Expert Systems With Applications 92 (2018) 106–123

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature selection in multiword expression recognition

Senem Kumova Metin

Izmir University of Economics, Faculty of Engineering, No.156, 35330, Balçova-İzmir, Turkey

ARTICLE INFO

Article history:
Received 22 December 2016
Revised 15 September 2017
Accepted 19 September 2017
Available online 20 September 2017

Keywords:
Multiword expression
Multiword expression recognition
Learning algorithms
Feature selection

ABSTRACT

In multiword expression (MWE) recognition, there exist many studies where different learning methods are employed to decide whether given word combination is a multiword expression. The recognition methods commonly utilize a number of features that are extracted from a data source, frequently from the given text. Though the recognition methods and the features are well studied, we believe that to achieve the best possible performance with a learning method, different subsets of features should also be considered and the best performing subset must be selected.

In this paper, we propose a procedure that covers the performance comparison of well-known feature selection methods to obtain the best feature subset in MWE recognition. The evaluation tests are performed on a Turkish MWE data set and the performance is measured by precision, recall and *F1* values. The highest *F1* value =0.731 is obtained by *C4.5* classifier employing either wrapper or filtering method in feature selection. In the regarding setting(s), it is examined that the performance is increased by 1.11% compared to the setting where all features are employed in classification.

Based on the experimental results, it may be stated that feature selection improves the performance of MWE recognition by eliminating the noisy/non-effective features. Moreover, it is obvious that proposed feature selection method contributes to the overall MWE recognition system by reducing the measurement and storage requirements due to the lower number of features in classification, providing a faster and more-cost effective learning model.

1. Introduction

Multiword expressions (MWEs) are combinations of words that are conventional representations of concepts and/or facts. Those combinations are built from lexemes of sequentially ordered (uninterrupted) or interrupted units in language. Starting from Firth (1957) a number of MWE definitions (Bisht, Dhami, & Tiwari, 2006; Hoey, 1991; Manning & Schütze, 2000; Sinclair 1991) are provided that emphasize different properties of MWEs. For instance, Firth (1957) stated that MWE is the traditional co-occurrence of words. Sag, Baldwin, and Bond (2002) described MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". Though, the researchers tend to define the concept of MWE in different ways, there exists a common understanding/agreement that the set of MWEs encloses idioms, collocations, named-entities and domain-dependent terms. One other commonly accepted fact is that some properties indicate the presence of a MWE. Those properties are language dependency, unitization, domain-dependency and arbitrariness.

Language dependency is generally realized when an expression is translated from one language to another. In translation of MWEs, it is observed that word-to-word translation commonly fails since in different languages, different words combine to represent same concepts. For example, in English, the term "wisdom teeth" expresses the teeth that erupt between the ages of 17 and 25. Word-to-word translation of wisdom teeth results with an expression,"akıl dişleri", that is never used in Turkish for this concept. In Turkish, the matching MWE is "yirmi yaş dişleri" which may be translated to English as "the teeth of 20 years old".

The unitization principle in MWEs is that the words in MWEs unite building a new semantic/syntactic unit in language. This is why, when the text that includes MWEs are processed, such words must be accepted as a single unit. The most salient examples of unitization are observed in idioms since the composing words may change their meanings completely when they unite. For instance, in Turkish, "çamura yatmak" is an idiom that may be translated as "to be in mud" ignoring the unitization principle. But actually, the expression means idiomatically "not to keep up one's word".

Due to the domain dependency principle in MWEs, some expressions that belong to a specific domain may have a completely different meaning that cannot be extracted from the meanings of the composing words. For instance, the expression "terzi kası" in Turkish is a domain dependent MWE that means "Sartorius muscle" in medicine domain but in everyday language it may be trans-

E-mail address: senem.kumova@ieu.edu.tr

https://doi.org/10.1016/j.eswa.2017.09.047
0957-4174/

lated/understood as “tailor’s muscle” rationally due to the lack of domain knowledge.

The arbitrariness property of MWEs that is firstly assigned to collocations indicates that the words arbitrarily unite to build a MWE. In other words, it cannot be explained why/how/which words unite to compose a MWE. For example, “canı tez” is a MWE in Turkish that means “impatient”, but there is no syntactic/semantic reason why “canı süratli” is not a MWE though “süratli” (in English fast) is given as the synonym of the word “tez” in dictionaries.

The ambiguity in the MWE concept and the lack of rules that recognize the MWEs directed researchers to identify MWEs based on some evidences. Those evidences are actually linguistic and/or statistical features that indicate the presence of a MWE in the given text and/or decide if the given word combination is a MWE.

In feature-based MWE recognition, firstly a feature-value that points out how close is the candidate (word combination) to be a MWE is measured from a data source (e.g. corpus, web) for regarding feature. Secondly, based on the feature-values, candidates may be assessed relative to the others or each candidate may be classified as MWE/non-MWE. In previous studies (e.g. Kumova Metin, 2016; Kumova Metin & Karaoğlan, 2010; Pecina, 2008; Tsvetkov & Wintner, 2013) many different features are reported to be effective in MWE identification. On the other hand, there exist a number of studies that several MWE features are utilized together by machine-learning methods (e.g. Pecina, 2008; Tsvetkov & Wintner, 2013). Though it is observed that the use of features together commonly increases the performance in identification of MWEs, it has also some drawbacks. One of the major drawbacks is observed when there are a large number of features. In such cases, the overall effort and the total time required in training and/or measuring the feature-values may reach to such high scores that directs the researchers to a mandatory simplification of the recognition model. One other important drawback is that when all MWE features are employed together, some features may fail in MWE identification and reduce the overall performance though they may succeed individually.

The main aim of our study is two-fold. First is to demonstrate that in feature-based MWE recognition a prior feature selection process improves the performance. Second is to present a systematic way of feature selection in order to determine the best set of features. We believe that identification of best MWE recognition feature set will lead manifold contributions to the natural language processing applications where MWE recognition is a prior task to be performed. The first contribution is that the overall performance of application will be changed in parallel to the performance increase in MWE recognition step. The second is that the total response time of application will be reduced due to the less number of features to be measured and processed. The third contribution will be on data storage. Namely, the simplified learning model and the lower number of features will require less amount of storage space. And the last is that the application will be simplified/improved in terms of coding.

In this study, a language independent feature selection procedure is proposed where well known feature selection methods; wrappers and filters; are utilized with many different learning algorithms/evaluators. In our experiments, a set of 27 statistical and 10 linguistics features are assessed with recall, precision and F-measures on a Turkish MWE data set of 8176 candidates (48.26% MWE, 51.74% non-MWE). We also presented some modifications on a group of linguistic features that are already defined for English in order to be used in MWE recognition in Turkish.

The experimental results showed that the best (reduced) feature set for both wrapper and filtering methods improves MWE identification performance when compared to the whole set of features. Furthermore, it is examined that the proposed feature selection procedure enhances the overall performance in MWE identification. To our knowledge, no systematic research exists addressing the feature selection in MWE recognition to this scope and offering a procedure to select best features in recognition. In addition, there is no study that offers a best performing feature set in Turkish MWE recognition.

This paper is organized as follows. In Section 2, we review the related work on MWE identification methods. Section 3 introduces the MWE features considered in our study. In Section 4, proposed procedure and feature selection methods are presented. Section 5 covers the experimental settings where data set, evaluation measures and set-up are explained. The experimental results are given in Section 6 and the paper is concluded in Section 7.

2. Related work

The MWE identification is defined simply as scoring the candidate word combinations from a given corpus according to their potential to be a MWE (Bouma, 2010) The identification procedure commonly includes 3 stages. Briefly, in the first stage, the candidates are selected to create a data set. Secondly, the candidates are ranked (Seretan, 2011) or classified based on the relations among the words and/or some linguistic features. The last stage includes the evaluation of identification performance. In this section, MWE recognition stages will be explained briefly and different approaches followed in each stage will be given. Table 1 presents summary information on aforementioned three stages for a number of previous works.

The preparation of candidate MWE set includes three important requirements to be satisfied. First is that a corpus that include a wide range of texts that may represent the language must be provided. In earlier studies on MWE recognition, it is explicit that due to the lack of a large corpus in different languages, the researchers tend to run their methods on English corpora. But currently, large corpora in different languages are available and this enabled researchers proposing methods specific to different languages (e.g. Kim, Yoon, & Song, 2001; Li, Lu, & Liu, 2007 respectively Korean and Chinese). The second requirement is the selection of word combinations/candidates of MWE data set. There exist several methods to select the candidates. For example, Evert and Krenn (2001) employed part of speech (POS) tags and selected the uninterrupted two-word sequences (bigrams) that are tagged as *adjective+noun* as candidates (as given in Table 1). Kumova Metin and Karaoğlan (2010) selected the candidates in data set based on statistical measures such as occurrence frequency, mutual information and chi-square similar to Pearce (2002). The third requirement in MWE data set preparation is the annotation of data set that includes both positive and negative examples. The annotation is defined simply as the procedure that the candidates are labelled as MWE (positive example) or non-MWE (negative example) by multiple judges. The purpose of employing multiple judges in annotation is simply having reliable and commonly agreed labels for the candidates (Schneider et al., 2014). The works of Pecina (2008) and Tsvetkov and Wintner (2013) may be given as examples where multiple annotators such as domain experts are employed. On the other hand, in a group of studies such as Pearce (2002) and Kumova Metin (2016) several dictionaries are used to label the candidates in order to have a more objective and reliable annotation of the data set.

In literature, there exist a number of studies where a variety of measures/features are used in ranking or classifying the MWE candidates as the second stage of MWE identification. In ranking approach, the candidates are sorted based on the predefined feature or a group of features. The expectation in ranking is that the candidates that hold the lower ranks in sorted lists have a higher potential to be a MWE compared to the candidates that

Table 1
A number of previous MWE identification studies.

Study	MWE Data Set Preparation				MWE Recognition		Evaluation
	Data source/ Language	*Generation Method*	*MWE type*	*Annotation*	*Recognition Method*	*Number- Type of features*	*Measure*
Frantzi et al. (2000)	Medical corpus / English	POS tag & stop word filtering	n-gram	Manual by domain expert	Ranking (N-best list)	2-statistical	Precision
Evert & Krenn (2001)	German corpora / German	POS tag filtering	bigram trigram	Manual by annotators	Ranking (N-best list)	5-statistical	Precision Recall
Pearce (2002)	British National corpus / English	Ranking	bigram	Dictionary	Ranking (N-best list)	6-statistical	Precision Recall
Pecina (2008)	Frankfurter Rundschau corpus/ German and Prague Dependency Treebank/Czech	POS tag filtering	bigram	Manual by domain experts & annotators	Classification (logistic regression, LDA, neural networks)	55-statistical	Precision
Kumova Metin and Karaoğlan (2010)	Turkish corpora	Ranking	bigram	Dictionary	Ranking (N-best list)	5-statistical	Precision Recall
Tsvetkov and Wintner (2013)	Hebrew-English Bilingual corpus / English	Ranking	bigram	Manual by annotators	Classification (Bayesian network)	8-linguistic	Precision Recall Accuracy F1
Kumova Metin (2016)	Leipzig corpora / English	Ranking	bigram trigram	Dictionary	Ranking	1-statistical	F1
Oflazer et al. (2004)	Turkish corpora / Turkish	POS tag filtering	n-gram	Manual by researchers	Rule based	1110-linguistic (Rules)	Precision Recall

hold higher ranks. The major disadvantage of ranking approach is the need to determine a threshold value/rank/score that splits the list in two groups as MWE and non-MWE. In classification and/or clustering approach, the candidates are labelled either as MWE or non-MWE considering different methods. For example, several machine-learning methods; such as linear discriminant analysis, logistic regressions; are used with 55 statistical features in MWE identification in the study of Pecina (2008). Tsvetkov and Wintner (2013) presented a Bayesian network considering only linguistic properties to detect MWEs. Though, the methods are important in MWE identification, employed features have also a very important role in performance for both ranking and classification approaches. Therefore, in this study, we offer a feature selection procedure prior to classification and/or ranking. To the best of our knowledge, no MWE recognition study exists where a subset of best performing features is build by such a systematic way.

The features to determine MWEs are roughly classified in two groups: statistical and linguistic features. Korkontzelos (2010) categorized the statistical features as unit-hood and term-hood features. The unit-hood/unitization features, commonly referred as association measures, are measuring the strength of the ties between the words composing the word combination. The strong ties between the words indicate that the word combination is a MWE and vice versa. In literature, there exist numerous unitization features such as point-wise mutual information, occurrence frequency, conditional probability, T-test and Pearson chi-square test. The unitization features are well studied and discussed in many previous works (e.g. Manning & Schütze, 2000; Pearce, 2002; Pecina, 2008; Seretan, 2011). The term-hood features are simply measuring the potential of the given word combination to create a term by assessing the strength of ties surrounding the combination. It is accepted that if the surrounding ties (e.g. the tie(s) between the neighbouring word(s) and the candidate combination) are weak, the combination tends to generate a term; as a result the candidate combination is expected to be a MWE. The well-known examples of term-hood features are *C* and *NC* values presented by Frantzi, Ananiadou, and Mima (2000) A similar effort to identify MWEs based on statistical features includes a group of studies where the distribution of the neighbour/surrounding words is examined by vector-based approaches (Kiela & Clark, 2013; Krčmář, Ježek, & Pecina, 2013; Reddy, McCarthy, Manandhar, & Gella, 2011; Salehi, Cook, & Baldwin, 2014). The ultimate goal of these studies is determining the semantic compositionality of word combinations. Since it is assumed that the non-compositional word combinations are MWEs (Krcmár, Jezek, & Pecina, 2013) the methods presented are also used in MWE identification (Kumova Metin, 2016).

