.15	56.69	56.64	56.69	47.65	73.98	53.85
1000	75.31	75.18	75.50	**75.53**	60.26	57.63	57.63	57.63	43.76	75.18	54.40
2000	**75.92**	75.86	75.64	75.59	63.75	64.76	64.67	64.76	50.09	75.88	55.21
3000	**76.04**	75.97	75.72	75.73	65.70	69.63	69.61	69.63	60.99	75.97	55.69
4000	**76.13**	76.06	75.68	75.71	69.31	73.67	73.72	73.67	65.32	76.08	56.08
5000	76.08	76.10	75.67	75.70	71.52	74.73	74.73	74.73	69.75	**76.11**	56.44
6000	**76.12**	**76.12**	75.76	75.76	73.53	75.57	75.56	75.57	71.23	76.11	57.03
7000	76.06	76.10	75.82	75.84	74.57	76.02	76.01	76.02	72.74	**76.11**	58.18
8000	76.10	76.07	75.96	75.93	74.83	**76.11**	76.09	**76.11**	74.47	76.07	59.99
9000	76.09	76.09	76.05	76.03	75.33	**76.12**	76.11	**76.12**	75.46	76.08	63.94
Ave.	**75.83**	75.75	75.69	75.69	68.60	70.09	70.08	70.09	63.15	75.76	57.08
t-test											
SMR	X	0.78	0.47	0.48	**0.00**	**0.03**	**0.03**	**0.03**	**0.00**	0.79	**0.00**
t-test											
UMA	0.78	X	0.80	0.80	**0.00**	**0.03**	**0.03**	**0.03**	**0.00**	0.99	**0.00**
Time	2236″	12,567″	110,094″	109,453″	457″	543″	547″	587″	579″	18,947″	409″

since they all belong to health domain, namely DNA, Pregnancy, Prognosis, Risk Factors, and Tomography. It is important to note that there are no overlapping terms in these 10 terms.

5.2. Comparison of MBFS methods with existing well-known feature selection methods

In order to compare to new MBFS methods, extensive experiments are performed by using six different datasets, two different classifiers: "multinomial Naïve Bayes (MNB)" and "Support Vector Machines (SVM)", nine different existing feature selection methods and the different number of selected features ranging from 500 to 10,000. For evaluation of classifiers performance, we use 10-fold cross-validation method. A large number of experiments are performed by using the combination of different parameters.

As explained in Section 3, a meaning score is computed for each term in a class in order to apply supervised MBFS. These class-based meaning scores for each term can be combined into one score using three different methods: "Rank", "Max", and "Average". As it can be seen in Section 5.4, our results show that "Rank" method usually gives better results. Therefore we only report the results of "Rank" method, called supervised meaning rank (SMR) in this paper. For unsupervised MBFS, a meaning score is computed for each in a document. Again, a single score for entire collection can be obtained by using "Rank", "Max" and "Average" methods. Among these three methods, "Average" approach usually works better for unsupervised MBFS. Consequently, only experimental results computed by "Average" method, called unsupervised meaning average (UMA), are reported in this paper.

The results are organized and reported in tables. We evaluate the performance of MNB or SVM classifiers after applying different feature selection methods. We also report performance of classifiers on original dataset without applying any feature selection method which can be accepted as a baseline in order to compare the effect of feature selection methods. The feature selection methods EOR, MC_OR, CDM, MOR, and WOR are developed by considering the characteristics of Naïve Bayes (NB) algorithm. Therefore, these special methods are only used while evaluating the performance of MNB classifier and are not applied before SVM classifier. The IG and χ^2 are very commonly used feature selection methods in text classification and they are included in all our result tables along with the unsupervised and supervised term weighting methods of TF-IDF and TF-ICF. We especially include TF-IDF method to the set of feature selection methods for benchmarking since the meaning measure has similarities to TF-IDF method. As discussed in Section 2.1, the meaning measure approach has much better theoretical background than TF-IDF method. The TF-ICF is a supervised version of TF-IDF that uses class information. In some respects, the TF-ICF is similar to the SMR and TF-IDF is similar to the UMA. More information can be found about these in the related work section.

In Tables 6–17, we report the accuracy of the classifiers. We start by selecting the best 500 features and go up to 10,000 features as long as the total number of features for that dataset permits. Each row shows the number of the selected features and performance of the classifier for a feature selection method given in that column. We mark the best performance values on each row by making typeface of its values as bold.

We evaluated the performance of MNB and SVM classifiers after reducing the number of features at each step. Tables 6, 8, 10, 12, 14 and 16 show the performance of MNB classifier on "tweet65k", "ohscal", "news3", "la1s", "wap" and "1150Haber" datasets respectively. Tables 7, 9, 11, 13, 15 and 17 show the performance of SVM classifier on "tweet6k", "ohscal", "news3", "la1s", "wap" and "1150Haber" datasets respectively. Remember that SMR is the supervised MBFS and UMA is the unsupervised MBFS method proposed in this paper. In Tables 6–17, each cell gives the classification accuracy of a classifier on the related feature selection method listed in each column. On each row, the values marked bold typeface shows the best performance values for that row. The "Ave." row gives the average of the classification accuracies in each column. The

Table 7
Comparison of different feature selection methods using the accuracy of SVM on "tweet65k" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	76.33	75.74	77.06	**77.10**	75.69	53.83
1000	77.41	77.27	**77.73**	**77.73**	77.27	54.36
2000	**77.87**	77.80	77.64	77.66	77.81	55.13
3000	**77.82**	77.75	77.44	77.46	77.77	55.56
4000	**77.78**	77.65	77.31	77.41	77.62	55.90
5000	**77.60**	77.53	77.33	77.36	77.53	56.23
6000	77.48	**77.51**	77.24	77.22	77.49	56.68
7000	**77.51**	77.47	77.25	77.28	77.50	57.43
8000	**77.50**	77.41	77.25	77.24	77.41	58.81
9000	77.44	77.37	77.27	77.29	**77.45**	62.71
Ave.	**77.47**	77.35	77.35	77.38	77.35	56.66
t-test SMR	X	0.60	0.43	0.52	0.62	**0.00**
t-test UMA	0.60	X	0.99	0.90	0.99	**0.00**
Time	16,315″	150,874″	317,338″	320,024″	212,829″	36,277″

Table 8
Comparison of different feature selection methods using the accuracy of MNB on "ohscal" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	74.31	71.92	74.29	74.65	20.20	**74.80**	74.56	74.71	20.52	71.94	56.75
1000	74.34	72.86	74.47	74.66	23.37	**75.66**	75.35	75.43	23.06	73.20	60.98
2000	74.43	73.47	74.74	74.96	27.87	**74.90**	74.68	74.63	27.84	74.01	64.85
3000	74.63	74.05	**75.06**	74.92	32.05	74.74	74.68	74.72	31.79	74.05	66.12
4000	74.62	74.30	74.74	74.76	35.76	**74.78**	74.77	74.66	35.35	74.38	66.43
5000	74.68	74.40	**74.85**	74.78	38.66	74.78	74.44	74.42	38.89	74.35	67.23
6000	74.70	74.58	74.72	74.76	42.64	**74.79**	74.74	74.74	42.46	74.65	67.57
7000	74.66	74.67	74.51	74.51	46.69	**74.74**	**74.74**	**74.74**	47.01	74.79	68.68
8000	74.81	74.71	74.74	74.74	54.09	**74.84**	74.76	74.76	54.29	74.78	69.29
9000	**74.96**	74.75	74.63	74.63	58.81	74.94	74.93	74.92	58.95	74.82	69.79
10,000	74.86	74.82	74.77	74.77	62.96	74.91	**74.88**	**74.88**	62.88	74.81	70.11
Ave.	74.64	74.05	74.68	74.74	40.28	**74.90**	74.78	74.78	42.00	74.16	66.16
t-test SMR	X	0.05	0.60	0.17	**0.00**	**0.02**	0.15	0.15	**0.00**	0.10	**0.00**
t-test UMA	0.05	X	**0.04**	**0.02**	**0.00**	**0.01**	**0.02**	**0.02**	**0.00**	0.77	**0.00**
Time	271″	806″	29,415″	29,286″	175″	181″	182″	183″	175″	1631″	178″

"t-test" rows give two-tailed distribution paired t-test results of SMR and UMR methods compared with other feature selection methods. The "Time" row gives the sum of the total execution times (in seconds) of experiments in each column by applying both the feature selection method and classifier.

