International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Analysis of Web-based frequency in Multiword Expression Detection

Hande Aka-Uymaz[1], Senem Kumova-Metin*[1]

Accepted : 05/06/2017 Published: 30/09/2017

Abstract:Multiword expressions (MWEs) are syntactic and/or semantic units in language, where the meaning of whole is limitedly connected to the meanings of the constituting units. The most prominent property that distinguishes MWEs from random word combinations is the recurrence. The recurrence is commonly measured by the occurrence frequencies of the MWE and the constituting words. Though occurrence frequency measures are known to be best in distinguishing MWEs from random combinations, the performance of those measures depend mainly on the quality and size of the data source where frequencies are obtained. The main goal of this study is to provide a detailed analysis on the change in performance of frequency based measures when the traditional frequency source, corpus, is swapped with a massive and dynamic data source, the World Wide Web. In order to use the web as a frequency source, the constituting words and word combinations are queried among a popular search engine, and the number of results for each query is accepted to be web-based frequency for the regarding word/word combination. In this study, the web-based frequencies are employed in three different MWE detection-related experiments utilizing a Turkish data set. In first group of experiments, the individual performances of 20 well-known frequency metrics in ranking/sorting MWE candidates based on their tendency to be a MWE is examined. Secondly, the most successful frequency metrics are determined by a feature selection method: filtering. Lastly, MWE detection is accepted to be a classification problem. Eight supervised methods are applied in order to show the combined performance of frequency metrics when the frequency is obtained from web. In all experiments, the performance of web-based frequencies in identification of MWEs is compared to the performance of traditional corpus based frequencies. The experimental results showed that the use of web-based frequency in identification of MWEs reveals promising results.

Keywords:multiword expressions, occurrence frequency, web based-frequency,feature selection, supervised learning.

1. Introduction

The multiword expressions (MWEs) are defined as idiosyncratic interpretations that cross word boundaries [1]. They are known to be a combination of two or more words that correspond to some conventional way of saying things where the meaning of the whole may not be predicted simply by interpreting individual meanings of constituting words [2]. One of the first definitions of the notion MWE that overlaps with the term*collocation* is given by Firth, in 1967 [3]. He states “collocations of a given word are statements of the habitual or customary places of that word”. Since the researchers do not commonly agree on a single definition of MWEs, they tend to study on distinguishing properties of MWEs to build MWE identification tools. The commonly accepted property of the MWEs, especially a subset of MWEs, is that the words constituting the MWE co-occur more than the words in random combinations. This property is named as occurrence frequency. The occurrence frequency of MWE candidates and the constituting words are measured from a data source to be employed in a variety of different ways. Each different use of measured frequencies is named as a statistical (frequency) measure/metric. A data source that provides reliable frequency values is essential while statistical metrics are to be used. The data source that is commonly utilized in MWE extraction is the corpus, which refers to an organized collection of written texts that are built by different methods by collecting the texts in a given period of time.The quality and the size of the corpus have a prominent effect on the performance of MWE extraction methods. If the corpus involves a wide range of texts, such as texts on different topics, articles written by different authors, and texts with different writing styles etc., the corpus is accepted to represent the natural language better that may result with more reliable frequency values for the words and MWEs. For example, it is difficult to come across to the term “BeyazSaray” (Eng. The White House) in a corpus that consists of articles about the medicine or the term “Genetic Engineering” in collection of history texts.Though these MWEs are both well-known and frequently used word combinations in their own domains. One other drawback in corpus use is observed when the corpus is not dynamically extended with new texts. For example, when a newly born word (e.g “selfie”) is searched in a static corpus, the occurrence frequency will be zero though the word may be used frequently in language. In our previous work presented in [4], we proposed to overcome the drawbacks of traditional static corpus usage in frequency-based MWE extraction methods and introduced a new notion: web-based frequency. In this paper, in order to strengthen our hypothesis on the use of dynamic corpus; web; for MWE extraction, we extend our previous work by two folds:

- Application of machine learning methods to analyse the combined performance of web-based frequency features
- Determination of features succeeding in MWE detection by feature filtering.

In addition, related work is also enhanced in this extended paper.
The alternative data source, web, in our experiments contains heterogeneous live data and is assumed to be the richest resource for human language technologies storing the highest number of texts in language. The web may be used as a data source in many different ways. For example, researchers may collect a set of texts to build a general-purpose corpus or a more specific one that includes only domain specific texts. In statistical MWE detection, the occurrence frequencies are required to measure frequency-based metrics. In our experiments, the occurrence frequencies are obtained by the use of a search engine. Simply, we propose to send MWE candidates and their constituting words as queries to

[1]*Department of Software Engineering, Faculty of Engineering, İzmir, TURKEY*
** Corresponding Author:Email: senem.kumova@ieu.edu.tr*

the search engine and employ retrieved page counts as occurrence frequencies. Google, currently being the most widely used search engine, is employed in order to retrieve page counts in our experiments.

The use of web-based frequency in MWE identification is examined in three sets of experiments. In first set of experiments, we utilized 20 different frequency-based metrics individually in order to generate a single sorted list of MWE candidates based on each metric. The sorted lists of MWE candidates based on different metrics are examined to compare the performance of each metric when web-based frequencies are employed. In first set of experiments, we also measured the same metrics based on corpus-based frequencies and presented the performance change in MWE detection.