The idea of employing linguistics features in MWE identification grounds on the change in surface forms of word combinations when they are used as MWEs. In other words, the linguistic structure of MWEs may differ from the structure of random word combinations due to some linguistic rules that are applied specifically for MWEs. For example, Tsvetkov and Wintner (2013) proposed to accept the dash "-" character as an indicator of MWEs since it is permitted to use a dash "-" between the words composing a MWE. As a result, if there exists a dash between the constituents of a candidate, it is accepted that the candidate has a potential to be a MWE. One other well-known set of linguistic features encounters the part of speech tag sequences of the word combinations. For example, Oflazer, Cetinoglu, and Say (2004) defined several POS (part of speech) sequences such as "adjective+noun", "noun+noun", as indicators of MWEs in Turkish and build a rule-based system based on POS pattern matching.

The evaluation of MWE identification performance is commonly performed by precision, recall and F1-measures. Evert and Krenn (2001) discussed that there exists some weakness in the evaluation of the identification systems. Later, Evert (2004) proposed the use of significance tests in performance evaluation.

3. MWE features

In this study, MWE recognition features are categorized as statistical and linguistic features. The statistical class includes the occurrence frequency based measures where the frequencies are obtained from the corpus. The linguistic features focus the structural differences of MWEs that are observed using different sources such as the corpus, dictionaries and linguistic rules. The following subsections detail the statistical and linguistic features respectively.

Table 2
Statistical features – association features.

Feature	Formula
1. Joint probability (*JP*)	$P(w_1w_2)$
2. Conditional probability (*CP*)	$P(w_2\|w_1)$
3. Reverse conditional probability (*RCP*)	$P(w_1\|w_2)$
4. Pointwise mutual information (*PMI*)	$\log \frac{P(w_1w_2)}{P(w_1)P(w_2)}$
5. Mutual dependency (*MD*)	$\log \frac{P(w_1w_2)^2}{P(w_1)P(w_2)}$
6. Log frequency biased *MD* (*LFMD*)	$\log \frac{P(w_1w_2)^2}{P(w_1)P(w_2)} + \log P(w_1w_2)$
7. Normalized expectation (*NE*)	$\frac{2f(w_1w_2)}{f(w_1)+f(w_2)}$
8. S cost (*Scost*)	$\log(1+\frac{\min(f(w_1\ \overline{w_2}),\ f(\ \overline{w_1}w_2\ ))}{f(w_1w_2)+1})$
9. U cost (*Ucost*)	$\log(1+\frac{\min(f(w_1\ \overline{w_2}),\ f(\ \overline{w_1}w_2\ ))+\ f(w_1w_2)}{\max(f(w_1\ \overline{w_2}),\ f(\ \overline{w_1}w_2\ ))+\ f(w_1w_2)})$
10. R cost (*Rcost*)	$\log(1+\frac{f(w_1w_2)}{(f(w_1w_2)+f(w_1\ \overline{w_2}\ ))}) + \log(1+\frac{f(w_1w_2)}{(f(w_1w_2)+f(\ \overline{w_1}\ w_2))})$
11. First Kulczynsky (*FK*)	$\frac{f(w_1w_2)}{f(w_1\ \overline{w_2}\ )+\ f(\ \overline{w_1}\ w_2)}$
12. Second Kulczynsky (*SK*)	$\frac{1}{2}(\frac{f(w_1w_2)}{(f(w_1w_2)+f(w_1\ \overline{w_2}\ ))} + \frac{f(w_1w_2)}{(f(w_1w_2)+f(\ \overline{w_1}\ w_2))})$
13. Braun-Blanquet (*BB*)	$\frac{f(w_1w_2)}{\max(f(w_1w_2)+f(w_1\ \overline{w_2}\ ),\ f(w_1w_2)+f(\ \overline{w_1}\ w_2)}$
14. Simps (*Simps*)	$\frac{f(w_1w_2)}{\min(f(w_1w_2)+f(w_1\ \overline{w_2}\ ),\ f(w_1w_2)+f(\ \overline{w_1}\ w_2)}$
15. Driver-Kroeber (*DK*)	$\frac{f(w_1w_2)}{\sqrt{(f(w_1w_2)+f(w_1\ \overline{w_2}\ )\cdot\ (f(w_1w_2)+f(\ \overline{w_1}\ w_2)}}$
16. Piatersky-Shapiro (*PS*)	$P(w_1w_2) - P(w_1)P(w_2)$
17. J (*J*)	$\frac{f(w_1w_2)}{f(w_1w_2)+\ f(w_1\ \overline{w_2}\ )+\ f(\ \overline{w_1}\ w_2)}$
18. Second Sokal-Sneath (*SSS*)	$\frac{f(w_1w_2)}{f(w_1w_2)+\ 2(f(w_1\ \overline{w_2}\ )+\ f(\ \overline{w_1}\ w_2))}$
19. Mountord (*Mount*)	$\frac{2f(w_1w_2)}{2f(w_1\ \overline{w_2}\ )f(\ \overline{w_1}\ w_2)+\ f(w_1w_2)f(w_1\ \overline{w_2}\ )+f(w_1w_2)f(\ \overline{w_1}\ w_2)}$
20. Fager (*F*)	$\frac{f(w_1w_2)}{\sqrt{(f(w_1w_2)+f(w_1\ \overline{w_2}\ )\cdot\ (f(w_1w_2)+f(\ \overline{w_1}\ w_2)}} - \frac{1}{2}\ \max(f(w_1\ \overline{w_2}\ )f(\ \overline{w_1}\ w_2))$

3.1. Statistical features

The strength of ties between the composing words and the weakness of the ties with the surrounding words are properties that distinguish MWEs from the other word combinations. The features that measure the strength/weakness of ties based on occurrence frequencies are accepted to be statistical features. The statistical features are considered in two types: association and term-hood features. These two types of features are defined as follows.

3.1.1. Association features

Association features, known as association measures, aim to grade the degree of association between the words (Pecina, 2008). Simply, they are measuring the strength of ties between the composing words in word combination. The common assumption in the association features is that the words unite to build a new unit when they co-occur frequently in language. In this study, 20 association measures given in Table 2 are utilized. The association measures given in Table 2 enable the ranking of MWE candidates. It is accepted that as if association value of a candidate is higher/lower than the other, the regarding candidate has a higher tendency to form a MWE. In Table 2, w_1 and w_2 represent the first and the second word in given MWE candidate, respectively. $P(w_1w_2)$ is the probability of co-occurrence of two words; $P(w_1)$, $P(w_2)$ are the occurrence probabilities of first and the second words in candidate. $P(w_i|w_j)$ gives the conditional occurrence probability of w_i given that the word w_j is observed. $f(w_1w_2)$, (w_1), $f(w_2)$ are occurrence frequency of w_1w_2, w_1 and w_2 respectively.

3.1.2. Term-hood features

Association features focus on the assumption that the co-occurrence frequency of words composing a MWE is higher compared to random word combinations. One of the significant drawbacks of this assumption is observed for MWEs, such as technical terms, which are not frequently used in everyday language. This type of MWEs are identified by term-hood features, which consider ties of the given word combination with the surrounding words. In term-hood measures, it is assumed that the words composing the MWE have weaker ties with surrounding words compared to their inner ties. The term-hood features that are utilized in our experiments and their short descriptions are presented in Table 3.

3.2. Linguistic features

Linguistic features enable the identification of MWEs by examination of the properties that are extracted from written texts such as part of speech tags, inflectional/constructional suffixes of words, the use of upper/lower cases in words. Even though linguistic features are advantageous to be independent of occurrence frequency, they require pre-processing of written texts. For example, in order to employ a linguistic feature considering suffixes, the words in corpus must be stemmed. Moreover, the performance of stemming algorithm also affects indirectly the performance of MWE identification.

In this study, the linguistic features that do not require the use of a pre-processing tool are preferred. The features/ the group of features are as follows

1. Partial variety in surface form (PVSF)
2. Orthographical variety (*OV*)
3. Frozen form (FF)
4. Hapax-Fossil (*HF*)
5. The ratio of upper case letters (*UC*)
6. The suffix sequence (*SS*)
7. Named entity words (*NEW*)

3.2.1. Partial variety in surface forms (PVSF)

The surface form of words may vary due to the inflectional suffixes and some punctuation marks. In Turkish, different word forms may be obtained by concatenating inflectional morphemes (suffixes) to the stems. For instance, the word “defter” (notebook) may be observed in different surface forms such as “defter*inin*” (*his* notebook), “defter*de*” (*in the* notebook) and “defter*ler*” (notebooks).

In different definitions of MWE, it is stated that the words unite to build a new syntactical and/or semantic unit. Since MWEs are units in language, we accepted that in Turkish MWEs, there might be no change in surface form of former words and a limited number of changes in the last word of the MWE. To measure the partial variety in surface forms, the histogram presenting the occurrence frequencies of different surface forms belonging to the same MWE candidate is used as it is proposed in the study of Tsvetkov and Wintner (2013). If the surface form histogram shows a uniform distribution, it indicates that no form is dominant over the others meaning that the candidate is not a true MWE. On the other

Table 3
Statistical features – term-hood features.

No	Term-hood features	Description
1	Bigram Forward Variety (*BFV*)	*BFV* is the ratio of different number of words following the bigram ($v_f(w_1w_2)$) to the occurrence frequency of bigram ($f(w_1w_2)$) $BFV(w_1w_2) = \frac{v_f(w_1w_2)}{f(w_1w_2)}$ *BFV* value is expected to be higher for MWEs compared to random word combinations.
2	Bigram Backward Variety (*BBV*)	*BBV* is the ratio of different number of words preceding the bigram ($v_b(w_1w_2)$) to the occurrence frequency of bigram ($f(w_1w_2)$) $BBV(w_1w_2) = \frac{v_b(w_1w_2)}{f(w_1w_2)}$ *BBV* value is expected to be higher for MWEs compared to random word combinations.
3	Word Forward Variety (*WFV*)	*WFV* is the ratio of different number of words following the second word of bigram ($v_f(w_2)$) to the occurrence frequency of the same word ($f(w_2)$) $WFV(w_1w_2) = \frac{v_f(w_2)}{f(w_2)}$ *WFV* value is expected to be high for MWEs.
4	Word Backward Variety (*WBV*)	*WBV* is the ratio of different number of words preceding the first word of bigram ($v_b(w_1)$) to the occurrence frequency of the same word ($f(w_1)$) $WBV(w_1w_2) = \frac{v_b(w_1)}{f(w_1)}$ *WBV* value is expected to be high for MWEs.
5	Bigram/Word Forward Variety (*BWFV*)	*BWFV* is the ratio of different number of words following the bigram ($v_f(w_1w_2)$) to different number of words following the second word of bigram($v_f(w_2)$) $BWFV(w_1w_2) = \frac{v_f(w_1w_2)}{v_f(w_2)}$ *BWFV* value is expected to be high for MWEs
6	Bigram/Word Backward Variety (*BWBV*)	*BWBV* is the ratio of different number of words preceding the bigram ($v_b(w_1w_2)$) to different number of words preceding the first word of bigram($v_b(w_1)$) $BWBV(w_1w_2) = \frac{v_b(w_1w_2)}{v_b(w_1)}$ *BWBV* value is expected to be high for MWEs
7	Neighbourhood Unpredictability (*NUP*) (Kumova Metin, 2016)	*NUP* is combined feature that considers both *BWFV* and *BWBV* values as follows $FNUP(w_1w_2) = 1 - \frac{v_f(w_1w_2)-1}{v_f(w_2)-1}$ $BNUP(w_1w_2) = 1 - \frac{v_b(w_1w_2)-1}{v_b(w_1)-1}$ $NUP(w_1w_2) = \sqrt{FNUP(w_1w_2)^2 + BNUP(w_1w_2)^2}$ *NUP* feature gets the values in a predefined range. A low *NUP* value indicate that the regarding candidate is a MWE.

case, where a particular form has a higher number of occurrences, candidate is expected to be a true MWE.

This feature is supposed to be effective in identification of especially idioms and/or idiomatic expressions.

In this study, variety in surface forms is measured in two different ways that are called as *PVSFm* and *PVSFn* features. The below 4 step procedure is followed in for each MWE candidate in order to obtain *PVSFm* and *PVSFn* values:

I. In corpus, the occurrence frequency of each string (character sequences) that begins with the given candidate is measured individually. The frequencies are sorted in decreasing order and are stored in PVSF vector ($[f_1\ f_2\ f_3\ \ldots\ f_n]$) where n represents the number of different surface forms belonging to the regarding candidate.

II. The average surface form frequency of a candidate is the ratio of sum of frequencies (summation of frequencies of different surface forms belonging to the same candidate) in PVSF vector to the length of the regarding vector. The average value is used to construct the *uniform distribution vector* ($[fu_1\ fu_2\ fu_3\ \ldots\ fu_n]$). The *uniform distribution vector* is a synthetic vector that has same number of elements with PVSF vector where all the elements hold the same average value.

III. The distance PVSF vector to the uniform-distribution vector is measured by Manhattan distance. Manhattan distance between two vectors of n elements are given as

$$Manhattan\ (f, fu) = \sum_{i=1}^{i=n} |f_i - fu_i|$$

Manhattan value is higher for the candidates where the occurrence frequency of a particular surface form is higher compared to the other forms, meaning that this surface form is a MWE. In this study, Manhattan distance is accepted to be the first feature, *PVSFm*, considering the variety in surface forms.

IV. Normalized *PVSFm* value, *PVSFn*, is the ratio of *PVSFm* to the sum of elements (sum of frequencies) in PVSF vector. *PVSFn* is calculated as follows

$$PVSF_n = \frac{Manhattan\ (f, fu)}{\sum_{i=1}^{i=n} f_i} = \frac{PVSF_m}{\sum_{i=1}^{i=n} f_i}$$

PVSFn features being independent of the frequency enables the comparison of the candidates that occur in different frequencies in corpus.