The performance of MNB and SVM classifiers on the dataset "tweek65k" is given in Tables 6 and 7 respectively. Table 6 shows that the highest MNB classification accuracy obtained with SMR is 76.13 with 4000 attribute. In general, MNB classification accuracy after applying SMR and UMA feature selection methods are higher than EOR, MC_OR, CDM, MOR, WOR and TF-ICF and it is statistically significant according to t-test in most cases. The "t-test" results less than 0.05 is accepted as statically significant. If we do not apply feature selection (FS) method to the original data set, then the classification accuracy of SVM without FS is 76.09 which can be accepted as a baseline to see the effects of FS methods. From Table 6, we can see that the classification accuracies are very close to our baseline. In some cases, we obtain better classification accuracies than our baseline although we lose some of the information from original data set by applying FS methods.

Table 7 shows that the highest SVM classification accuracy obtained with SMR is 77.87 with 2000 attributes. SVM classification accuracy with SMR and UMA are always higher than TF-ICF and their results are statistically significant according to t-test. Although significance test results are not well enough, the performance of UMA and SMR is very close to IG, χ^2 and TF-IDF results. Although it not statistically significant, we can clearly see that SMR increases SVM classification accuracy in most cases. The classification accuracy of SVM without FS is 77.33 which is very close to the classification accuracies obtained after applying SMR and UMA methods.

Tables 8 and 9 show the performance of MNB and SVM classifiers on "ohscal" dataset. The MNB classifier has the best performance with MC_OR method which is a feature selection method specially developed for Naïve Bayes classifiers as seen in Table 8. The classification accuracy of MNB with SMR and UMA is higher than EOR, WOR and TF-ICF and it is statistically significant according to t-test.

Table 9
Comparison of different feature selection methods using the accuracy of SVM on "ohscal" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	**77.80**	74.28	77.20	77.38	74.82	50.61
1000	**77.71**	75.44	77.53	77.58	75.86	55.69
2000	76.76	75.22	77.60	**77.84**	75.95	61.21
3000	76.21	75.09	77.03	**77.07**	75.33	63.19
4000	75.74	75.11	76.17	**76.20**	75.26	63.47
5000	75.77	75.00	**76.21**	75.85	75.49	62.92
6000	**75.78**	75.04	75.30	75.24	75.27	62.91
7000	75.36	74.83	75.11	75.11	**75.40**	64.17
8000	**75.23**	74.75	74.55	74.54	75.07	64.16
9000	74.67	74.76	74.57	74.57	**75.09**	65.00
10,000	74.60	74.84	74.25	74.28	**75.02**	65.13
Ave.	**75.97**	74.94	75.96	**75.97**	75.32	61.68
t-test SMR	X	**0.01**	0.98	1.00	0.07	**0.00**
t-test UMA	**0.01**	X	**0.02**	**0.02**	**0.01**	**0.00**
Time	1624″	6920″	28,822″	28,751″	7178″	4599″

Table 10
Comparison of different feature selection methods using the accuracy of MNB on "new3s" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	76.72	66.72	75.62	**77.35**	12.58	68.35	70.01	70.11	13.61	65.82	67.28
1000	76.73	70.74	76.26	**77.99**	15.93	73.98	74.58	74.60	16.97	71.91	70.95
2000	76.96	75.64	76.96	**77.43**	17.64	75.96	76.47	76.19	19.20	74.49	73.95
3000	77.13	76.14	**77.54**	77.47	21.59	77.02	77.03	76.96	21.49	75.79	74.96
4000	77.37	76.65	77.88	**77.89**	23.02	77.35	77.07	76.96	23.82	76.11	76.21
5000	77.57	76.70	77.89	**77.90**	26.61	77.08	76.94	76.94	26.79	76.76	76.68
6000	77.84	77.01	78.10	**78.13**	28.48	77.25	77.18	77.19	29.03	77.12	76.97
7000	78.09	77.27	78.27	**78.31**	30.97	77.45	77.51	77.51	31.60	77.38	77.19
8000	78.16	77.68	78.52	**78.54**	33.64	77.80	77.68	77.68	33.80	77.62	77.52
9000	78.29	77.83	78.65	**78.66**	36.17	78.00	77.90	77.90	35.99	77.89	77.64
10,000	78.36	78.20	78.38	**78.46**	38.55	78.26	78.15	78.15	38.27	78.12	77.77
Ave.	77.57	75.51	77.64	**78.01**	25.93	76.23	76.41	76.38	26.42	75.36	75.19
t-test SMR	X	0.07	0.83	0.07	**0.00**	0.15	0.13	0.11	**0.00**	0.06	**0.03**
t-test UMA	0.07	X	0.07	**0.03**	**0.00**	0.61	0.49	0.50	**0.00**	0.93	0.83
Time	1333″	7370″	61,332″	61,365″	237″	326″	332″	370″	239″	4671″	523″

Table 11
Comparison of different feature selection methods using the accuracy of SVM on "new3s" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	57.03	35.38	57.40	**59.76**	26.79	45.74
1000	62.68	49.55	65.41	**66.28**	50.03	53.33
2000	68.92	66.46	71.44	**71.54**	62.12	60.11
3000	**72.58**	71.03	72.17	72.35	70.69	63.22
4000	**73.78**	72.68	71.98	71.95	72.82	65.24
5000	**74.48**	73.59	71.71	71.63	74.24	67.15
6000	**74.79**	74.07	71.48	71.45	74.56	68.17
7000	74.83	74.18	71.59	71.56	**74.84**	68.78
8000	74.80	74.57	71.46	71.52	**75.07**	69.20
9000	74.86	74.33	71.91	71.91	**74.89**	69.53
10,000	74.52	73.76	72.27	72.24	**75.10**	70.11
Ave.	**71.21**	67.24	69.89	70.20	66.47	63.69
t-test SMR	X	0.36	0.57	0.64	0.35	**0.02**
t-test UMA	0.36	X	0.53	0.47	0.90	0.44
Time	7520″	13,980″	46,856″	45,605″	11,703″	3574″

Table 9 shows that the highest SVM classification accuracy obtained with SMR is 77.80 with 500 attribute. The performance of SVM with SMR approximately is the same as IG and χ^2 according to t-test and higher than TF-ICF. If we compare the classification accuracy of SVM without FS, which is 74.28, with SMR and UMA methods, we can clearly see that SMR and UMA increase the SVM classification accuracy.

Table 12
Comparison of different feature selection methods using the accuracy of MNB on "la1s" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	83.86	82.08	**84.93**	84.68	43.70	78.25	78.18	78.18	37.89	83.21	70.54
1000	85.80	84.55	85.92	**86.20**	45.26	81.84	82.9	82.93	40.70	84.83	73.85
2000	86.77	85.74	**87.11**	87.08	47.53	85.30	85.99	85.92	46.88	86.17	75.94
3000	87.30	86.52	87.36	**87.55**	52.93	86.55	86.67	86.67	50.53	86.89	77.06
4000	87.64	86.95	87.30	87.42	55.46	**88.08**	87.23	87.23	53.84	87.27	78.87
5000	**88.05**	87.23	87.67	87.58	58.65	87.98	87.55	87.58	56.46	87.58	80.03
6000	88.05	87.48	87.70	87.70	62.02	87.95	88.02	**88.14**	61.27	87.80	81.93
7000	88.08	87.73	87.61	87.58	65.45	**88.11**	88.05	88.05	64.23	88.05	82.83
8000	88.08	87.98	87.42	87.42	69.82	88.14	**88.17**	**88.17**	67.38	88.23	83.27
9000	88.14	**88.20**	87.58	87.58	73.66	**88.20**	88.08	88.05	71.75	88.11	83.61
10,000	88.11	88.23	87.92	87.92	78.81	**88.30**	88.26	88.26	77.12	88.17	84.39
Ave.	**87.26**	86.61	87.14	87.16	59.39	86.25	86.28	86.29	57.10	86.94	79.30
t-test SMR	X	0.36	0.80	0.83	**0.00**	0.36	0.35	0.35	**0.00**	0.61	**0.00**
t-test UMA	0.36	X	0.41	0.40	**0.00**	0.75	0.77	0.77	**0.00**	0.66	**0.00**
Time	119″	273″	3825″	3787″	67″	68″	69″	69″	66″	527″	126″

Table 13
Comparison of different feature selection methods using the accuracy of SVM on "la1s" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	79.74	82.12	**83.71**	83.55	82.24	57.52
1000	84.02	83.68	**85.17**	84.86	83.99	61.77
2000	85.46	83.80	86.02	**86.17**	84.80	63.14
3000	84.74	84.64	85.55	**85.61**	84.64	64.73
4000	85.08	84.21	**85.14**	84.93	84.46	67.38
5000	**85.05**	84.33	84.24	84.24	84.74	68.95
6000	**84.71**	84.21	84.08	84.43	84.33	69.13
7000	**84.58**	84.43	84.02	83.99	84.27	71.16
8000	**84.30**	84.14	83.52	83.52	84.24	71.47
9000	**84.77**	84.33	83.77	83.77	84.11	72.35
10,000	84.24	84.43	83.55	83.55	**84.64**	73.22
Ave.	84.24	84.03	**84.43**	84.42	84.22	67.35
t-test SMR	X	0.68	0.73	0.75	0.97	**0.00**
t-test UMA	0.68	X	0.24	0.26	0.52	**0.00**
Time	477″	802″	4061″	4235″	1185″	573″

The experimental results of MNB and SVM classifiers on "new3s" dataset are in given Tables 10 and 11. Table 10 shows that χ^2 has the best performance but it is not statistically significant according to t-test with SMR. The classification accuracy of MNB with SMR is higher than EOR, WOR and TF-ICF and it is statistically significant according to t-test.

Table 11 shows that SVM classification accuracy with SMR is higher than TF-ICF and it is statistically significant according to t-test. In general, SVM classification accuracy with SMR is higher than other FS (feature selection) methods.

Tables 12 and 13 show the performance of MNB and SVM classifiers on "la1s" dataset. Table 12 indicates that MNB classification accuracy with SMR is usually higher than other FS methods. MNB classification accuracy with SMR and UMA are higher than EOR, WOR and TF-ICF and it is statistically significant according to t-test.

In Table 13, we can see that SVM classification accuracy with SMR and UMA are higher than TF-ICF and it is statistically significant according to t-test. IG has the best performance but it is not statistically significant according to t-test.

The performance of MNB and SVM classifiers on the dataset "wap" is given in Tables 14 and 15 respectively. Table 14 shows that the highest MNB classification accuracy obtained with SMR is 84.17 with 3000 attribute. The performance MNB with SMR approximately is the same as UMA. MNB classification accuracy with SMR and UMA are higher than EOR and WOR and it is statistically significant according to t-test. In general, MNB classification accuracy with UMA is higher than other FS methods. If we compare the classification accuracy of SVM without FS, which is 81.09, with SMR and UMA methods, we can clearly see that SMR and UMA increase the SVM classification accuracy.

Table 15 indicates that SMR outperforms all other feature selection methods and it is statistically significant according to t-test. The highest SVM classification accuracy obtained with SMR is 83.53 with 1000 attribute. If we compare the classification accuracy of SVM without FS, which is 82.18, with SMR method, we can clearly see that SMR increases the SVM classification accuracy.

Tables 16 and 17 give an evaluation of performances of MNB and SVM classifiers on "1150Haber" dataset. Table 16 shows that IG has the best performance but it is not statistically significant according to t-test. MNB classification accuracy with SMR is higher than EOR, WOR and TF-ICF and it is statistically significant according to t-test.

Table 14
Comparison of different feature selection methods using the accuracy of MNB on "wap" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	**82.82**	81.28	81.47	81.54	40.06	80.64	79.42	79.36	40.32	80.83	77.05
1000	82.37	**83.08**	82.69	82.44	46.47	81.86	82.88	82.82	46.60	81.67	79.29
2000	83.53	83.33	83.33	83.33	53.97	83.72	**83.78**	83.72	54.55	83.08	82.05
3000	**84.17**	83.91	82.50	82.50	58.78	83.78	83.91	83.91	58.72	82.95	82.37
4000	83.27	83.59	82.50	82.50	64.62	**83.72**	83.33	83.33	63.53	82.82	82.44
5000	82.44	**83.27**	82.37	82.37	69.29	83.14	82.95	82.95	66.79	82.37	82.18
6000	82.24	82.44	82.24	82.24	71.35	82.31	82.31	82.31	69.94	**82.56**	82.44
7000	81.67	82.12	82.12	82.12	74.68	82.24	82.05	82.12	71.73	**82.31**	82.05
8000	81.35	81.09	**81.99**	**81.99**	77.88	81.35	81.03	81.03	75.96	81.86	81.67
Ave.	82.65	**82.68**	82.36	82.34	61.90	82.53	82.41	82.39	60.90	82.27	81.28
t-test SMR	X	0.95	0.40	0.37	**0.00**	0.80	0.67	0.66	**0.00**	0.34	0.06
t-test UMA	0.95	X	0.41	0.37	**0.00**	0.77	0.65	0.63	**0.00**	0.34	0.07
Time	55″	82″	831″	825″	25″	25″	25″	26″	25″	148″	46″

Table 15
Comparison of different feature selection methods using the accuracy of SVM on "wap" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	**83.01**	81.09	81.99	81.86	81.41	76.47
1000	**83.53**	81.41	81.35	81.22	80.77	78.59
2000	**82.63**	81.60	81.28	81.28	81.60	79.68
3000	**82.69**	81.28	80.96	80.96	82.18	80.51
4000	**82.24**	81.73	81.22	81.22	81.60	80.32
5000	82.44	**82.63**	81.86	81.86	82.05	80.71
6000	82.24	82.05	82.31	82.37	**82.44**	80.51
7000	**82.76**	82.05	82.31	82.31	82.44	80.96
8000	**82.44**	82.31	82.37	82.37	81.79	81.09
Ave.	**82.66**	81.79	81.74	81.72	81.81	79.87
t-test SMR	X	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
t-test UMA	**0.00**	X	0.83	0.76	0.95	**0.00**
Time	328″	315″	1030″	1039″	514″	223″

Table 16
Comparison of different feature selection methods using the accuracy of MNB on "1150haber" dataset.

|F|	SMR	UMA	IG	χ^2	EOR	MC_OR	CDM	MOR	WOR	TF-IDF	TF-ICF
500	**93.39**	91.57	92.87	92.78	62.17	92.70	92.61	92.70	62.35	92.52	86.70
1000	93.57	93.30	**93.91**	93.65	71.65	93.83	93.74	93.83	71.57	93.39	89.22
2000	93.13	93.74	93.48	93.65	79.57	**93.83**	**93.83**	**93.83**	79.57	93.39	90.70
3000	93.30	93.57	**94.00**	93.91	83.39	93.83	93.74	93.83	83.39	93.30	91.30
4000	93.74	93.65	**94.00**	**94.00**	87.13	93.83	93.83	93.83	87.13	93.48	92.52
5000	**93.91**	93.83	**93.91**	**93.91**	89.57	**93.91**	**93.91**	**93.91**	89.57	93.65	**93.91**
6000	94.17	94.00	**94.26**	**94.26**	91.91	94.09	94.09	94.09	91.91	93.91	**94.26**
Ave.	93.60	93.38	**93.78**	93.74	80.77	93.72	93.68	93.72	80.78	93.38	91.23
t-test SMR	X	0.53	0.45	0.56	**0.01**	0.61	0.74	0.61	**0.01**	0.31	**0.04**
t-test UMA	0.53	X	0.29	0.34	**0.01**	0.36	0.43	0.36	**0.01**	0.99	0.06
Time	15″	29″	365″	366″	8″	9″	9″	9″	8″	56″	9″

Table 17
Comparison of different feature selection methods using the accuracy of SVM on "1150haber" dataset.