In Fig. 1, the histogram representing the PVSF vector of the candidate “organik madde” (organic substance) is given as example. Since “organik madde” is observed

- 84 times as “organik madde”
- 5 times as “organik madde*ce*” (according to the organic substance)
- 1 time as “organik madde*den*” (from the organic substance)
- 1 time as “organik madde*dir*” (it is the organic substance)
- 38 times as “organik madde*ler*” (organic substances)
- 2 times as “organik madde*lerce*” (according to the organic substances)
- 5 times as “organik madde*lerden*” (from the organic substances)
- 5 times as “organik madde*lerdir*” (they are the organic substances)
- 2 times as “organik madde*lere*” (to the organic substances)
- 6 times as “organik madde*leri*” (organic substances in … form)
- 31 times as “organic madde*lerin*” (accusative form of the organic substances)
- 8 times as “organik madde*lerle*” (with the organic substances)
- 11 times as “organik madde*nin*” (… of the organic substance)
- 1 time as “organik madde*si*” (the organic substance of …)
- 2 times as “organik madde*ye*”(to the organic substance)
- 7 times as “organik madde*yi*” (accusative form of the organic substance)

in corpus, the resulting PVSF vector is f= [84 38 31 11 8 7 6 5 5 5 2 2 2 1 1 1] that has 16 elements and the average frequency is calculated as 209/16=13,0625. As a result, the uniform_distribution vector (fu) is composed of 209 elements of the average value (13.0625). The Manhattan value *PVSFm* is $PVSF_m = |84 - 13.0625| + |38 - 13.0625| + |31 - 13.0625| + \ldots |1 - 13.0625| + |1 - 13.0625| + |1 - 13.0625| = 227.625$ and the normalized Manhattan value is $PVSF_n = 227.625/209 = 1.089$ for the candidate “organic madde”.

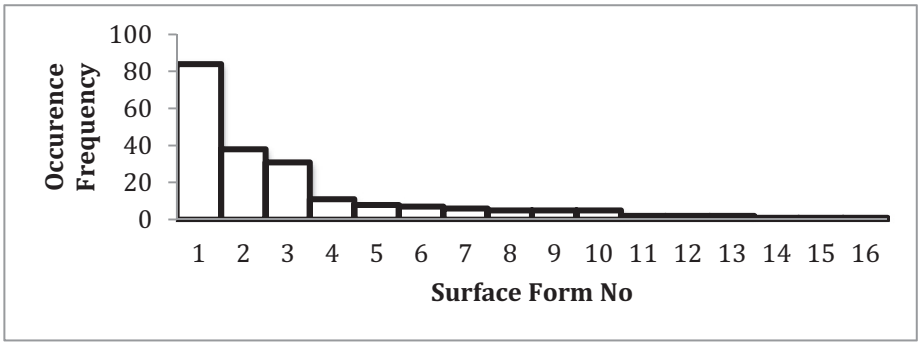

Fig. 1. The surface form frequency histogram of the candidate "organik madde".

3.2.2. Orthographical variety

In Turkish texts, MWEs are examined to hold orthographical changes to the some of the punctuation marks. The most commonly observed mark is the hypen"-" symbol that is used to connect the constituents of MWEs. For instance, the term e-mail has two different forms: "e-posta" and "e posta". One other symbol that changes the surface form is the aposthrophe " ' ". The aposthrophe is used if a suffix is to be added to a proper noun. For example, in sentence "Ali'nin kitabı kaybolmuş" (Ali's book is lost), apostrophe is used between the noun and the possessive suffix. *OV* is examined by two different features: orthographical variety due to the use of hyphen (*OV_h*) and orthographical variety due to the apostrophe (*OV_a*) symbol.

OV_h value is the proportion of the two occurrence frequencies of candidate that is formed with a hyphen and without a hyphen. The lower occurrence frequency is divided by the higher frequency value, in order to obtain an *OV_h* value in range [0 1]. For instance, assuming that the term "bağ kur", which is a type of public insurance system in Turkey, is observed as "bağ kur" 239 times and as "bağ-kur" 854 times, the *OV_h* value is assigned to 239/854=0.289.

OV_a value of a given candidate is measured in a similar manner to *OV_h* value. The number of occurrences of the second word in candidate with and without apostrophe symbol is counted for each candidate. The lower of these two frequencies is divided by the higher frequency value. For example, for the MWE candidate "gito yaylasından" (from the Gito plateau), the occurrences of the form "yaylasından" and "yaylası'ndan" are observed as 167 and 109 respectively. *OV_a* value of the candidate is 109/167=0.65.

*3.2.3. Frozen Form (*FF*)*

The words in some of the MWEs are observed in a single surface form, which is called as frozen form (Tsvetkov & Wintner, 2013). The constituting words in frozen formed MWEs may be used in same location in different MWEs but when they are used together they do not change their surface forms. For instance, the expression "arkası arkasına" (sequentially) is an MWE of frozen form type. The concatenation of any suffix to the expression may change the meaning (e.g. "arkası arkasınalar" means "one is behind the other"). *FF* is a binary valued feature that *FF* value of a candidate is 1 if it has a single surface form in corpus and 0 vice versa.

*3.2.4. Hapax-Fossil (*HF*)*

"Hapax legomenon" is a word in Greek, that stands for the words in language that occurs only once within a context. The identification of such words is important in the analysis of meaning, disambiguation of the language of text. In the study of Tsvetkov and Wintner (2013) it is stated that some of the MWEs are hapax legomena. In other words, the words composing a MWE, may be the words that do not occur with other words. For example, "hokus pokus" (hocus-pocus) is a MWE that holds constituents that do never co-occur with words except each other. The constituents in hapax legomenon are commonly called as hapax words. Similar to the hapax words, a group of words in language loses their frequent usage in language in time. Such words are named as fossil words. Fossil words, similar to the hapax words, may build MWEs.

In order to identify the MWEs that are composed of hapax or fossil words, it is required to have a list of hapax and fossil words. Since hapax-fossil word lists are not available, a list of HF bigrams together with their occurrence frequencies is built from the corpus. The *HF* bigrams are the bigrams that include words that do not occur in same location in a bigram more than 3 times in the corpus. *HF* value of a given candidate that is a *HF* bigram is the ratio of the occurrence frequency of the candidate to the maximum occurrence frequency in HF bigram list. If the given candidate is not a *HF* bigram the *HF* value is accepted to be zero.

*3.2.5. The ratio of upper case letters (*UC*)*

In identification of MWEs, the use of capital letters is considered as a MWE indicator since they are commonly observed in a special type of MWEs such as personal and location names. *UC* is the ratio of occurrence frequency of the candidate where capital letters are used to the total frequency of the candidate. *UC* value ranges in [0 1], as the value reaches to 1 the expectation of being a MWE increases.

UC value of the candidate bigram "araştırma komisyonunun" (Eng. belonging to the research commission) that is observed in following surface forms

i. "araştırma komisyonunun" (occurrence frequency=22)
ii. "Araştırma komisyonunun" (occurrence frequency=4)
iii. "Araştırma Komisyonunun" (occurrence frequency=145)

is calculated as *UC*("araştırma komisyonunun") =(145+4)/(145+4+22) =0.871.

An exceptional case of capital letter use is observed in the first sentences of the sentence. Since the sentence segmentation is not provided in the context of this study, such exceptional cases are not considered.

*3.2.6. The suffix sequence (*SS*)*

The idea grounding the *SS* feature is the expectation of a number of suffixes or suffix sequences to be used with MWEs more than random word/word combinations. In order to determine such suffixes/suffix sequences, firstly an annotated list of MWEs is required. Secondly, this list of MWEs must be stemmed to extract the suffixes. Lastly, it must be validated that the use of these suffixes with MWEs is not random.

Tablo 4
Top most 30 suffix sequences that are used with idioms in combined corpus.

No	Occurrence frequency	SS	No	Occurrence frequency	SS	No	Occurrence frequency	SS
1	15,982	arak	**11**	10,155	erek	**21**	5645	meye
2	15,274	mak	**12**	9945	diği	**22**	5245	tiği
3	15,131	mek	**13**	9082	mesi	**23**	5152	mış
4	14,803	duğunu	**14**	8904	ecek	**24**	4986	ıldı
5	13,276	duğu	**15**	8716	iyor	**25**	4986	miş
6	12,613	ıyor	**16**	8272	diğini	**26**	4853	uyor
7	12,282	acak	**17**	6185	ındı	**27**	4800	ınan
8	11,989	ması	**18**	6135	dığını	**28**	4596	ilmesi
9	11,194	ildi	**19**	5866	maya	**29**	4494	anı
10	10,222	dığı	**20**	5695	tığı	**30**	3339	ici

Since a MWE annotated and stemmed corpus/list is not available in Turkish to the best of our knowledge, we employed the idioms from the Turkish idiom dictionary to calculate values of two different *SS* features: *SS_h* and *SS_n*. Both *SS* features require a suffix sequence list of idioms. The list of idiom suffix sequences is obtained in five steps, defined as follows:

I. All bigrams that include an idiom are listed in the corpus. For example, the corpus contains 3 matching bigrams: "aceleye geldi", "aceleye gelmiş", "aceleye gelmez" that include the idiom "aceleye gel".
II. The matching idiom is stripped out from the surface forms and the remaining sequences of length in range [3 11] are added to the list of idiom suffix sequences. The range is determined empirically.
For example, when the matching idiom is stripped out from the forms "aceleye gel*di*", "aceleye gel*miş*", "aceleye gel*mez*", we obtain "di", "miş" and "mez" suffix sequences, respectively. "miş" and "mez" are added to the list of idiom suffix sequences since their string lengths are in range [3 11].
III. The idiom suffix sequences are sorted in decreasing order of the occurrence frequency. Table 4 presents the top 30 suffix sequences in the sorted list that is obtained from the combined corpus[1].
IV. Observing the suffix sequences of idioms, it is explicitly seen that some of the suffixes in the list are used frequently not only with idioms. For instance, "mez" that is observed in "aceleye gel*mez*" is the negation suffix that is used frequently with almost all the verbs in Turkish.
The suffixes/suffix sequences that are frequently used with all type of words must be eliminated from the lists of idiom suffix sequences. In this study the frequently occurring suffix sequences are determined by the use of a corpus. The last *n* characters of all the words in corpus are extracted and each different character sequence of length *n* is counted. The character sequences are sorted in decreasing order of their occurrence frequencies. In the resulting lists, it is accepted that frequently used suffixes/suffix sequences hold the lowest ranks. *n* is assigned with values in range [3 11]. In Tables 5 and 6, first 10 suffix sequences obtained from the combined corpus are presented.
The corpus suffix list is built by the first N=200 suffix/suffix sequences from the resulting lists. N=200 value is determined empirically.
V. The suffixes/suffix sequences that are both in idiom and corpus list are removed/filtered from the idiom suffix sequences list. The filtering left 256 suffix/suffix sequences that are validated to be frequently used with idioms. The lengths of remaining suffix sequences are in range [3 10]. Table 7 gives first (highest frequency) 100 sequences in the list. For example, the string "ındı" that has 4 characters (l=4) is the most frequent item (No=1) in the idiom suffix sequence list.

SS feature is classified in two sub-features: *SS_h* and *SS_n*. *SS_h* stands for the length of the suffix sequence in given candidate. *SS_h* takes the value 0 or an integer value in range [3 10]. *SS_n* is obtained by dividing the *SS_h* value of the candidate to the maximum suffix length (max=10) in idiom suffix sequence list. *SS_h* varies in range 0 and [0.3 1].

*3.2.7. The use of named-entity words (*NEW*)*

Named-entities (NEs) are information units like names, including person, organization and location names, and numeric expressions including time, date, money/percent expressions (Nadeau & Sekine, 2007). Named-entities are considered as a type of MWEs since they have specific names that have similar properties with the other types of MWEs. *NEW* feature is generating a value that shows the occurrence ratio of named-entity words in the given candidate. As the value gets higher, the expectation that the candidate is a true MWE increases and decreases otherwise. In order to obtain the *NEW* value, a list of named-entity words is required. Since, there is not a NE-word list in Turkish, we proposed an approach to compile NE-words list by using different data sources. The compilation covers the use of 3 different types of sources: NE data sets that are prepared in previous NE recognition studies, popular name and surname lists, addresses/location names provided by public institutions.

First data source of NE-words is the data sets of Küçük and Yazici (2012) and Tatar and Cicekli (2011). In Table 8, the content of data sets provided in both studies are given.