|F|	SMR	UMA	IG	χ^2	TF-IDF	TF-ICF
500	**90.35**	85.91	89.22	90.09	87.22	80.70
1000	90.35	87.91	**90.78**	90.00	87.57	80.87
2000	90.35	89.22	90.17	**90.78**	89.83	81.04
3000	**91.13**	89.22	89.74	89.65	90.43	81.57
4000	**90.26**	89.30	88.96	88.96	89.57	84.26
5000	**90.26**	89.39	90.00	90.00	**90.26**	86.78
6000	90.26	89.91	90.35	**90.35**	90.17	88.52
Ave.	**90.42**	88.69	89.89	89.98	89.29	83.39
t-test SMR	X	**0.01**	0.07	0.09	0.05	**0.00**
t-test UMA	**0.01**	X	0.06	**0.04**	0.42	**0.00**
Time	59″	75″	405″	433″	107″	44″

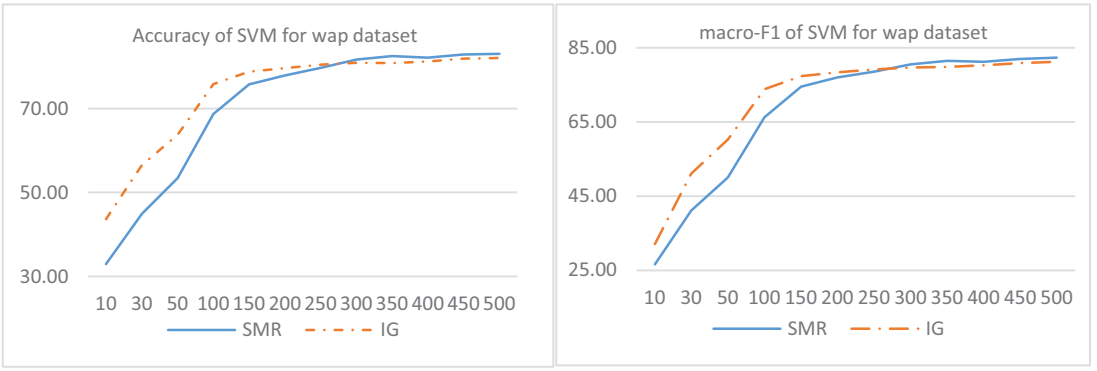

Fig. 3. Accuracy and macro-F1 measure of SMR and IG feature selection methods for "wap" dataset with SVM classifiers. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

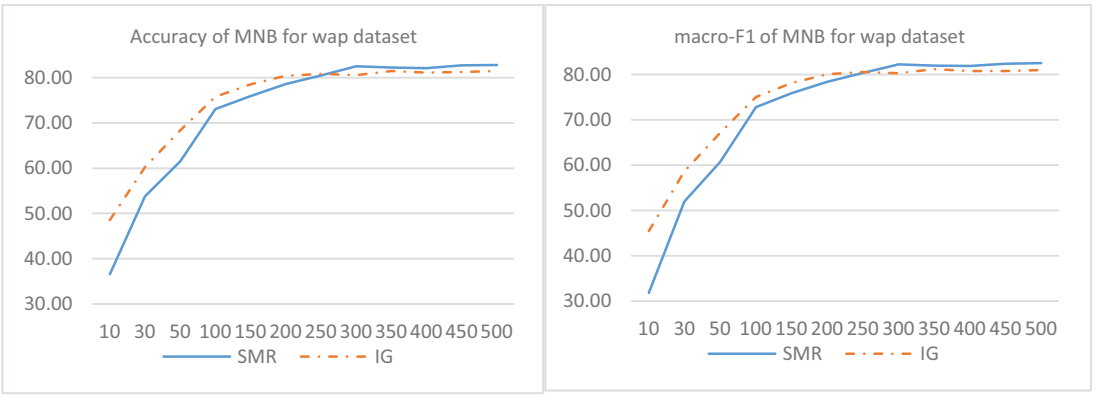

Fig. 4. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "wap" dataset with MNB classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

Table 17 shows that the highest SVM classification accuracy obtained with SMR is 91.13 with 3000 attribute and it shows the best classification performance in most cases. If we compare the classification accuracy of SVM without FS, which is 90.17, with SMR method, we can clearly see that SMR increases the SVM classification accuracy.

5.3. Comparison of SMR method with IG on smaller feature sizes

In previous section on Tables 6–17, we analyzed the performance of new meaning based SMR and UMA methods by comparing them with several well-known feature selection methods. At each experiment, we selected a large number of features in number ranging from 500 to 10,000 depending on the dataset and compared performance results. It is also important to analyze the performance of the new meaning based method with a smaller feature size. We reduce the number of features into 10, 30, 50, 100, 150, 200, 250, 300, 350, 400, 450 and 500 at each step with new proposed SMR and popular IG methods. We selected IG to compare with SMR because it is well-known and commonly used feature selection method in order to reduce high dimensionality of dataset (Altınel, Ganiz, & Diri, 2013, 2014a, 2014b; Ganiz et al., 2011; Poyraz, Kilimci, & Ganiz, 2012, 2014; Ganiz, Lytkin, & Pottenger, 2009)

In Figs. 3–10, the performance of MNB and SVM classifiers is evaluated by using accuracy and Macro-F_1 measures. In all Figs. 3–10, in left hand side graph, x-axis represents number of features and y-axis represents accuracy of classifiers and in right hand side graph, x-axis represents number of features and y-axis represents Macro-F_1 of classifiers. Macro-F_1 is a measure often used for performance evaluation of multiclass classifiers. Macro-F_1 measure is computed for each different class first and then the average for all classes is taken. This measure gives equal weight to each class regardless of size. The classes with the less number of features have the same influence as with large number feature on the classifier's performance.

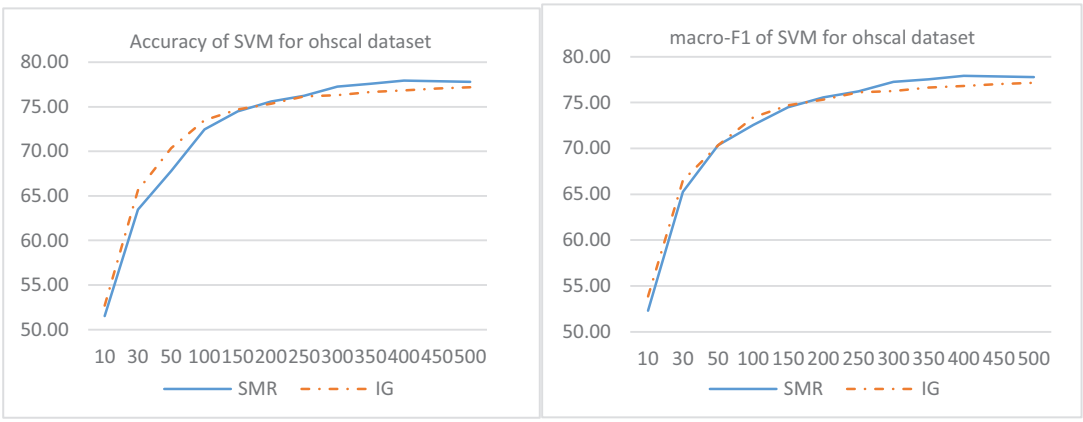

Fig. 5. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "ohscal" dataset with SVM classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

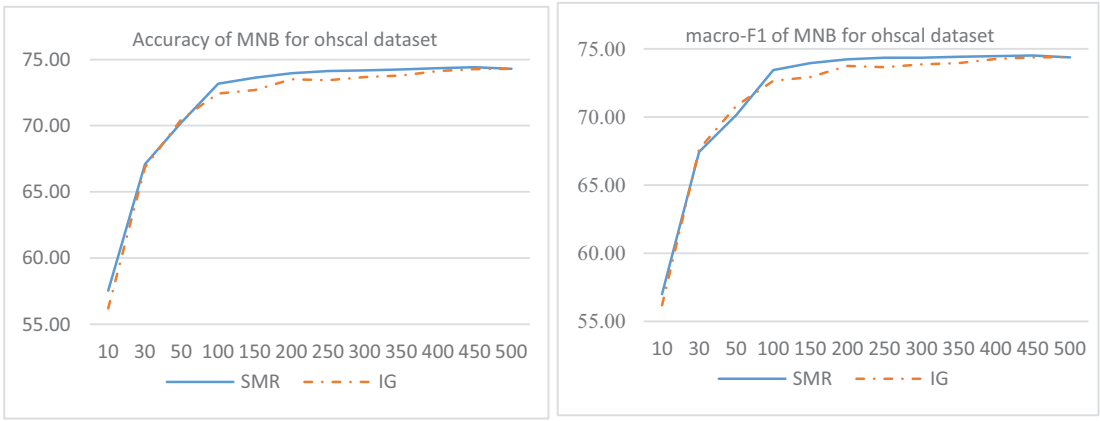

Fig. 6. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "ohscal" dataset with MNB classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

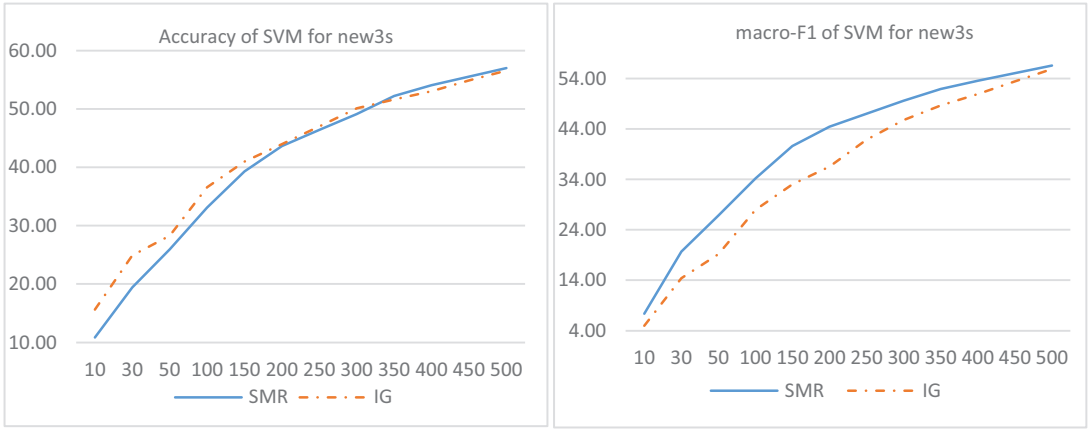

Fig. 7. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "new3s" dataset with SVM classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

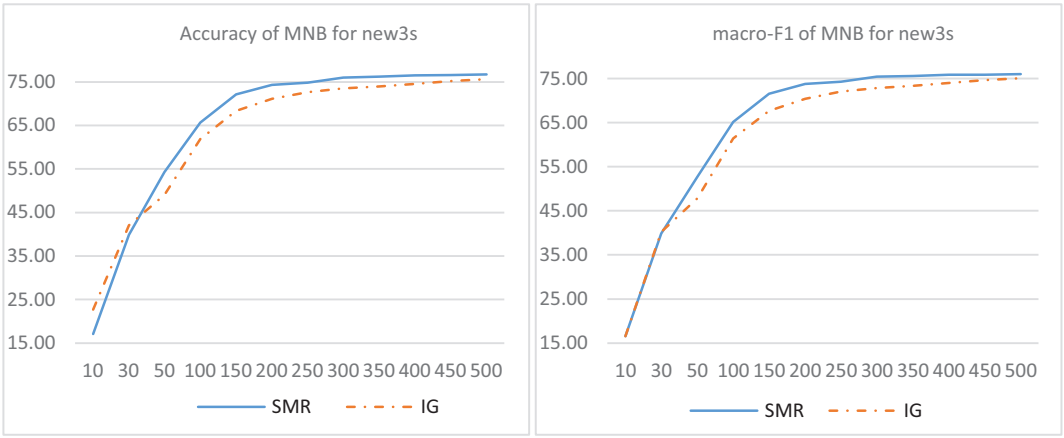

Fig. 8. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "new3s" dataset with MNB classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

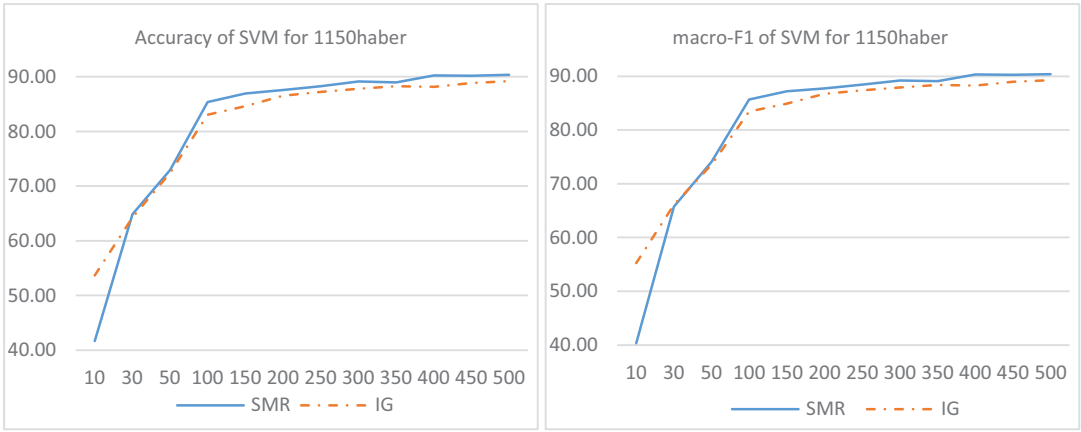

Fig. 9. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "1150haber" dataset with SVM classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

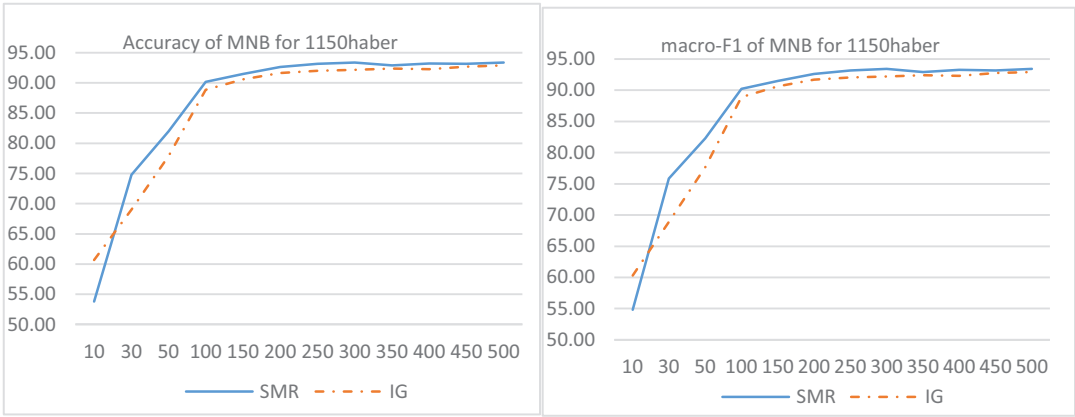

Fig. 10. Accuracy and macro-F1 measures of SMR and IG feature selection methods for "1150haber" dataset with MNB classifier. x axis shows the number of features, y axis shows the performance in terms of accuracy or macro-F1.

Table 18
Comparison of different MBFS methods using the accuracy of MNB on "tweet65k" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**74.45**	74.40	72.92	54.76	54.33	73.97
1000	75.31	**75.38**	74.92	55.94	57.40	75.18
2000	75.92	**75.96**	75.67	60.85	70.03	75.86
3000	**76.04**	**76.04**	75.82	66.44	74.92	75.97
4000	**76.13**	76.12	75.98	69.64	75.13	76.06
5000	76.08	**76.11**	75.97	71.90	75.13	76.10
6000	**76.12**	76.04	75.96	73.00	75.22	**76.12**
7000	76.06	**76.11**	76.10	73.70	75.24	76.10
8000	76.10	**76.11**	76.07	74.85	75.26	76.07
9000	76.09	**76.14**	76.07	75.51	75.32	76.09
Ave.	75.83	**75.84**	75.55	67.66	70.80	75.75
Time	2236″	772″	740″	19,456″	13,967″	12,567″

Table 19
Comparison of different MBFS methods using the accuracy of SVM on "tweet65k" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	76.33	76.35	74.69	54.35	54.12	75.74
1000	77.41	77.46	76.79	55.09	56.33	77.27
2000	77.87	77.77	77.59	59.94	70.98	77.80
3000	77.82	77.86	77.65	66.64	76.81	77.75
4000	77.78	77.62	77.49	70.24	76.74	77.65
5000	77.60	77.58	77.47	72.87	76.91	77.53
6000	77.48	77.41	77.37	74.12	76.77	77.51
7000	77.51	77.50	77.45	74.63	76.59	77.47
8000	77.50	77.40	77.46	75.86	76.59	77.41
9000	77.44	77.35	77.45	76.71	76.56	77.37
Ave.	77.47	77.43	77.14	68.04	71.84	77.35
Time	265,315″	213,363″	222,474″	183,248″	126,815″	150,874″

Fig. 3 shows plots of accuracy and Macro-F_1 measures respectively by using SVM classifier with SMR and IG feature selection methods on "wap" dataset. Fig. 4 gives the same performance results for MNB classifier. As we can see from the figures., the new SMR method gives better performance with small number of features. The performance of MNB and SVM classifiers for the "ohscal" dataset is given in Figs. 5 and 6. Figs. 7 and 8 show the performance for "news3" dataset. The performance results for "1150Haber" dataset are given in Figs. 9 and 10. We can observe that the new proposed SMR method usually has a similar or better performance than IG method in many cases.