The NEs in this study are categorized in 5, based on the studies of Küçük and Yazici (2012) and Tatar and Cicekli (2011). The categories of NEs are as follows:

1. Personal Name
2. Surname
3. Location/Address
4. Institution/Foundation/Organization
5. Structure (Structure tagged expressions include building/street/highway names such as "Beyoğlu Adliyesi" (The Court House of Beyoğlu), "Güllibağ İstasyonu" (Güllübağ Station), "E-5 Karayolu" (E-5 highway))

The words composing NEs in data sets of Küçük and Yazici (2012), Tatar and Cicekli (2011), except the words that are observed in NEs tagged as date, number and time, are utilized as primary data set in our study. The set of location words (e.g. "… Oteli" ("Hotel …"), "... Okul" ("... School"), "… Hastanesi" ("… Hospital")) is enlarged by examining school names/addresses that are listed

[1] The combined corpus is obtained by combining the texts in Leipzig Quasthoft et. al., 2005), Ege (Tür el. al. 2003) and Bilkent corpora , the details on these corpora are given in Section 5.1

Table 5
Most frequently occurring suffix sequences in combined corpus (length n=3 to 7).

n=3	*f*	*n=4*	*f*	*n=5*	*f*	*n=6*	*f*	*n=7*	*f*
-mak	15,274	-arak	15,982	-mişti	3243	-duğunu	14,803	-acağını	3852
-mek	15,131	-duğu	13,276	-mıştı	2929	-diğini	8272	-eceğini	2262
-mış	5152	-ıyor	12,613	-acağı	2832	-dığını	6135	-acaktır	2089
-miş	4986	-acak	12,282	-anına	2674	-ilmesi	4596	-masının	1609
-anı	4494	-ması	11,989	-madan	2572	-tığını	4489	-dukları	1519
-ici	3339	-ildi	11,194	-mayan	2325	-abilir	3489	-mesinin	1449
-yan	2957	-dığı	10,222	-meden	2102	-tiğini	3485	-tikleri	1288
-muş	1713	-erek	10,155	-irken	1912	-miştir	2909	-mesiyle	1106
-mez	1706	-diği	9945	-anlar	1715	-ılması	2892	-dıkları	1097
-dik	1598	-mesi	9082	-arken	1706	-irildi	2718	-dikleri	1089

Table 6
Most frequently occurring suffix sequences in combined corpus (length n=8 to 11).

n=8	*f*	*n=9*	*f*	*n=10*	*f*	*n=11*	*f*
-madığını	4454	-duklarını	3186	-acaklarını	887	-abileceğini	1078
-maktadır	3668	-diklerini	2317	-eceklerini	668	-madıklarını	633
-mektedir	2962	-dıklarını	2208	-mayacağını	608	-amayacağını	304
-ildiğini	2354	-ındığında	1470	-ilmektedir	464	-abilirsiniz	304
-ındığını	1926	-tiklerini	1106	-meyeceğini	460	-ebileceğini	271
-ıldığını	1648	-abileceği	978	-irildiğini	420	-mektedirler	184
-mediğini	1519	-tıklarını	866	-ilemeyecek	249	-maktadırlar	177
-ilmesini	1466	-ilmelidir	588	-ınmadığını	243	-abilecektir	168
-abilecek	1400	-ileceğini	572	-ilmediğini	220	-mediklerini	166
-ilmiştir	1191	-ınacağını	571	-ülmektedir	178	-abileceğine	159

Table 7
Idiom suffix sequences (first 100 sequences).

No	Suffix sequence	l	No	Suffix sequence	l	No	Suffix sequence	l	No	Suffix sequence	l	No	Suffix sequence	l
1	-ındı	4	21	-diler	5	41	-duğuna	6	61	-ıyordu	6	81	-ıyoruz	6
2	-ınan	4	22	-ırken	5	42	-ilerek	6	62	-ınca	4	82	-duk	3
3	-ülen	4	23	-dik	3	43	-miyor	5	63	-maktan	6	83	-ınmasını	8
4	-ici	3	24	-dukları	7	44	-tiler	5	64	-enler	5	84	-tılar	5
5	-üldü	4	25	-medi	4	45	-laşma	5	65	-malı	4	85	-alım	4
6	-anına	5	26	-ındığında	9	46	-irilen	6	66	-mayacak	7	86	-iyordu	6
7	-ınması	6	27	-ulan	4	47	-müş	3	67	-duğunca	7	87	-acağına	7
8	-madan	5	28	-meyen	5	48	-makta	5	68	-ülüyor	6	88	-mamız	5
9	-düğü	4	29	-madı	4	49	-lamda	5	69	-ılıyor	6	89	-cın	3
10	-ınarak	6	30	-üşme	4	50	-memesi	6	70	-ulduğunu	8	90	-meli	4
11	-ındığı	6	31	-tikleri	7	51	-sun	3	71	-ınmış	5	91	-ilirken	7
12	-meden	5	32	-dim	3	52	-utlu	4	72	-lam	3	92	-ilmişti	7
13	-meyi	4	33	-dık	3	53	-mekte	5	73	-üntü	4	93	-ulacak	6
14	-ındığını	8	34	-ulması	6	54	-ilmiş	5	74	-mıyor	5	94	-makla	5
15	-muştur	6	35	-muştu	5	55	-dılar	5	75	-ılmış	5	95	-urken	5
16	-irken	5	36	-mesiyle	7	56	-amaz	4	76	-mekten	6	96	-irildiği	8
17	-ınacak	6	37	-acağız	6	57	-umlu	4	77	-dım	3	97	-müştü	5
18	-muş	3	38	-ardı	4	58	-eceğiz	6	78	-anının	6	98	-uyoruz	6
19	-arken	5	39	-maması	6	59	-ıcı	3	79	-duğumuz	7	99	-ılarak	6
20	-mez	3	40	-mediği	6	60	-uyorum	6	80	-diğinde	7	100	-ınacağını	9

Table 8
The data sets in Küçük and Yazici (2012), Tatar and Cicekli (2011) studies.

NE Tag	Tatar and Çicekli (2011)	Küçük and Yazıcı (2012)
Date	58	–
Instrument	1	–
Location	281	571
Number	148	–
Organization	165	456
Person	137	398
Structure	30	–
Time	68	–
Vehicle	4	–
TOTAL	892	1425

in the official web pages of Ministry of Education, agency/member names/addresses of private insurance companies and chambers of Commerce. The words that are commonly used in addresses such as

"... Iskele*si*"	(... Port),
"... Camis*i*"	(The ... Mosque),
"... Caddes*i*"	(... Street),
"... Çarşısı"	(... Bazaar)

are all included into the set of location words.

The set of names and surnames is enhanced by the list of 50 most popular name and surname provided by Directorate General of Civil Registration and Citizenship Affairs and 3334 popular baby names provided by different web sources.

Table 9 presents the content of the extended list of NE-words together with some examples and the number of words in each category.

The *NEW* value of a MWE candidate is obtained by examining whether the constituents of candidate are in the NE-words list. The first constituent of candidate is queried among the personal names and the second is searched in the other categories of NE-words. Since NE-words list has 5 categories, theoretically max-

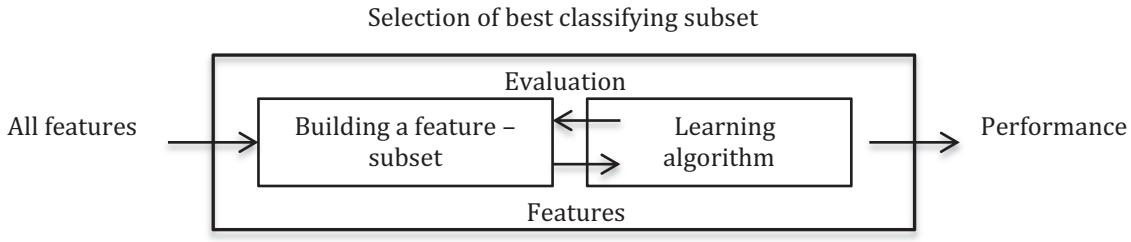

Fig. 2. The flow chart of wrappers.

Table 9
The statistics of NE-words list.

Category	Number of words	Examples
Personal Names	3354	"Abdullah", "Salime", "Recep"
Surnames	151	"Yüksel", "Yılmaz", "Özkan", "Kılıç"
Location/Address	79	"Mahallesi", "İlkokulu", "Caddesi"
Institution/Foundation/ Organization	36	"Bankası", "komitesi", "Vakfı"
Structure	6	"Adliye", "Demiryolu", "Karayolu"

imum number of matches is 5 and minimum is zero. The *NEW* value is the number of matches divided by the maximum number of matches. For instance, the candidate "Aral Gölü" (Aral Lake) returned 2 matches, one match to personal names other to location/address list, resulting with *NEW* ("Aral Gölü")=2/5=0.4.

4. Feature selection procedure: the use of wrappers and filtering

The feature selection in machine learning is a pre-process that enables the simplification of classification models, reducing the training time and understanding of the data set. The feature selection methods fall into three categories (Guyon & Elisseeff, 2003): wrappers, filtering and embedded methods where wrappers and filtering methods are integrated.

In this study, we consider wrappers and filtering in determination of MWE feature subsets. The main aim in wrappers is the selection of the feature subset that gives the highest classification performance. In wrappers, the first task is generation of the subsets. Following, for each subset the classification model is trained and classification performance is measured on a testing (hold-out) set. Fig. 2 gives the procedure followed in wrappers.

As the wrappers require the generation of all subsets and the training of the classification model for each new subset, they are computationally intensive. One other factor that is closely related to the overall performance of the wrappers is the searching algorithm that is used in determination of order of subsets. The algorithms that reach to the best subset faster decrease the execution time of the overall classification system.

In determination of MWE feature set by the wrappers, we propose the use of k-fold approach with different learning algorithms. In this procedure, firstly, predefined learning algorithms are executed individually to extract a subset of features in a k-fold manner. In k-fold approach, each learning algorithm returns a list of features together with the number of folds that the given feature is successful in classification. Following, the results of learning algorithms are compared and the features that are observed commonly in a high number of folds in different learning methods are merged to build the MWE feature subset. The proposed k-fold approach has the advantage of considering different proportions of the data set and learning algorithms individually. On the other hand the number of operations to determine the best subset may be seen as a disadvantage of the method.

In filtering, the classification performance of each feature is assessed individually by a predetermined attribute evaluator. In other words, the features are determined regardless of the machine-learning model. There are a number of attribute evaluators, such as Pearson correlation, mutual information and Kendall correlation. In Fig. 3, the filtering procedure is presented.

The filtering approach enables the sorting of features based on decreasing order of their attribute/evaluator score. It is possible to select the subset of features in different ways based on the sorted feature lists. One approach is that the features that generate scores higher than a user-defined threshold score may be merged to build a set of features. In other approaches, a desired number of features or a predefined proportion of features may build the subset. In filtering, lastly the performance of feature subset is measured by the application of a learning algorithm that is also determined by the researchers. The significant drawbacks of filtering are dependency on empirical thresholds of subset sizes and ignoring the relations between the features by assessing them individually, which may result with the selection of redundant features in classification.

In determination of MWE subset by filtering, we propose two different approaches to determine the threshold scores based on the common agreement of different evaluators. The proposed approaches differ in the way they assess the agreement among the evaluators. In the first approach, the agreement is measured by Kendall-*W* value, accepting evaluators as the raters and MWE features as the data to be ranked. Second approach bases on measuring the agreement on features individually. Simply, in each iteration, a different feature is assessed whether it is commonly accepted to be successful in MWE detection for a large number of evaluators. The steps to follow for both approaches are presented in detail in Sections 5 and 6 by experimental results on MWE data set.

The feature selection procedure ends with selecting the best classifying subset by the performance comparison of the subsets offered by wrappers and filters. The following subsections briefly explain wrappers, filtering and detail the evaluators/learning algorithms utilized in our study.

4.1. Wrappers

The ultimate goal in wrappers is scoring feature subsets and determining the best performing subset as mentioned in preceding section. In this study, in order to determine the best subset of features in MWE classification seven learning algorithms

1. Naive Bayes (NB)
2. *C4.5*
3. K-nearest neighbour (K-nn)
4. Logistic Regression (LR)
5. Random Forests (RF)
6. Support Vector Machine (SVM)
7. Multilayer Perceptron (MP)

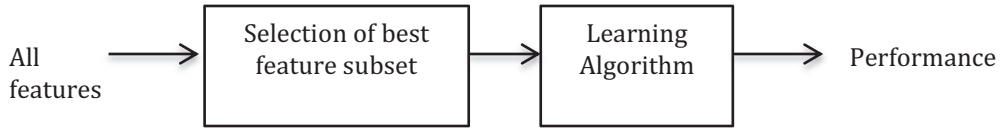

Fig. 3. The flow chart of filtering.

are implemented and the subsets are generated by the best first search strategy. The experiments are validated in 5-folds.

Naive Bayes is a classifier grounding to the Bayes Theorem where the features are assumed to be independent of each other and equally important. In Bayes theorem, it is accepted that each sample is represented by a feature vector. For example, $x=[x_1\ x_2\ x_3 \ldots x_{n-1}\ x_n]$ is a sample with n features. And the conditional probability of the sample $x=[x_1\ x_2\ x_3 \ldots x_{n-1}\ x_n]$ to be assigned to class S_k given the feature vector of x is represented as $P(S_k|\ \mathrm{x})$. In order to assign a class to the sample, conditional probabilities of classes must be compared pairwise. Finally, if

$$P(S_i|\ \mathrm{x}) > \ P(S_j|\ \mathrm{x})\ \forall j \neq i \tag{1}$$

in other words

$$P(x|\ S_i)P(S_i) > \ P(x|S_j)\ P\left(S_j\right)\ \forall j \neq i \tag{2}$$

then it is accepted that x belongs to the class S_i. In classifiers, assuming $(x|\ S_i) \approx \prod_{k=1}^{n} P(x_k|\ S_i)$, the formula (2) is modified to

$$\mathrm{P}(S_i)\prod_{k=1}^{n} P(x_k|\ S_i)\ > P\left(S_j\right)\prod_{k=1}^{n} P(x_k|\ S_j)\ \forall j \neq i \tag{3}$$

C4.5 is an algorithm, proposed by Quinlan (1992) to generate a decision tree. In decision trees, samples are split into different groups by a decision criterion (Mitchell, 1997). Classification goes on till all the samples in the group are assigned to the same class. The splitting criterion determines the type of the decision tree. There exist two main categories of decision trees: decision trees that base on entropy (e.g. ID3, *C4.5*), regression trees (e.g. CART). In *C4.5*, an improved version of ID3 that considers information gain, the splitting criterion is the gain ratio (Quinlan, 1996).

In k-nearest neighbour algorithm, each sample is assigned to the majority class of its k number of neighbours. In classification, k is determined empirically. If k=1, then the sample is assigned to the class of its nearest neighbour.

In logistic regression method (Cox, 1958), the variable z (known as logit) is provided to the learning system. The variable z is given as

$$z = b_1x_1 + \ b_2x_2 + \ b_3x_3 + \ldots + b_kx_k + c \tag{4}$$

where x_i represents the i^{th} feature and $\mathrm{b_i}$ is the weight assigned to the regarding feature. The classification result is determined by $f(z)$ value that is calculated as follows

$$f(z) = \frac{e^z}{1-e^z} \tag{5}$$

$f(z)$ is in range [0 1] and simply it shows the class that the sample is to be assigned. In our study, a modified logistic regression model with a ridge estimator, provided in WEKA tool (Hall et al., 2009) is employed. The regarding model bases on the algorithm proposed in Cessie, Houwelingen, and Society (1992) where further details on LR may be found.