5.4. Comparison of new meaning based feature selection methods

In Tables 18–29, we analyzed the performance of all MBFS methods; supervised approaches, SMR, SMM and SMA and unsupervised approaches UMA, UMM and UMR methods by comparing them each other. At each experiment, we selected a large number of features in number ranging from 500 to 10,000 depending on the dataset and compared performance results. For evaluation of classifiers performance, we use 10-fold cross-validation method. These tables show that generally supervised MBFS methods especially SMR outperforms unsupervised MBFS methods.

Table 30 gives a summary of averages of the classification accuracy of classifiers MNB and SVM on different data sets after applying supervised SMR, SMM, SMA, and unsupervised UMR, UMM and UMA methods listed in the last second row of Tables 18–29. The average values in each cell are found by taking average of classification accuracy of classifiers with feature sizes ranging from 500 up to 10,000 obtained with feature selection methods. The column |F| gives the number of features in each data set.

The "NO-FS" column gives the classification accuracy of MNB and SVM classifiers on whole dataset without applying any feature selection methods. This can be considered as a baseline to compare the effects of feature selection methods. By reducing feature sizes, the feature selection methods loses some of information on the original dataset. Although we lose some information by application of feature selection methods, the average classification accuracies are very close to NO-FS values. For example, if compare the average classification accuracy of SMR method with average NO-FS value, they have are very close values (82.29 and 81.93). On average, we only see 0.43% decrease on the classification accuracy of MNB after

Table 20
Comparison of different MBFS methods using the accuracy of MNB on "ohscal" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**74.31**	74.23	60.75	39.84	20.01	71.92
1000	**74.34**	74.23	68.08	53.08	25.88	72.86
2000	**74.43**	74.40	71.23	61.03	36.11	73.47
3000	**74.63**	74.41	72.25	66.96	44.77	74.05
4000	**74.62**	74.61	73.16	70.53	53.16	74.30
5000	**74.68**	74.52	73.60	72.18	59.68	74.40
6000	**74.70**	74.63	73.98	73.15	64.48	74.58
7000	74.66	**74.82**	74.29	73.74	67.41	74.67
8000	74.81	**74.87**	74.36	74.78	69.56	74.71
9000	**74.96**	74.84	74.47	74.92	73.83	74.75
10,000	74.86	74.89	74.85	**74.98**	74.91	74.82
Ave.	**74.64**	74.59	71.91	66.84	53.62	74.05
Time	271″	215″	190″	34,514″	2137″	806″

Table 21
Comparison of different MBFS methods using the accuracy of SVM on "ohscal" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**77.80**	77.58	63.69	32.72	15.59	74.28
1000	**77.71**	77.49	70.53	50.05	17.11	75.44
2000	**76.76**	76.63	73.06	60.15	22.76	75.22
3000	**76.21**	76.13	74.00	68.03	32.65	75.09
4000	75.74	**76.02**	74.05	72.61	44.09	75.11
5000	**75.77**	75.30	74.65	74.31	54.35	75.00
6000	**75.78**	75.54	74.57	75.00	61.84	75.04
7000	75.36	**75.52**	74.52	75.20	65.46	74.83
8000	75.23	75.43	74.65	**76.38**	68.80	74.75
9000	74.67	75.21	74.50	**76.19**	74.66	74.76
10,000	74.60	74.99	74.53	**75.34**	75.26	74.84
Ave.	75.97	**75.99**	72.98	66.91	48.42	74.94
Time	16,242″	4480″	5429″	43,744″	5342″	6920″

Table 22
Comparison of different MBFS methods using the accuracy of MNB on "new3s" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**76.72**	74.52	56.95	36.83	15.92	66.72
1000	**76.73**	76.05	65.05	51.89	24.10	70.74
2000	**76.96**	75.94	71.70	62.96	37.10	75.64
3000	**77.13**	76.37	74.61	67.32	48.54	76.14
4000	77.37	76.84	75.53	**77.77**	57.10	76.65
5000	77.57	77.06	76.05	**79.57**	63.65	76.70
6000	77.84	77.27	76.14	**80.00**	68.90	77.01
7000	78.09	77.50	76.48	**80.58**	72.17	77.27
8000	78.16	77.86	77.09	**80.82**	74.96	77.68
9000	78.29	78.04	77.52	**80.86**	76.71	77.83
10,000	78.36	78.23	77.60	**81.09**	78.26	78.20
Ave.	**77.57**	76.88	73.16	70.88	56.13	75.51
Time	1333″	253″	226″	103,097″	8384″	7370″

applying SMR method. In some cases, we can even get better classification accuracies after application of feature selection methods as it is seen in Tables 6–17.

From Table 30, we can observe that SMR (Supervised Meaning Rank) usually produces better classification accuracies among supervised feature selection methods (SMR, SMM and SMA). Among unsupervised feature selection methods (UMR,

Table 23
Comparison of different MBFS methods using the accuracy of SVM on "new3s" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**57.03**	53.43	26.99	19.29	09.06	35.38
1000	**62.68**	58.56	42.30	29.38	11.60	49.55
2000	**68.92**	67.30	54.51	40.60	16.88	66.46
3000	**72.58**	70.62	64.34	47.10	25.27	71.03
4000	**73.78**	73.72	68.35	62.48	34.57	72.68
5000	74.48	**74.59**	70.00	69.20	42.36	73.59
6000	74.79	**74.91**	70.82	71.36	49.17	74.07
7000	74.83	**75.30**	71.81	72.46	55.59	74.18
8000	74.80	**75.54**	72.29	73.14	61.70	74.57
9000	74.86	**75.48**	72.66	73.81	65.38	74.33
10,000	74.52	**75.30**	72.72	74.44	67.88	73.76
Ave.	**71.21**	70.43	62.44	57.57	39.95	67.24
Time	7520″	5441″	7189″	115,246″	4960″	13,980″

Table 24
Comparison of different MBFS methods using the accuracy of MNB on "la1s" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**83.86**	83.61	77.15	65.36	45.19	82.08
1000	85.80	**86.24**	80.65	75.87	55.59	84.55
2000	**86.77**	86.70	83.83	82.55	69.10	85.74
3000	**87.30**	**87.30**	85.24	85.92	76.47	86.52
4000	**87.64**	87.39	86.52	87.05	81.15	86.95
5000	**88.05**	87.55	86.83	87.48	84.68	87.23
6000	**88.05**	87.80	87.14	87.89	85.46	87.48
7000	**88.08**	**88.08**	87.27	87.83	86.27	87.73
8000	88.08	**88.36**	87.58	87.89	86.77	87.98
9000	88.14	**88.20**	87.67	88.08	87.58	**88.20**
10,000	88.11	**88.23**	88.02	88.02	87.92	**88.23**
Ave.	**87.26**	87.22	85.26	83.99	76.93	86.61
Time	119″	68″	69″	5194″	267″	273″

Table 25
Comparison of different MBFS methods using the accuracy of SVM on "la1s" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	79.74	76.37	75.97	54.43	33.77	**82.12**
1000	**84.02**	81.99	79.37	65.61	40.70	83.68
2000	**85.46**	84.71	82.52	75.84	52.15	83.80
3000	**84.74**	84.49	83.55	80.34	60.58	84.64
4000	**85.08**	84.68	83.90	82.62	69.04	84.21
5000	**85.05**	84.49	83.74	84.96	75.72	84.33
6000	84.71	84.58	83.68	**85.42**	79.74	84.21
7000	84.58	84.36	84.05	**85.33**	81.87	84.43
8000	84.30	84.21	83.71	**84.64**	83.99	84.14
9000	84.77	84.24	83.52	**85.27**	85.27	84.33
10,000	84.24	84.27	84.14	**85.17**	83.99	84.43
Ave.	**84.24**	83.49	82.56	79.06	67.89	84.03
Time	477″	475″	591″	7020″	637″	802″