Random Forest is a decision tree type algorithm where a set of decision trees is constructed and their common decision is considered. In the algorithm, instead of best performing subsets of features, random subsets are provided to the trees in order to have independent decisions and the decision trees are not pruned (Archer & Kimes, 2008) (Breiman, 2001). Random forest is stated to be faster and more resistant to over fitting compared to alternative approaches and the number of trees may be pre-determined (Breiman & Cutler, 2007). Further information on random forests may be found in Breiman (2001).

In binary classification, support vector machine primarily shows how to draw a line between two groups of instances that are on the same plane in order to classify them. In its most known version, a line is drawn between the instances of groups to be classified, which is the farthest possible to them to create a clear cut between different groups. In general, an SVM is a machine that constructs a hyper plane (or set of hyper planes) in a high or infinite dimensional space. These planes can be used for classification, regression or similar tasks (Shevade, Keerthi, Bhattacharyya, & Murthy, 2000). In our experiments, C-SVC classifier of LibSVM library (Chang & Lin, 2001) is called with default values from WEKA tool to classify samples in the data set. Briefly, in C-SVM (known as SVM Type 1) model, training involves the minimization of an error function that includes a capacity constant C subject to the constraints. The details on the mathematical background of C-SVC and its implementation in WEKA are given in Hsu, Chang, and Lin (2008), Chang and Lin (2001) and Koutroumbas and TheoDoridis (2003).

Multilayer perceptron is a neural network that uses backpropagation. In our experiments, default setting of WEKA for MP classifier that has one input and one output layer together with *number of features* + *number of classes*/2 number of hidden layers where sigmoid is the activation function is employed. For further details on MP classifier, readers may refer to Haykin (1994) and Gardner and Dorling (1998).

4.2. Filtering

The classification performance/efficiency of MWE features is determined by measuring the score generated by an attribute/feature evaluator.

Feature filtering in MWE identification is performed in two phases. In the first phase statistical and linguistic features are ranked in their own categories. In the second phase all MWE features (statistical features + linguistics features) are compared to each other regardless of the category. The experiments are performed by 5-fold cross validation. In this study, 4 different attribute evaluators are employed. These are

- Information gain
- Gain ratio
- Relief-F
- Chi-square

evaluators.

The notion of information gain is closely related to the entropy concept. Entropy is a measure of uncertainty in samples for a given feature. For example, if all the samples in data set hold a different value for a feature, the uncertainty reaches to its maximum value. Information gain is the reduction of uncertainty in samples based on a specific feature (Mitchell, 1997).This is why; as the information gain gets higher the uncertainty gets lower supporting the effective classification.

Table 10
The statistics on collection of corpora (*The statistics on Bilkent corpus is given eliminating the tags.).

Corpus	Corpus Size (tokens)	Vocabulary size (tokens)	Description
BilCol	39,545,399	739,380	BilCol2005 Turkish news corpus contains ~200.000 news, which are collected from five different Turkish news web sources throughout the year 2005.
Bilkent*	719,665	96,453	Bilkent corpus (Tur et al., 2003)is a Turkish corpus that is composed of news and articles published in the period of time that the corpus is built. The corpus is morphologically analysed and the sentence segmentation in corpus is done by a finite state machine (Tur et al., 2003). We employed the corpus that is re-organized by Dinçer (2004).
Ege	2493,213	232,420	The corpus is compiled in International Computing Institute, Ege University, to be used in natural language processing studies. It contains 875 texts that are classified in 9 different topics. The corpus has labels for class and sub-classes.
Leipzig	13,625,149	671,104	The corpus is a part of Leipzig corpus collection (Quasthoff et al., 2006). It contains 1.000.000 Turkish sentences that are collected from the newspapers in year 2005. The corpus is not annotated except the sentences.
Metu	1987,447	202,810	Metu (Say et al., 2002) is the corpus that contains texts in Turkish written after the year 1990. The corpus has the XCES2 tags.
Muder	638,546	82,144	Muder corpus is prepared in Muğla Sıtkı Koçman University (Dinçer, 2004). It includes 41,862 sentences collected texts from several sources. There exists to labelling in the corpus.
TOTAL	**59,009,419**	**2024,311**	

Information gain (IG) is calculated as follows

$$H(S) = -\sum_{i=1}^{n} p_i \ log_2 p_i \tag{6}$$

$$IG(S,A) = \ H(S) - H(S|A) \tag{7}$$

Given that A is the feature, H(S) is the entropy of class S. The term p_i in equation of entropy H(S) is the probability of *i*th class. In filtering, features are sorted in decreasing order of IG values.

The second information theoretic evaluator is the gain ratio. Gain ratio stands for the ratio of information gain to the entropy of the feature. Gain ratio is of the form

$$GR(S,A) = \ \frac{H(S) - H(S|A)}{H(A)} \tag{8}$$

where *H(A)* is the entropy of regarding feature regardless of class tags. *H(A)* represents the uncertainty among the samples considering the feature value independent of other features or the class tags. Similar to *IG* evaluator, the features are ranked based on the *GR* value. The most valuable feature is accepted to be the one that generates highest *GR* value.

The attribute evaluator Relief–F, firstly proposed by Kira and Rendell (1992), does not consider the notion of entropy. In this algorithm, in each iteration, a sample is taken from the data set and a feature vector is built. The closest neighbour in sample's own class and the opposing class(es) are determined by a distance metric such as Euclidean distance. The nearest instance in its own class is named as "near-hit" and the other(s) is "near-miss". The sample distances to the near-hit and near-miss are calculated. As the near-hit distance gets lower compared to the near-miss distance, the Relief-F value gets higher, meaning that the feature is accepted to have a higher classification power. Relief-F value varies in range [-1 1]. The attribute evaluator Relief–F in this study is the relief-F algorithm proposed by Kononenko (1994) in which there exists improvements such as the replacement of Euclidean distance to Manhattan distance and the use of absolute distance.

The chi-square (χ^2) statistics is an overall measure of the independence of two events *A* and *B*. *A* and *B* are defined to be independent if *P(AB)=P(A)P(B)* or equivalently, *P(A|B)=P(A)* and, *P(B|A)=P(B)*. In general, the chi-square statistic is of the form

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{9}$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency. More specifically in MWE feature selection, χ^2 is used to test the independence of occurrence of a specific class and a specific feature. The rationale of χ^2 feature selection is that if the two events are dependent, then the occurrence of a specific feature makes the occurrence of the class more likely (or less likely), so it should be helpful as a feature. In chi-square feature selection, all features are sorted in decreasing order of the chi-square value enabling the comparison of importance of features.

5. Experimental set-up

5.1. Data set

In MWE identification, an annotated MWE data set, which includes both positive and negative samples, is required to test the proposed methods. We merged 6 different Turkish corpora in order to build a collection of Turkish corpora that includes samples of different types of MWEs (e.g. names entities, idioms). The collection includes BilCol (Can et al., 2010) Bilkent (Tur, Hakkani-Tur, & Oflazer, 2003), Ege, Leipzig (Quasthoff, Richter, & Biemann, 2006), Metu (Say, Zeyrek, Oflazer, & Umut, 2002) and Muder (Dinçer, 2004) corpus.

Table 10 presents approximate corpus/vocabulary sizes and short descriptions of the regarding corpora.

In the study, we build out the MWE data set by using 4 different methods. Firstly, a group of frequency-based methods (occurrence frequency, point-wise mutual information, t-test and χ^2 test) are employed and the bigrams in MWE candidates are sorted by the methods as in Kumova Metin (2016). Top most 200 bigrams are selected from each sorted list and included in MWE data set. Secondly, in order to detect bigrams that hold the MWE-type linguistics properties, we defined a set of MWE patterns. The set includes 11 patterns (e.g. adverb+noun, noun+pronoun, noun+verb, adjective+noun) that are extracted from previous studies on Turkish (e.g. Özkan, 2010) and other languages (e.g. Augusto, Boos, Prestes, & Villavicencio, 2014; Evert & Kermes, 2003). We retrieved all bigrams from Bilkent corpus with a matching pattern and occurring more than or equal to 3 times in corpus. Third and fourth methods in MWE data set generation aim to detect MWE candidates that have idiomatic and term-like properties respectively. Thus a list of idioms and a list of terms are compiled from several web sources. The list of idioms includes 10,216 two-worded idioms and the list of terms has 47,805 two-worded terms. Ege, Bilkent and Leipzig corpus are merged to obtain a combined corpus and the bigrams in corpus is listed. For each two-worded idiom, the first word of the idiom is compared with the first words of the bigrams in the list. All the bigrams with matching first words are retrieved. In the retrieved list of bigrams, the bigram that has the highest frequency and a non-stop second word is inserted to the MWE data set. The same procedure to find the matching bigrams

Table 11
Content summary of MWE data set.

Methods	*MWE*	*non-MWE*	Total	Fleiss Kappa
Statistical	1165 (54.82%)	960 (45.18%)	2125	~0738
Linguistic	475 (30.59%)	1078 (69.41%)	1553	~0786
Idiomatic	893 (63.29%)	518 (36.71%)	1411	~0.678
Term-like	2315 (54.14%)	1961 (45.86%)	4276	~0.853
MWE Data Set	3946 (48.26%)	4230 (51.74%)	8176	

is followed for the term list and the candidates that hold term-like properties are inserted to the data set.

Table 11 depicts content summary of the MWE data set that is labelled using a rule-based procedure by 3-4 judges based on a guideline[2] provided by the researchers. In rows of Table 10, the numbers of MWEs and non-MWEs obtained by four methods (given in first column) are listed individually. The agreement among the judges is measured by Fleiss kappa value. As the Fleiss kappa values listed in Table 11 is observed, it may be stated that there exists a reliable agreement among judges. The last row of Table 11 provides the information on the final MWE data set.

5.2. Evaluation measures and set-up

The feature selection methods are evaluated simply by executing the classification methods once using the subset of features and once using the whole set of features. It is accepted that feature selection method succeeds if the proposed subset increases the classification performance. The performance is measured commonly by precision, recall and F1-values. In classification evaluation, precision (P), recall (R) and F1-metric are defined as

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{10}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{11}$$

$$\mathrm{F1} = 2\frac{P \cdot R}{P + R} \tag{12}$$

where *TP* is the number of candidates that are both annotated and expected (by the classifier) to be in same class, *FP* is the number of candidates that are expected to belong to one class but is annotated in the other class. And *FN* stands for the number of candidates that belong to one class but assigned to the opposite by the classifier.

The procedures to follow in wrappers and the filtering methods may be different since the outcomes of the methods may vary. In this study, the wrappers' performance is measured by comparing the performance of proposed subsets to the performance of whole set of features by application of Naïve Bayes and *C4.5* classifiers.

The evaluation of filtering requires two pre-processing tasks to be performed. The tasks are

Task 1. The feature rankings of different attribute evaluators must be examined to obtain an agreed subset of features. In this task, the aim is building an optimal set of features that includes the features that are commonly offered by methods.

Task 2. The classification performance of feature subset must be evaluated/compared to the whole set of features by classification methods.

[2] Simply in the guideline, online sources such as dictionaries (e.g. Tureng) and encyclopaedias (e.g. Wikipedia) that must be used in annotation are listed and exceptional cases are defined. The guideline may be provided on demand.

In this study, we propose two approaches to determine the features that different attribute evaluators agree on. First approach is grounding to Kendall's *W* analysis (Kendall & Smith, 1939) Kendall's *W* analysis is a non-parametric test used for assessing agreement among raters. Kendall's *W* ranges from 0 (no agreement) to 1 (complete agreement). In filtering, different attribute evaluators are accepted to be the raters that provide a ranking of a group of MWE identification features.

In the analysis, if r_{ij} is the rank given to the *i*th feature by *j*th attribute evaluator (judge) in a set of *k* features, then the summation of the ranks of the *i*th feature is obtained as $R_i = \sum_{j=1}^{m} r_{ij}$. The mean value of the total ranks is

$$\bar{R} = \frac{1}{k}\sum_{1}^{k} R_i \tag{13}$$

and the sum of squared deviations, S, is defined as

$$S = \sum_{1}^{k} \left(R_i - \bar{R}\right)^2 \tag{14}$$

and then Kendall's *W* is given as

$$W = \frac{12S}{m^2\left(k^3 - k\right)} \tag{15}$$

If *W* value is close to maximum value, then it may be stated that attribute evaluators commonly agree on the rankings, in other words they provided almost same rankings to the features. Thus the rankings are such reliable that the features holding the lower ranks (e.g. first/top most feature) in lists of evaluators may be selected to construct the feature subset. In this approach, a threshold rank value must be determined to limit number of features in the subset.

The second approach in determination on agreed features bases on the assumption that the features that commonly have higher ranks (e.g. last feature) in the lists, do not perform well in MWE classification. We propose to eliminate such failing features following the procedure defined as follows

1. A threshold rank (*tr*) is calculated as $tr = N^*c$ where *N* represents the total number of features and *c* is a constant in range [0 1] that is determined empirically. *tr* is actually the maximum rank value that a feature in MWE subset may have.
2. For each feature *M*, the ranks in lists of different evaluators are observed and stored in vector *rank(M)*.
3. If all the ranks in *rank(M)* is lower than the threshold rank *tr* the feature is inserted to the MWE feature set.

For example, in a set of 10 features and three evaluators, if *c=0.3*, threshold rank is calculated as *tr=10*0.3=3* meaning that the MWE feature set will include the features that are the first, second and/or third top feature in the lists of three evaluators. Assuming that the feature M_1 and M_2 hold the ranks 3, 2, 2 and 2, 7, 3 respectively, we obtain the vectors $rank(M_1)$=[3 2 2] and $rank(M_2)$=[2 7 3]. As the values in vectors are compared individually with *tr=3*, it is observed that the feature M_1 must be inserted the MWE feature and the feature M_2 must be eliminated since $7 \not\leq 3$ in $rank(M_2)$ vector.

In second task of filtering evaluation, the proposed subsets of features are utilized in Naïve Bayes and *C4.5* methods and the performance is observed as it is done in wrappers.

6. Experimental results

To demonstrate the effectiveness of MWE feature selection, we conducted two sets of experiments using different feature selection approaches; filtering and wrappers; individually. In the experiments WEKA (Hall et al., 2009) tool is used to observe the perfor-

Table 12
Wrapper results – statistical features.

NB		C4.5		K-nn (k=1)		K-nn (k=5)		K-nn (k=10)		LR		RF		SVM		MP	
f	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*
5	*BBV*	5	*NUP*	4	*WFV*	5	*NUP*	5	*NUP*	5	*NUP*	4	*WFV*	5	*NUP*	5	*WBV*
5	*BFV*	5	*Fager*	3	*PS*	5	*Fager*	5	*Fager*	5	*BBV*	3	*PS*	5	*Fager*	5	*WFV*
5	*MD*	5	*JP*	2	*NUP*	4	*BWFV*	5	*MD*	5	*BFV*	2	*NUP*	5	*PS*	5	*BBV*
0	*NUP*	4	*WFV*	2	*BBV*	4	*MD*	4	*Mount*	5	*SK*	2	*BBV*	2	*BFV*	5	*BFV*
0	*BWBW*	4	*BBV*	2	*Fager*	3	*WFV*	4	*PS*	5	*MD*	2	*Fager*	1	*WFV*	5	*MD*
0	*BWFV*	4	*SSS*	1	*WBV*	3	*BBV*	2	*WFV*	4	*Fager*	1	*WBV*	1	*FK*	4	*NUP*
0	*WBV*	4	*CP*	1	*BFV*	3	*BFV*	2	*BBV*	4	*PS*	1	*BFV*	0	*BWBW*	4	*BWBW*
0	*WFV*	3	*BWFV*	1	*CP*	3	*PS*	2	*BFV*	4	*PMI*	1	*CP*	0	*BWFV*	4	*BWFV*
0	*Fager*	3	*BFV*	1	*PMI*	3	*DK*	2	*LFMD*	3	*WBV*	1	*PMI*	0	*WBV*	4	*Fager*
0	*Mount*	3	*Mount*	0	*BWBW*	3	*SK*	2	*PMI*	3	*WFV*	0	*BWBW*	0	*BBV*	4	*Mount*
0	*SSS*	3	*DK*	0	*BWFV*	3	*FK*	1	*BWFV*	3	*FK*	0	*BWFV*	0	*Mount*	4	*PS*
0	*J*	3	*SK*	0	*Mount*	3	*JP*	1	*SSS*	3	*RCP*	0	*Mount*	0	*SSS*	4	*SK*
0	*PS*	3	*Rcost*	0	*SSS*	2	*Mount*	1	*J*	2	*BWBW*	0	*SSS*	0	*J*	4	*Scost*
0	*DK*	3	*Ucost*	0	*J*	2	*SSS*	1	*BB*	2	*BWFV*	0	*J*	0	*DK*	4	*NE*
0	*Simps*	3	*Scost*	0	*DK*	2	*J*	1	*SK*	2	*J*	0	*DK*	0	*Simps*	4	*RCP*
0	*BB*	3	*LFMD*	0	*Simps*	2	*Rcost*	1	*NE*	2	*DK*	0	*Simps*	0	*BB*	4	*CP*
0	*SK*	3	*MD*	0	*BB*	2	*LFMD*	0	*BWBW*	2	*Ucost*	0	*BB*	0	*SK*	4	*JP*
0	*FK*	2	*BWBW*	0	*SK*	2	*PMI*	0	*WBV*	2	*LFMD*	0	*SK*	0	*Rcost*	4	*PMI*
0	*Rcost*	2	*WBV*	0	*FK*	1	*WBV*	0	*DK*	2	*CP*	0	*FK*	0	*Ucost*	3	*DK*
0	*Ucost*	2	*J*	0	*Rcost*	1	*CP*	0	*Simps*	1	*Mount*	0	*Rcost*	0	*Scost*	3	*Simps*
0	*Scost*	2	*PS*	0	*Ucost*	0	*BWBW*	0	*FK*	1	*SSS*	0	*Ucost*	0	*NE*	3	*Rcost*
0	*NE*	2	*BB*	0	*Scost*	0	*Simps*	0	*Rcost*	1	*BB*	0	*Scost*	0	*LFMD*	3	*Ucost*
0	*LFMD*	2	*NE*	0	*NE*	0	*BB*	0	*Ucost*	1	*Rcost*	0	*NE*	0	*MD*	3	*LFMD*
0	*RCP*	1	*Simps*	0	*LFMD*	0	*Ucost*	0	*Scost*	1	*Scost*	0	*LFMD*	0	*RCP*	2	*SSS*
0	*CP*	1	*FK*	0	*MD*	0	*Scost*	0	*RCP*	1	*JP*	0	*MD*	0	*CP*	2	*J*
0	*JP*	1	*RCP*	0	*RCP*	0	*NE*	0	*CP*	0	*Simps*	0	*RCP*	0	*JP*	0	*BB*
0	*PMI*	1	*PMI*	0	*JP*	0	*RCP*	0	*JP*	0	*NE*	0	*JP*	0	*PMI*	0	*FK*

Table 13
Wrapper results – linguistic features.

NB		C4.5		K-nn (k=1)		K-nn (k=5)		K-nn (k=10)		LR		RF		SVM		MP	
f	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*
5	*FF*	5	*PVSFn*	5	*PVSFn*	5	*PVSFn*	5	*PVSFn*	5	*PVSFn*	5	*NEW*	5	*DF*	5	*NEW*
5	*SS_h*	5	*PVSFm*	5	*PVSFm*	5	*PVSFm*	5	*PVSFm*	5	*SS_h*	5	*SS_h*	5	*PVSFm*	5	*SS_h*
4	*BHK*	5	*OV_a*	5	*OV_h*	5	*OV_h*	5	*FF*	5	*BHK*	5	*BHK*	4	*BHK*	5	*BHK*
3	*OV_a*	5	*HF*	5	*OV_a*	5	*OV_a*	5	*HF*	5	*DF*	5	*HF*	2	*SS_h*	5	*DF*
2	*PVSFm*	5	*BHK*	5	*HF*	5	*FF*	5	*BHK*	3	*OV_a*	5	*OV_a*	0	*NEW*	4	*OV_h*
2	*OV_h*	5	*SS_h*	5	*BHK*	5	*BHK*	5	*SS_h*	3	*OV_h*	5	*OV_h*	0	*SS_n*	3	*SS_n*
2	*NEW*	5	*NEW*	5	*SS_h*	5	*SS_h*	5	*NEW*	3	*PVSFm*	5	*PVSFm*	0	*HF*	3	*PVSFm*
1	*SS_n*	4	*OV_h*	5	*NEW*	5	*SS_n*	4	*OV_a*	2	*NEW*	5	*PVSFn*	0	*OV_a*	3	*PVSFn*
0	*PVSFn*	2	*FF*	2	*FF*	5	*NEW*	2	*OV_h*	0	*SS_N*	3	*DF*	0	*OV_h*	1	*HF*
0	*HF*	0	*SS_n*	0	*SS_n*	4	*HF*	1	*SS_n*	0	*HF*	0	*SS_n*	0	*PVSFn*	1	*OV_a*

Table 14
Wrapper results – statistical+linguistic features.

NB		C4.5		K-nn (k=1)		K-nn (k=5)		K-nn (k=10)		LR		RF		SVM		MP	
f	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*	*f*	*feature*
5	*MD*	5	*PS*	4	*WFV*	5	*Fager*	5	*MD*	5	*DF*	4	*SS_h*	5	*NUP*	5	*NEW*
5	*BFV*	5	*Fager*	4	*SS_h*	5	*FF*	5	*Fager*	5	*BBV*	4	*WFV*	5	*PS*	5	*BHK*
5	*BBV*	5	*BBV*	3	*PS*	4	*NUP*	5	*NUP*	5	*BFV*	3	*PS*	4	*PVSFm*	5	*DF*
5	*FF*	5	*NUP*	2	*Fager*	3	*JP*	5	*FF*	5	*MD*	2	*NUP*	4	*Fager*	5	*PVSFn*
2	*BHK*	5	*PVSFn*	2	*BBV*	3	*PS*	4	*OV_h*	4	*SS_h*	2	*BBV*	3	*BHK*	5	*NUP*
2	*NEW*	5	*HF*	2	*NUP*	3	*BWFV*	3	*FK*	4	*Fager*	2	*Fager*	2	*BFV*	5	*WFV*
1	*NE*	4	*SK*	1	*PMI*	3	*PVSFm*	3	*J*	4	*FK*	1	*WBV*	1	*WFV*	5	*BBV*
1	*Ucost*	4	*Mount*	1	*CP*	3	*OV_h*	3	*SSS*	3	*OV_a*	1	*BFV*	1	*FK*	5	*BFV*
1	*BB*	4	*BFV*	1	*BFV*	3	*HF*	3	*BFV*	3	*OV_h*	1	*CP*	1	*JP*	5	*Fager*
1	*NUP*	4	*OV_a*	1	*WBV*	3	*SS_h*	3	*BBV*	3	*PVSFn*	1	*PMI*	0	*NEW*	5	*PMI*
1	*PVSFn*	4	*NEW*	0	*JP*	2	*SSS*	3	*NEW*	3	*PS*	0	*NEW*	0	*SS_n*	4	*SS_h*
1	*OV_h*	3	*PMI*	0	*RCP*	2	*Mount*	2	*JP*	3	*Simp*	0	*SS_n*	0	*SS_h*	4	*PVSFm*
1	*OV_a*	3	*JP*	0	*MD*	2	*WFV*	2	*LFMD*	3	*RCP*	0	*BHK*	0	*HF*	4	*BWBW*
0	*PMI*	3	*MD*	0	*LFMD*	2	*NEW*	2	*NE*	3	*PMI*	0	*HF*	0	*DF*	4	*WBV*
0	*JP*	3	*NE*	0	*NE*	1	*PMI*	2	*BB*	2	*BHK*	0	*DF*	0	*OV_a*	4	*DK*
0	*CP*	3	*Scost*	0	*Scost*	1	*MD*	2	*DK*	2	*PVSFm*	0	*OV_a*	0	*OV_h*	4	*FK*
0	*RCP*	3	*Ucost*	0	*Ucost*	1	*LFMD*	2	*Mount*	2	*BWBW*	0	*OV_h*	0	*PVSFn*	4	*LFMD*
0	*LFMD*	3	*Rcost*	0	*Rcost*	1	*NE*	2	*WFV*	2	*WBV*	0	*PVSFm*	0	*BWBW*	4	*MD*
0	*Scost*	3	*WFV*	0	*FK*	1	*Scost*	2	*PVSFn*	2	*SSS*	0	*PVSFn*	0	*BWFV*	3	*OV_h*
0	*Rcost*	3	*BHK*	0	*SK*	1	*Rcost*	2	*PVSFm*	2	*DK*	0	*BWBW*	0	*WBV*	3	*Mount*
0	*FK*	2	*CP*	0	*BB*	1	*FK*	2	*OV_a*	2	*CP*	0	*BWFV*	0	*BBV*	3	*SK*
0	*SK*	2	*FK*	0	*Simps*	1	*SK*	2	*HF*	2	*JP*	0	*Mount*	0	*Mount*	3	*Rcost*
0	*Simps*	2	*BB*	0	*DK*	1	*Simps*	2	*BHK*	1	*HF*	0	*SSS*	0	*SSS*	3	*Ucost*
0	*DK*	2	*Simps*	0	*J*	1	*DK*	2	*SS_h*	1	*NUP*	0	*J*	0	*J*	3	*RCP*
0	*PS*	2	*WBV*	0	*SSS*	1	*J*	1	*PS*	1	*WFV*	0	*DK*	0	*DK*	3	*CP*
0	*J*	2	*PVSFm*	0	*Mount*	1	*BFV*	0	*PMI*	1	*Rcost*	0	*Simps*	0	*Simps*	2	*HF*
0	*SSS*	2	*OV_h*	0	*BWFV*	1	*BBV*	0	*CP*	1	*Ucost*	0	*BB*	0	*BB*	2	*BWFV*
0	*Mount*	2	*FF*	0	*BWBW*	1	*PVSFn*	0	*RCP*	1	*Scost*	0	*SK*	0	*SK*	2	*SSS*
0	*Fager*	1	*SSS*	0	*PVSFn*	1	*BHK*	0	*Scost*	1	*LFMD*	0	*FK*	0	*Rcost*	2	*J*
0	*WFV*	1	*BWFV*	0	*PVSFm*	1	*SS_n*	0	*Ucost*	0	*NEW*	0	*Rcost*	0	*Ucost*	2	*PS*
0	*WBV*	1	*BWBW*	0	*OV_h*	0	*CP*	0	*Rcost*	0	*SS_n*	0	*Ucost*	0	*Scost*	2	*Simps*
0	*BWFV*	0	*RCP*	0	*OV_a*	0	*RCP*	0	*SK*	0	*BWFV*	0	*Scost*	0	*NE*	2	*Scost*
0	*BWBW*	0	*LFMD*	0	*FF*	0	*Ucost*	0	*Simps*	0	*Mount*	0	*NE*	0	*LFMD*	2	*NE*
0	*PVSFm*	0	*DK*	0	*HF*	0	*BB*	0	*WBV*	0	*J*	0	*LFMD*	0	*MD*	1	*SS_n*
0	*HF*	0	*J*	0	*BHK*	0	*WBV*	0	*BWFV*	0	*BB*	0	*MD*	0	*RCP*	1	*OV_a*
0	*SS_h*	0	*SS_h*	0	*SS_n*	0	*BWBW*	0	*BWBW*	0	*SK*	0	*RCP*	0	*CP*	1	*BB*
0	*SS_n*	0	*SS_n*	0	*NEW*	0	*OV_a*	0	*SS_n*	0	*NE*	0	*JP*	0	*PMI*	1	*JP*

mance of 27 statistical and 10 linguistics features in MWE identification.