Table 26
Comparison of different MBFS methods using the accuracy of MNB on "wap" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**82.82**	79.36	80.00	60.90	46.28	81.28
1000	82.37	82.05	82.37	70.06	61.03	**83.08**
2000	**83.53**	83.08	83.14	76.22	73.46	83.33
3000	**84.17**	83.78	83.59	79.29	76.73	83.91
4000	83.27	83.46	**83.78**	81.73	80.19	83.59
5000	82.44	83.21	**83.53**	82.31	81.67	83.27
6000	82.24	**82.56**	82.31	82.50	82.24	82.44
7000	81.67	81.92	81.99	**82.18**	81.73	82.12
8000	**81.35**	81.03	81.22	**81.35**	80.83	81.09
Ave.	82.65	82.27	82.44	77.39	73.80	**82.68**
Time	55″	25″	25″	971″	69″	82″

Table 27
Comparison of different MBFS methods using the accuracy of SVM on "wap" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	**83.01**	79.55	78.85	52.31	31.41	81.09
1000	**83.53**	81.86	79.74	65.13	50.90	81.41
2000	**82.63**	81.67	81.35	72.88	66.54	81.60
3000	**82.69**	81.86	80.77	77.44	72.69	81.28
4000	82.24	**82.88**	81.67	80.64	77.76	81.73
5000	82.44	81.35	81.92	**82.76**	79.81	82.63
6000	82.24	82.56	82.05	**83.33**	82.76	82.05
7000	82.76	82.12	82.05	**83.53**	83.01	82.05
8000	82.44	**82.50**	82.18	82.31	82.37	82.31
Ave.	**82.66**	81.82	81.18	75.59	69.69	81.79
Time	328″	237″	253″	1158″	310″	315″

Table 28
Comparison of different MBFS methods using the accuracy of MNB on "1150haber" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	93.39	94.00	84.43	67.83	42.26	91.57
1000	93.57	93.13	88.87	83.22	68.00	93.30
2000	93.13	93.39	93.22	91.91	86.52	93.74
3000	93.30	93.30	93.57	93.22	92.70	93.57
4000	93.74	93.39	93.48	93.48	94.00	93.65
5000	93.91	93.57	93.65	93.83	93.74	93.83
6000	94.17	93.74	93.83	94.17	94.00	94.00
Ave.	93.60	93.50	91.58	88.24	81.60	93.38
Time	15″	10″	10″	415″	27″	29″

UMM and UMA), UMA (Supervised Meaning Rank) usually produces better classification accuracies. We can conclude that SMR and UMA can be preferred methods for supervised and unsupervised meaning based feature selection respectively. Some of feature selection methods can produce very different and poor performance results on different datasets as in EOR method in Tables 6 and 8. Another observation that can be concluded from experiments is that both SMR and UMA methods usually results in good and consistent performances irrespective of different datasets.

Table 29
Comparison of different MBFS methods using the accuracy of SVM on "1150haber" dataset.

|F|	SMR	SMM	SMA	UMR	UMM	UMA
500	90.35	88.78	77.65	59.30	31.57	85.91
1000	90.35	89.65	81.39	76.09	51.04	87.91
2000	90.35	89.57	86.87	87.91	79.74	89.22
3000	91.13	89.22	88.09	89.74	87.30	89.22
4000	90.26	89.57	89.22	90.17	90.87	89.30
5000	90.26	89.65	89.04	90.78	91.57	89.39
6000	90.26	89.39	90.26	90.09	90.78	89.91
Ave.	90.42	89.40	86.07	83.44	74.70	88.69
Time	59″	50″	58″	536″	60″	75″

Table 30
Comparison of different MBFS methods.

Classifier	Dataset	|F|	NO-FS	SMR average	SMM average	SMA average	UMR average	UMM average	UMA average
MNB	tweet65k	9905	76.09	75.83	**75.84**	75.55	67.66	70.80	**75.75**
MNB	ohscal	11,466	74.94	**74.64**	74.59	71.91	66.84	53.62	**74.05**
MNB	new3s	26,833	79.24	**77.57**	76.88	73.16	70.88	56.13	**75.51**
MNB	la1s	13,169	88.23	**87.26**	87.22	85.26	83.99	76.93	**86.61**
MNB	wap	8461	81.09	**82.65**	82.27	82.44	77.39	73.80	**82.68**
MNB	1150haber	6656	94.17	**93.60**	93.50	91.58	88.24	81.60	**93.38**
MNB average	–	–	82.29	**81.93**	81.72	79.98	75.83	68.81	**81.33**
SVM	tweet65k	9905	77.33	**77.47**	77.43	77.14	68.04	71.84	**77.35**
SVM	ohscal	11,466	74.28	75.97	**75.99**	72.98	66.91	48.42	**74.94**
SVM	new3s	26,833	71.94	**71.21**	70.43	62.44	57.57	39.95	**67.24**
SVM	la1s	13,169	84.24	**84.24**	83.49	82.56	79.06	67.89	**84.03**
SVM	wap	8461	82.18	**82.66**	81.82	81.18	75.59	69.69	**81.79**
SVM	1150haber	6656	90.17	**90.42**	89.40	86.07	83.44	74.70	**88.69**
SVM average	–	–	80.02	**80.33**	79.76	77.06	71.77	62.08	**79.01**

6. Conclusion

In this study, we introduce and evaluate the performance of new feature selection methods, called meaning based feature selection (MBFS). The new methods use the meaning measure. It is based on the Helmholtz principle from the Gestalt theory of human perception. Helmholtz principle states that a structure can be perceived or meaningful in an image whenever some large deviation from randomness occurs. In the context of text mining, this means that meaningful (informative) features or interesting events appear as large deviations from randomness in a given text. By using Helmholtz principle, a meaning measure is defined to assign a level (score) of meaningfulness to important terms in a given text. The meaning measure is applied to rapid change detection, keyword extraction, and document summarization etc. previously.

We adapt and use the meaning measure as a new method for supervised and unsupervised feature selection in this paper. An extensive comparative study is carried out in order to assess the performance of new supervised and unsupervised meaning based feature selection methods (MBFS). We have defined three supervised MBFS methods called SMR, SMM and SMA, and three unsupervised MBFS methods called UMR, UMM, UMA. From experimental studies, we observe that SMR and UMA results in better classification accuracies.

SMR and UMA methods are compared with nine different and well-known feature selection methods on six different datasets. The performance of supervised SMR and unsupervised UMA methods is almost the same as the performance of many well-known feature selection algorithms and produces better results in many cases. The new methods have stable performance on different datasets and different feature spaces. They can be used as an effective feature selection method without hesitation on different datasets and feature spaces.

The execution times of new MBFS methods are compared with other methods. Experimental results show that MBFS methods have higher speed than the popular feature selection methods IG, TF-IDF and χ^2.

As a future work, we plan to use meaning score as metric for text classification. As our first attempt for text classification, a new classifier, called Supervised Meaning Classifier (SMC), is proposed in (Ganiz, Tutkan, & Akyokuş, 2015).

Previously, the new MBFS methods are introduced by us at Turkish symposiums: ASYU[1] and ELECO[2]. In ASYU we presented the supervised MBFS with "max" approach (Tutkan, Ganiz, & Akyokuş, 2014a,b). In ELECO, we presented the unsupervised MBFS method with "max" approach (Tutkan et al., 2014a,b).

[1] Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu (ASYU): Symposium of innovations on intelligent systems and applications.

[2] Elektrik – Elektronik, Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu (ELECO): Symposium of electric electronic computer and biomedical engineering.

Acknowledgements

This work is supported in part by The Scientific and Technological Research Council of Turkey (TÜBİTAK) grant number 111E239. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the TÜBİTAK.