In first set of experiments, wrappers are employed with MWE features. Tables 12–14 give the results with statistical, linguistic and statistical+linguistic features respectively. All the experiments are performed with 5-fold cross-validation to examine whether the features are stable under changes in input data. The *f* columns in Tables 12–14 express the number of folds that the regarding feature is observed. It may be stated that the higher the number of the folds, the more effective (successful) is the feature in classification. In tables, the features that are observed at least in 3 folds are highlighted. For example, in the set of statistical features, there exists 25, 19 and 10 features that have $f \geq 3$, $f \geq 4$ and $f=5$ respectively.

In linguistic features, it is observed that in k-nearest neighbour method, all the features are effective in classification when k=5. In

Table 15
Evaluation results of wrappers.

Feature type	Feature Subsets	Subset Size	C4.5				Naive Bayes			
			P	R	F1	F1 increase (%)	P	R	F1	F1 increase (%)
Statistical features	$f \geq 0$	27	0.709	0.706	0.706	–	0.642	0.564	0.637	–
	$f \geq 3$	25	0.706	0.703	0.703	−0,42	0.649	0.649	**0.648**	1,73
	$f \geq 4$	19	0.712	0.709	0.708	0,28	0.651	0.648	**0.648**	1,73
	$f=5$	10	0.714	0.711	**0.711**	0,71	0.696	0.640	0.618	−2,98
Linguistic features	*All features*	10	0.647	0.647	**0.647**	–	0.564	0.532	0.424	–
	All features excluding SS_n	9	0.647	0.647	**0.647**	0.00	0.560	0.531	**0.425**	0.24
Statistical+Linguistic features	$f \geq 0$	37	0.726	0.723	0.723	–	0.639	0.628	0.614	–
	$f \geq 3$	35	0.725	0.722	0.722	−0,14	0.646	0.638	0.630	2,61
	$f \geq 4$	24	0.727	0.725	0.725	0,28	0.674	0.674	0.673	9,61
	$f=5$	14	0.734	0.731	**0.731**	1,11	0.697	0.687	**0.685**	11,56

addition, *SS_n* feature is to be excluded from the feature set since it is never observed or observed in a single fold, except multilayer perceptron and k-nearest neighbour algorithm with k=5. In wrapping of statistical+linguistic features, it is examined that 35 features are observed in at least 3-folds, 24 is in at least 4-folds and 14 in 5-folds.

C4.5 and Naïve Bayes classification results of the feature subsets are given in Table 15. In Table 15, $f \geq n$ subset expresses the set of features that are observed at least in n number of folds in all wrappers. Simply, the results in $f \geq 0$ rows demonstrate the performance values when all features are employed. These results are actually the values that are obtained when no feature selection is performed.

In Table 15, the highest *F1* values are given in bold for each feature type. For example, when only statistical features are considered, the highest *F1* value for classifier *C4.5*; 0.711 is obtained by the $f=5$ feature subset. *F1 increase (%)* columns represent the change (in percentages) in *F1* value compared to *F1* values when the whole set of features are used (no feature selection is performed) in classification. For instance, when all statistical and linguistic features are used in *Naïve Bayes* classification, *F1* value is 0.614 and the increase is measured as 11.56% when the feature set is replaced with $f=5$ subset ($F1$=0.685). Overall wrapper results show that

1. The highest *F1* values are obtained with $f=5$ subsets as expected, except Naïve Bayes classification with statistical features. When the detailed evaluation results are observed in order to extract the reason behind this exceptional case, it is examined that though the Naïve Bayes classifier with f=5 subsets returned a relatively high *F1* value for MWE pairs ($F1_{MWE}$=0.705), the resulting *F1* for non-MWE pairs was lower ($F1_{nonMWE}$=0.537) due to low recall values for non-MWE pairs. As a result, it may be stated that, the only reason for this unexpected decrease in *F1* value is the unsuccessful recall of non-MWE pairs.
2. The average increase in *F1* value is 0.26% for *C4.5* and 3.5% for Naive Bayes classifier. Hence, we believe that though the feature selection provides a remarkable improvement in Naive Bayes classifier's performance (especially when both statistical and linguistic features are employed) on average, it is not such effective if the classifier is a decision tree type *C4.5* classifier that sorts the features and uses them one by one in decreasing order of their classification power as default.
3. For both classifiers, the highest evaluation scores (*F1*) are reached when statistical and linguistic features are considered together.

The second group of experiments contain the evaluation tests on filtering methods. Tables 16 and 17 include the classification results of filtering methods on sets of statistical and linguistic features respectively. The tables present the ranks given to the features by each of four attribute evaluators. The features are listed in decreasing order of the average ranks given in the last columns of the tables. For example, *NUP* is the first/best feature with the average rank 3.5 and *Mount* is the worst with the average rank 25.

Table 18 presents the ranks assigned to the features when they are considered all together. It is observed from the results that sta-

Table 16
Statistical features – filtering results.

Feature	Feature Type	Rank				
		Information Gain	Gain ratio	Relief-F	Chi-Square	Average
NUP	Term-hood	1	2	10	1	3.50
MD	Association	2	6	9	2	4.75
LFMD	Association	4	14	3	4	6.25
BWBW	Term-hood	5	1	13	8	6.75
DK	Association	3	9	18	3	8.25
PMI	Association	12	3	8	13	9.00
NE	Association	6	13	21	6	11.50
J	Association	7	15	23	5	12.50
Rcost	Association	13	4	24	11	13.00
SK	Association	11	12	17	12	13.00
WBV	Term-hood	20	7	5	20	13.00
SSS	Association	9	11	25	9	13.50
FK	Association	8	16	27	7	14.50
BB	Association	10	17	22	10	14.75
RCP	Association	16	10	16	17	14.75
BFV	Term-hood	25	8	1	25	14.75
Simps	Association	14	18	14	14	15.00
Scost	Association	15	19	11	15	15.00
Fager	Association	18	22	7	18	16.25
BWFV	Term-hood	17	21	12	16	16.50
BBV	Term-hood	24	20	2	24	17.50
WFV	Term-hood	21	23	6	21	17.75
PS	Association	22	5	26	22	18.75
CP	Association	19	24	15	19	19.25
Ucost	Association	26	25	4	26	20.25
JP	Association	23	26	20	23	23.00
Mount	Association	27	27	19	27	25.00

Table 17
Linguistic features – filtering results.

Feature	Rank				
	Information Gain	Gain ratio	Relief-F	Chi-Square	Average
PVSFn	1	2	2	1	1.50
BHK	2	4	1	2	2.25
PVSFm	4	3	5	4	4.00
FF	3	1	10	3	4.25
SS_h	5	5	6	5	5.25
SS_n	6	6	7	6	6.25
NEW	8	7	4	8	6.75
OV_h	9	9	3	9	7.50
OV_a	7	8	9	7	7.75
HF	10	10	8	10	9.50

Table 18
Statistical+Linguistic features – filtering results.

Feature	Type	Rank				
		Information Gain	*Gain ratio*	*Relief-F*	*Chi-Square*	*Average*
NUP	Statistical	1	2	14	1	4.50
MD	Statistical	2	7	11	2	5.50
LFMD	Statistical	4	15	7	4	7.50
BWBW	Statistical	5	1	16	8	7.50
PMI	Statistical	12	3	8	13	9.00
DK	Statistical	3	9	22	3	9.25
NE	Statistical	6	13	23	6	12.00
WBV	Statistical	20	6	5	20	12.75
J	Statistical	7	14	28	5	13.50
SK	Statistical	11	12	21	12	14.00
Rcost	Statistical	13	4	31	11	14.75
SSS	Statistical	9	11	33	9	15.50
BB	Statistical	10	18	24	10	15.50
Scost	Statistical	16	19	12	15	15.50
RCP	Statistical	15	10	20	17	15.50
Simps	Statistical	14	17	18	14	15.75
FK	Statistical	8	16	37	7	17.00
Fager	Statistical	18	23	9	18	17.00
BFV	Statistical	29	8	3	29	17.25
BWFV	Statistical	17	22	17	16	18.00
WFV	Statistical	21	24	6	21	18.00
FF	Linguistic	24	20	4	24	18.00
BHK	Linguistic	23	29	1	23	19.00
BBV	Statistical	27	21	2	27	19.25
CP	Statistical	19	25	19	19	20.50
PVSFn	Linguistic	22	26	15	22	21.25
PS	Statistical	25	5	35	25	22.50
Ucost	Statistical	30	28	10	30	24.50
JP	Statistical	26	30	29	26	27.75
PVSFm	Linguistic	28	27	30	28	28.25
NEW	Linguistic	35	35	13	35	29.50
SS_n	Linguistic	32	32	26	32	30.50
SS_h	Linguistic	33	31	27	33	31.00
Mount	Statistical	31	33	34	31	32.25
OV_h	Linguistic	36	36	25	36	33.25
OV_a	Linguistic	34	34	32	34	33.50
HF	Linguistic	37	37	36	37	36.75

Table 19
Kendall-*W* results.

Features	*Kendall-W*	k	m	*p*
Statistical	*0.434*	*27*	*4*	*0.4895*
Linguistic	*0.698*	*10*	*4*	*0.0105*
Statistical+Linguistic	*0.614*	*37*	*4*	*0.0006*

tistical features perform better compared to linguistic features in information gain, gain ratio and chi-square evaluators.

The results of Kendall-*W* analysis that determines commonly agreed features are given in Table 19. Kendall-*W* value of statistical features is calculated as W=0.439 meaning that evaluators disagree on the ranks of statistical features. When the analysis results are examined for linguistic and/or statistical+linguistic features, it is observed that though *W* values are high, regarding *p* values are greater than 0.001 meaning that these *W* values are not statistically significant. Consequently, it can be stated that Kendall-*W* analysis and the approach to build a subset of features based on Kendall-*W* analysis cannot be used in MWE feature selection in our data set.

In the second approach, we determine commonly agreed features based on the assumption that the features that commonly have lower ranks in the lists, succeed in MWE classification and vice versa. In this empirical approach, it is required to split the feature lists in two; succeeding features and failing features; by a threshold score. For any feature, if it has a rank that is lower than the threshold score in lists of all evaluators, it is accepted to be in succeeding set otherwise it is included in the failing features set to be eliminated from the MWE feature set. Threshold value (*th*) is defined as $th = N^{*}c$ where *N* is the total number of features and *c* is a variable that takes the values in range [*0.1 0.9*] with *0.1* increments. In the experiments, we initially assign the minimum value to the *c*, and than increase it by *0.1* till the maximum value *0.9*. This procedure generated 9 different threshold scores. It is observed that for at least 6 of threshold scores

- *Mount* and *JP* in statistical feature set,
- *HF* and *OV_a* in linguistics feature set
- *Mount, JP, HF, OV_a, OV_h, SS_n, SS_h* and *PVSFm* in statistical+linguistic feature set

are assigned to the failing features set. Therefore, they are accepted to be the candidates that will be eliminated from the MWE feature sets. The performance of the sets containing the whole and the eliminated (reduced) set of features are measured by utilizing Naive Bayes and *C4.5* classifiers. Table 20 presents average precision (*P*), recall (*R*) and *F1* results of 5-fold cross-validated classification. For each set of features, the highest *F1* results are given in bold, in Table 20. For instance, in the set of statistical features, the highest *F1* is obtained when the feature *JP* is excluded from the feature set in *C4.5* classifier.

For each feature, we examined whether the elimination of the feature decreases *F1* value of classification. In case of a decrease, we decide not excluding the regarding feature. If the elimination of feature increases or does not change *F1* value then it is accepted to be a candidate in elimination. Finally, all elimination candidates are excluded from the set.

In our experiments, examining the results individually for each candidate to be eliminated, it may stated that

1. Considering the set of statistical features, excluding *Mount* and/or *JP* from the feature set, we obtained an increase in *F1* value for both classifiers, thus these features are decided to be removed from the feature set.
2. Considering the set of linguistic features, since elimination of *HF* results with an increase in *F1* values for Naïve Bayes (*0.516*>*0.424*) and *F1* value does not change in *C4.5* classifier, *HF* feature is decided to be excluded from the feature set.
3. Considering the set of linguistic features, the removal of *OV_a* decreases the *F1* values for both classifiers. In addition, when both *OV_a* and *HF* are removed, it is observed that *F1* values are lower compared to the case where only *HF* is removed from the set. As a result, it is decided that *OV_a* must stay in the feature set.
4. In statistical+linguistic feature set, it is observed that the elimination of *OV_a, JP* and *OV_h* features decreases the performance in one of the classifiers. Therefore, they must be used in MWE feature set. On the other hand, when *Mount, HF, SS_n, SS_h* and *PVSFm* features are removed from the set, it is examined that classifiers perform better or the performance do not change.
5. By the overall evaluation of the results, it may be stated that when statistical and linguistic features are employed together in classification, both classifiers generate the highest *F1* scores.

There are four remarkable outcomes of the experiments. Firstly, it is observed that for both wrappers and filters, the highest evaluation scores are obtained when statistical and linguistics features are employed together. Therefore, it may be stated that when statistical and linguistic features are employed together, it has a positive impact on MWE recognition when the regarding problem is accepted as a classification task. In other words, though linguistic features are observed to be less successful in classification compared to the statistical features, when linguistic and statistical features are employed together, linguistic features contribute to the performance. Secondly, comparing the proposed sets of features

Table 20
Filtering–classification results of threshold based approach.