References

Altınel, B., Ganiz, M. C., & Diri, B. (2013). A novel higher-order semantic kernel. In *Proceedings of the IEEE 10th international conference on electronics computer and computation (ICECCO)* (pp. 216–219).
Altınel, B., Ganiz, M. C., & Diri, B. (2014). A semantic kernel for text classification based on iterative higher–order relations between words and documents. In *Proceedings of the 13th international conference on artificial intelligence and soft computing (icaisc), lecture notes in artificial intelligence(LNAI)* (pp. 505–517).
Altınel, B., Ganiz, M. C., & Diri, B. (2014). A simple semantic kernel approach for svm using higher-order paths. In *Proceedings of IEEE International Symposium on INnovations in intelligent sysTems and applications (INISTA)* (pp. 431–435).
Altınel, B., Ganiz, M. C., & Diri, B. (2015). A corpus-based semantic kernel for text classification by using meaning values of terms. *Engineering Applications of Artificial Intelligence, 43*, 54–66.
Amasyalı, M. F., & Beken, A. (2009). Türkçe kelimelerin anlamsal benzerliklerinin ölçülmesi ve metin sınıflandırmada kullanılması. *IEEE signal processing and communications applications conference, SİU-2009*.
Baccianella, S., Esuli, A., & Sebastiani, F. (2013). Using micro-documents for feature selection: The case of ordinal text classification. *Expert Systems with Applications, 40*(11), 4687–4696.
Balinsky, A. A., Balinsky, H. Y., & Simske, S. J. (2010). On Helmholtz's principle for documents processing. In *Proceedings of the 10th ACM symposium on document engineering* (pp. 283–286). ACM.
Balinsky, A., Balinsky, H., & Simske, S. (2011). *On the Helmholtz principle for data mining*. Hewlett-Packard Development Company, LP, HP Laboratories, Technical Report HPL-2010-133, Palo Alto, CA.
Balinsky, A., Balinsky, H., & Simske, S. (2011). Rapid change detection and text mining. *The 2nd IMA conference on mathematics in defence, defence academy*.
Balinsky, H., Balinsky, A., & Simske, S. (2011). Document sentences as a small world. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on* (pp. 2583–2588). IEEE.
Balinsky, H., Balinsky, A., & Simske, S. J. (2011). Automatic text summarization and small-world networks. In *Proceedings of the 11th ACM symposium on Document engineering* (pp. 175–184). ACM.
Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*(3), 5432–5435.
Chen, Y. T., & Chen, M. C. (2011). Using chi-square statistics to measure similarities for text categorization. *Expert systems with applications, 38*(4), 3085–3090.
Dadachev, B., Balinsky, A., & Balinsky, H. (2014). On automatic text segmentation. In *Proceedings of the 2014 ACM symposium on Document engineering* (pp. 73–80). ACM.
Dadachev, B., Balinsky, A., Balinsky, H., & Simske, S. (2012). On the helmholtz principle for data mining. In *Emerging Security Technologies (EST), 2012 Third International Conference on* (pp. 99–102). IEEE.
Desolneux, A., Moisan, L., & Morel, J. M. (2007). *From gestalt theory to image analysis: A probabilistic approach*: Vol. 34. Springer, New York.
Ganiz, M. C., George, C., & Pottenger, W. M. (2011). Higher order Naive Bayes: A novel non-IID approach to text classification. *Knowledge and Data Engineering, IEEE Transactions on, 23*(7), 1022–1034.
Ganiz, M. C., Lytkin, N. I., & Pottenger, W. M. (2009). Leveraging Higher Order Dependencies Between Features For Text Classification. In *Proceedings of the Conference Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)* (pp. 375–390).
Ganiz, M. C., Tutkan, M., & Akyokuş, S. (2015). A Novel Classifier Based on Meaning for Text Classification. In *Innovations in Intelligent Systems and Applications (INISTA 2015), September* (pp. 2–4). Spain: Madrid.
Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1–12.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter, 11*(1), 10–18.
Han, E. H. S., & Karypis, G. (2000). *Centroid-based document classification: Analysis and experimental results* (pp. 424–431). Berlin Heidelberg: Springer. Online avalible of datasets: http://www.cs.waikato.ac.nz/ml/weka/datasets.html (19 multi-class (1-of-n) text datasets donated by George Forman/Hewlett-Packard Labs.
Han, E. H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., & Moore, J. (1998). Webace: A web agent for document categorization and exploration. In *Proceedings of the second international conference on Autonomous agents* (pp. 408–415). ACM.
Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. *Morgan kaufmann*. San Francisco, CA.
Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94* (pp. 192–201). London: Springer.
Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137–142). Berlin Heidelberg: Springer.
Joachims, T. (2001). A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 128–136). ACM.
Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11–21.
Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373–397.
Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (pp. 453–459). Association for Computational Linguistics.
Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management, 42*(1), 155–165.
Lertnattee, V., & Theeramunkong, T. (2004). Analysis of inverse class frequency in centroid-based text classification. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on: Vol. 2* (pp. 1171–1176). IEEE.
Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98* (pp. 4–15). Berlin Heidelberg: Springer.
Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. In FSDM (pp. 4–13).
McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization, Vol. 752*, 41–48.
Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. In *Proceedings of the 16th international conference on machine learning (ICML*.
Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods* (pp. 185–208). MA, USA: MIT press.
Poyraz, M., Kilimci, Z. H., & Ganiz, M. C. (2014). Higher-order smoothing: A novel semantic smoothing method for text classification. *Journal of Computer Science and Technology, 29*(3), 376–391.
Poyraz, M., Kilimci, Z. H., & Ganiz, M. C. (2012). A novel semantic smoothing method based on higher-order paths for text classification. In *IEEE international conference on data mining (ICDM)* (pp. 615–624).

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1992). Gamma Function, Beta Function, Factorials, Binomial Coefficients and Incomplete Beta Function, Student's Distribution, F-Distribution, Cumulative Binomial Distribution § 6.1 and 6.2 in Numerical Recipes in FORTRAN: The Art of Scientific Computing.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81–106.

Quinlan, J. R. (1993). C4. 5: Programs for machine learning. *Morgan kaufmann*. San Francisco, CA.

Rennie, J.D., Shih, L., Teevan, J., & Karger, D.R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616–623).

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of documentation, 60*(5), 503–520.

Schneider, K. M. (2004). A new feature selection score for multinomial naive Bayes text classification based on KL-divergence. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics* (p. 24).

Shang, C., Li, M., Feng, S., Jiang, Q., & Fan, J. (2013). Feature selection via maximizing global information gain for text classification. *Knowledge-Based Systems, 54*, 298–309.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications, 33*(1), 1–5.

Shannon, C.E., & Weaver, W. (1949). The mathematical theory of communication (Urbana, IL.)

Tasci, S., & Gungor, T. (2009). LDA-based Keyword Selection in Text Categorization. In *24th International symposium on computer and information sciences, ISCIS 2009* (pp. 230–235). 2009.

Tutkan, M., Ganiz, M.C., & Akyokuş, S. (2014a). (symposium paper in Turkish). Metin Sınıflandırma için Yeni Bir Eğitimli Anlamsal Özellik Seçimi Yöntemi. ASYU 2014 (Akıllı Sistemlerde Yenilikler ve Uygulamalar Sempozyumu) , Ekim 9-10 2014, İzmir Katip Çelebi Üniversitesi, İzmir, Türkiye.

Tutkan, M., Ganiz, M.C., & Akyokuş, S.(2014b). (symposium paper in Turkish). Metin Sınıflandırma için Yeni Bir Eğitimsiz Anlamsal Özellik Seçimi Yöntemi. ELECO 2014 (Elektrik - Elektronik, Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu), 27–29 Kasım 2014, Bursa, Türkiye.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226–235.

Yang, J., Liu, Y., Liu, Z., Zhu, X., & Zhang, X. (2011). A new feature selection algorithm based on binomial hypothesis testing for spam filtering. *Knowledge-Based Systems, 24*(6), 904–914.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management, 48*(4), 741–754.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML: Vol. 97* (pp. 412–420).

Zhou, Q., Zhao, M., & Hu, M. (2004). Study on feature selection in Chinese text categorization. *Journal of Chinese Information Processing, 18*(3), 17–23.

Zhou, X., Hu, Y., & Guo, L. (2014). Text categorization based on clustering feature selection. *Procedia Computer Science, 31*, 398–405.