Features		C4.5				Naive Bayes			
		P	*R*	*F1*	*F1 Increase (%)*	*P*	*R*	*F1*	*F1 Increase (%)*
Statistical	*All*	0.709	0.706	0.706	–	0.642	0.564	0.637	–
	Excluding Mount	0.710	0.708	0.708	0.28	0.644	0.643	0.640	0.47
	Excluding JP	0.715	0.709	**0.709**	0.42	0.644	0.643	0.641	0.63
	Excluding Mount+JP	0.714	0.709	0.708	0.28	0.647	0.647	**0.646**	1.41
Linguistic	*All*	0.647	0.647	**0.647**	–	0.564	0.532	0.424	–
	Excluding HF	0.647	0.647	**0.647**	0.00	0.552	0.548	**0.516**	21.70
	Excluding OV_a	0.646	0.646	0.645	−0.31	0.565	0.532	0.423	−0.24
	Excluding HF +OV_a	0.642	0.643	0.642	−0.77	0.546	0.541	0.501	18.16
Statistical+Linguistic	*All*	0.726	0.723	0.723	–	0.639	0.628	0.614	–
	Excluding HF	0.729	0.726	0.726	0.41	0.651	0.648	0.644	4.89
	Excluding OV_a	0.723	0.720	0.720	−0.41	0.640	0.627	0.614	0.00
	Excluding Mount	0.727	0.724	0.724	0.14	0.642	0.631	0.618	0.65
	Excluding JP	0.725	0.722	0.722	−0.14	0.640	0.631	0.619	0.81
	Excluding OV_h	0.726	0.723	0.723	0.00	0.637	0.624	0.610	−0.65
	Excluding SS_n	0.726	0.723	0.723	0.00	0.642	0.631	0.619	0.81
	Excluding SS_h	0.726	0.723	0.723	0.00	0.642	0.631	0.619	0.81
	Excluding PVSFm	0.726	0.723	0.723	0.00	0.641	0.630	0.617	0.49
	Excluding Mount, JP, HF, OV_a, OV_h, SS_n, SS_h, PVSFm	0.732	0.731	**0.731**	1.11	0.663	0.663	**0.662**	7.82

that are most effective in two approaches (given in Tables 15 and 20), it is examined that the performance of wrapping that employs only 14 of 37 features reaches to the performance of filtering with 29 features. When the features of two best performing settings are compared it is examined that *BBV, BFV, Fager, FF, MD, NUP, PS, PVSFn, WFV, PMI, NEW, FF* and *BHK* features (a total of 13 features) are commonly employed in both best feature subsets. Thirdly, even though feature selection requires a number of operations to be performed prior to classification, the experimental outcomes implied that it has a positive impact in terms of classification performance in almost all of the experiments. Lastly, it is examined clearly that decreasing the number of features in classification brings many advantages such as reduced overall time required for MWE recognition, facilitating data visualization and data understanding, owing the potential to avoid over fitting as mentioned in Gui et al. (2016)

7. Conclusion

The feature–based recognition is the major approach in determining MWEs in intelligent natural language systems. There exists a wide range of learning methods proposed to identify MWEs where a large number of features are employed. As the number of features increase, both the training time and the complexity in learning models gets higher. In this research, we begged the important question “Is it possible to increase or provide at least the same performance in MWE recognition by reducing the feature set size in order to overcome the disadvantages of employing a wide range of features?”. In this paper we presented our effort and proposed a procedure to select the subset of features that succeed in identifying multiword expressions. We employed well-known methods of feature selection; filtering and wrapping; using different evaluators/learning algorithms and compared the results in order to decide on a feature subset that is commonly agreed by different methods.

The comparison of feature selection methods is performed in two stages. Firstly, each selection method is assessed internally. In evaluation of subsets offered by wrappers, we assume that the features that are observed in a high number of folds for different methods are to be the most successful features. This assumption is tested on the data set in a 5-fold basis. The experiments showed that using the features that are observed in all 5-fold sets of different methods returns the highest performance results in almost all settings, as expected. The result of filtering method is a list of features sorted according to their individual performances in classification. In order to decide on a subset of features by filtering, we applied firstly Kendall-*W* test and secondly defined the threshold score empirically. It is observed that empirical approach succeeds in defining the features to be excluded from the feature set.

The experiments revealed that best wrapper and filtering subsets commonly contain 13 features (*BBV, BFV, Fager, FF, MD, NUP, PS, PVSFn, WFV, PMI, NEW, FF* and *BHK*) in our Turkish data set. These 13 features are of a total of 14 features in best wrapper subset. We believe that these results may be used in further Turkish MWE recognition studies. It is observed that the highest classification performance in MWE recognition is obtained by the best feature subset that is offered either by wrapper or filtering method in different experimental settings. Hereby, it may be stated that compared to the state of art methods where no feature selection is performed, it is possible to reach to higher evaluation results when features are subject to a prior selection process.

To conclude, even though the proposed feature selection procedure adds a number of prior operations to the MWE recognition system, it is demonstrated that it improves the whole system in terms of classification performance. We believe that it also contributes by reducing the measurement and storage requirements.

We have two main directions for future work. Firstly, we will run similar tests to select best performing subsets of features on different languages in order to examine if there exists any common features that reside on best subsets. Secondly, we plan to enlarge our data sets with MWEs of three words in order to observe whether the feature subset varies when number of words change in data set.

Acknowledgment

This work is carried under the grant of TÜBİTAK – The Scientific and Technological Research Council of Turkey to Project No: 115E469, Identification of Multi-word Expressions in Turkish Texts.

References

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis, 52*(4), 2249–2260. http://doi.org/10.1016/j.csda.2007.08.015.

Augusto, R., Boos, S., Prestes, K. V., & Villavicencio, A. (2014). Identification of multiword expressions in the brWaC. In *LREC* (pp. 728–735).

Bisht, R. K., Dhami, P. H. S., & Tiwari, N. (2006). An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. *Journal of Quantitative Linguistics, 13*(2–3), 161–175.

Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 conference short papers* (pp. 109–114).

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. http://doi.org/10.1023/A:1010933404324.

Breiman, L., & Cutler, A. Random forests — classification description: Random forests. http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H. C., & Uyar, E. (2010). New event detection and topic tracking in turkish. *Journal of the American Society for Information Science and Technology, 61*(4), 802–819. http://doi.org/10.1002/asi.21264.

Cessie, S. L., Houwelingen, J. C. Van, & Society, R. S. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 41*(1), 191–201. http://doi.org/10.2307/2347628.

Chang, C., & Lin, C. (2001). LIBSVM: A library for support vector machines. *Computer, 2*(3), 1–30. http://doi.org/10.1145/1961189.1961199.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). A practical guide to support vector classification. *BJU International, 101*(1), 1396–1400.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological), 20*(2), 215–242. http://doi.org/10.1007/BF03180993.

Dinçer, B. T. (2004). *Türkçe için istatistiksel bir bilgi geri-getirim sistemi*.

Evert, S. (2004). Significance tests for the evaluation of ranking methods. In *Proceedings of the 20th international conference on computational linguistics* (p. 945). http://doi.org/10.3115/1220355.1220491.

Evert, S., & Kermes, H. (2003). Experiments on candidate {D}ata for collocation extraction. In *Proceedings of the research note sessions of the 10th conference of the European chapter of the association for computational linguistics (EACL'03)* (pp. 83–86). http://doi.org/10.3115/1067737.1067754.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting on association for computational linguistics - ACL '01* (pp. 188–195). http://doi.org/10.3115/1073012.1073037.

Firth, J. (1957). *Modes of meaning. Papers in linguistics*. Oxford University Press.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries, 3*(2), 115–130. http://doi.org/10.1007/s007999900023.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment, 32*(14–15), 2627–2636. http://doi.org/10.1016/S1352-2310(97)00447-0.

Gui, J., Sun, Z., Ji, S., Member, S., Tao, D., & Tan, T. (2016). Feature selection based on structured sparsity : A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, 1–18. http://doi.org/10.1109/TNNLS.2016.2551724.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR), 3*(3), 1157–1182. http://doi.org/10.1016/j.aca.2011.07.027.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *SIGKDD Explorations Newsletter, 11*(1), 10. http://doi.org/10.1145/1656274.1656278.

Haykin, S. (1994). In M. Herrmann, H.-U. Bauer, & R. Der (Eds.), *Neural networks-A comprehensive foundation*. New York: IEEE Press *http://doi.org/10.1017/S0269888998214044*.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

Kendall, M. G., & Smith, B. B. (1939). The problem of *m* rankings. *The Annals of Mathematical Statistics, 10*(3), 275–287. http://doi.org/10.1214/aoms/1177732186.

Kiela, D., & Clark, S. (2013). Detecting Compositionality of multi-word expressions using nearest neighbours in vector space models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1427–1432 (October).

Kim, S., Yoon, J., & Song, M. (2001). Automatic Extraction of Collocations From Korean Text. *Computers and the Humanities, 35*, 273–297. http://doi.org/10.1023/A:1017507019909.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning* http://doi.org/10.1016/S0031-3203(01)00046-2.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Machine Learning: ECML-94, 784*, 171–182. http://doi.org/10.1007/3-540-57868-4.

Korkontzelos, I. (2010). *Unsupervised learning of multiword expressions*.

Koutroumbas, S., & TheoDoridis, K. (2003). *Pattern recognition. Pattern recognition*: 8 *http://doi.org/10.1016/B978-012369531-4/50016-0*.

Krcmár, L., Jezek, K., & Pecina, P. (2013). Determining compositionality of expressions using various word space models and methods. In *Proceedings of the workshop on continuous vector space models and their compositionality* (pp. 64–73).

Krčmář, L., Ježek, K., & Pecina, P. (2013). Determining compositionality of word expressions using various word space models and measures. In *In Proceedings of the workshop on continuous vector space models and their compositionality* (pp. 64–73). *August 9 2013*.

Kumova Metin, S. (2016). Neighbour unpredictability measure in multiword expression extraction. *Computer Systems Science & Engineering, 31*(3), 209–221.

Kumova Metin, S., & Karaoğlan, B. (2010). Collocation extraction in Turkish texts using statistical methods. In Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 6233 LNAI (pp. 238–249). http://doi.org/10.1007/978-3-642-14770-8_27.

Kumova Metin, S., & Karaoğlan, B. (2010). Collocation extraction in Turkish texts using statistical methods. In *Advances in natural language processing* (pp. 238–249). Springer.

Küçük, D., & Yazici, A. (2012). A hybrid named entity recognizer for Turkish. *Expert Systems with Applications, 39*(3), 2733–2742. http://doi.org/10.1016/j.eswa.2011.08.131.

Li, W., Lu, Q., & Liu, J. (2007). Chinese typed collocation extraction using corpus-based syntactic collocation patterns. In *IEEE NLP-KE 2007 - Proceedings of international conference on natural language processing and knowledge engineering* (pp. 248–255). http://doi.org/10.1109/NLPKE.2007.4368039.

Manning, C. D., & Schütze, H. (2000). *Foundations of natural language processing Reading*.

Mitchell, T. M. (1997). *Machine learning. WCB*. Boston: McGraw-Hill.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26. http://doi.org/10.1075/li.30.1.03nad.

Oflazer, K., Cetinoglu, O., & Say, B. (2004). Integrating morphology with multiword expression processing in Turkish. In *Proceedings of the second workshop on multiword expressions: Integrating processing* (pp. 64–71). http://doi.org/10.3115/1613186.1613195.

Özkan, B. (2010). Türkçenin öğretiminde sifatlarin eşdizim sözlüğü: yöntem ve uygulama. *International Journal of Educational Research, 1*, 51–65.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Third international conference on language resources and evaluation* (pp. 1530–1536).

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *LREC workshop towards a shared task for multi- word expressions* (pp. 54–61).

Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on language resources and evaluation: 17991802*.

Quinlan, J. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77–90. http://doi.org/10.1613/jair.279.

Quinlan, J. R. (1992). *C4.5: Programs for machine learning*: Vol. 1. Morgan Kaufmann San Mateo California *http://doi.org/10.1016/S0019-9958(62)90649-6*.

Reddy, S., McCarthy, D., Manandhar, S., & Gella, S. (2011). Exemplar-based word-space model for compositionality detection: Shared task system description. In *Proceedings of the workshop on distributional semantics and compositionality* (pp. 54–60). Association for Computational Linguistics.

Sag, I., Baldwin, T., & Bond, F. (2002). Multiword expressions: A pain in the neck for NLP. In *CICLing '02 proceedings of the third international conference on computational linguistics and intelligent text processing* (pp. 1–15). http://doi.org/10.1162/COLI_a_00139.

Salehi, B., Cook, P., & Baldwin, T. (2014). Detecting non-compositional mwe components using wiktionary. *Emnlp* Retrieved from *http://www.emnlp2014.org/papers/pdf/EMNLP2014189.pdf%5Cnpapers3://publication/uuid/1E3F8100-E323-428F-A6AE-18FF51803DF0* .

Say, B., Zeyrek, D., Oflazer, K., & Umut, Ö. (2002). Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the eleventh international conference of Turkish linguistics*.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., & Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the ninth international conference on language resources and evaluation* (pp. 455–461). Retrieved from. http://server01.ark.cs.cmu.edu/LexSem/mwecorpus.pdf.

Seretan, V. (2011). *Syntax-based collocation extraction*: 44 *Syntax-Based Collocation Extraction http://doi.org/10.1007/978-94-007-0134-2*.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks, 11*(5), 1188–1193. http://doi.org/10.1109/72.870050.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Tatar, S., & Cicekli, I. (2011). Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science, 37*(2), 137–151. http://doi.org/10.1177/0165551511398573.

Tsvetkov, Y., & Wintner, S. (2013). Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 1–20. http://doi.org/10.1162/COLI_a_00177.

Tur, G., Hakkani-Tur, D., & Oflazer, K. (2003). A statistical information extraction system for Turkish. *Natural Language Engineering, 9*(2), 181–210. http://doi.org/10.1017/s135132490200284x.