In second set of experiments, we applied filtering, which is a well-known feature selection approach in machine learning, in order to determine best performing features in MWE detection.

In third set of experiments, accepting MWE detection as a binary classification problem, well-known supervised machine learning algorithms are run employing web-based frequency metrics as features.

The performance of web-based frequency is measured utilizing Turkish MWE data sets by three metrics: precision, recall and F-measure.

The term MWE in this study is limited to the consecutive two-word combinations (bigrams) in text. A bigram is annotated as a MWE if it belongs to one of the following groups:

- *Phrasal verbs and idioms:*Phrasal verbs are MWEs that consist of a verb in combination with a preposition or adverb or both, the meaning of which is commonly different from the meaning of its constituents. For example, the phrasal verb "açığavurmak" in Turkish is "to reveal" though the constituents "açığa" is "open" and "vurmak" is "to knock". The term idiom refers to the group of words in a fixed order forming an expression whose meaning is not predictable from the usual meanings of its constituents.
- *Stock phrases:* A stock phrase is a MWE that is frequently and traditionally used by a group of persons and thus associated with them. For example, stock phrases in Turkish "sertkahve" and "acıgerçek" refer to "strong coffee" and "grim reality", respectively in English.
- *Technical terms:* The terms that have a specific meaning within a specific field of expertise are named as technical terms (e.g. "molekülergenetik" (Eng. molecular genetics), "antipsikotikilaç" (Eng. antipsychotic drug)).
- *Named entities and job titles:* The real-world objects such as persons (e.g. "Alan Turing"), locations (e.g. "New Castle"), organizations denoted by proper names are considered as named entities in the areas of natural language processing and information retrieval. In this study, we enhanced this group by including job titles such as "genelmüdür" (Eng. "general manager")

The contribution of the study is summarized as follows:

- The notion of web-based frequency is presented and its performance in MWE detection is examined over a Turkish MWE data set.
- A set of occurrence frequency based methods is applied on two different base sets and individual performances are compared based on the sorted lists of MWE candidates.
- A comparison of two different sources of occurrence frequency; web and corpus; is provided.
- The combined performance of frequency-based metrics are examined by supervised learning methods both for corpus based and web-based frequencies

This paper is organized as follows: Section 2 presents previous works on MWE extraction. In section 3, the proposed method is introduced. Section 4 details the experimental set-up procedures. Finally, in section 5 the results are given and the paper is concluded.

2. Related Work

In literature, there exista variety of MWE definitions. In this study, we accept the terms MWE and collocation as similar though in some studies, collocation is defined to be a type of MWEs in which the high recurrence is detected. Since each MWE definition focuses on particular features of MWEs, there are no known rules to construct all types of MWEs. However, there are some common properties that are accepted to shape MWEs.

The first property of the MWEs is known to be the recurrence, which is the most widely measured and the easiest property to observe. Almost all extraction techniques suggest that a MWE must differ from other word combinations in some kind of frequency metric ([5], [6], [7]). This property enables the use of occurrence frequency metrics in MWE recognition.

The second property is being language specific. For example, in Turkish, the English MWE "wisdom teeth" refers to "yirmiyaşdişleri" (Eng. the teeth of age 20)though the word-by-word translation of "wisdom teeth" to Turkish is "akıldişleri". Since the MWEs are language specific, it is not possible to translate MWEs simply in a word-by-word manner, which makes this property very important for machine translation. One other problem that arises due to language specificity property is observed when the language is to be automatically generated and/or understood. Since there are no known rules that define how a word chooses a particular word or word combinations from millions of different words in language while creating a MWE [8], it is not easy to understand or generate the language automatically. For instance, "sweet dreams" is a commonly usedMWE in English, but there is no clear explanation for why "sweet" is preferred instead of "candy" which is almost a synonym for sweet.As a result, the systems to generate/understand the language fail even if they have the information on word senses.

An other commonly accepted property is the meaning integrity of constituting words in MWE. This feature enables MWEs to create unit blocks of meaning where the meaning of the whole is commonly different than the meaning of the parts [8].

The last property of MWEs is being domain dependent. There are several domain specific MWEs in different domains such as science, medicine, art and sports. Smadja[7] described this property with an example from sailing domain. He exemplified the domain dependency with the MWE "dry suit". Dry suit is a term that refers to a special type of suit used by sailors to keep warm. However, comprehending these meaning easily is hard for even native speakers of the language.

MWE/collocation extraction studies can be categorized in a variety of different ways. For example, methods may be categorized based on the approach to decide on the MWEs (e.g. rule-based, supervised, unsupervised, ranking or any combination of these methods) or the type of information used (e.g. statistical, linguistics, statistical and linguistics or dictionary-based information) in the study. Table 1 presents some example studies and the regarding category information.

The first study in Table 1 is a rule-based system that employs linguistic information [9]. As it is stated in [9], several patterns of MWEs are defined and a semi-lexicalized rule based method is employed to detect those patterns. The study of Tsvetkov and Wintner[10] is another example for studies, which utilize linguistic information.

Tsvetkov and Winther [10] proposed the use of linguistically motivated features as classification features in order to classify given MWE candidates by a neural network.

In many of the previous MWE studies, it is observed that MWE extraction process typically proceeds by scoring collocation candidates with a frequency metric [11]. In such studies, the ultimate goal is generating a ranked list of MWE candidates using a variety of commonly statistical metrics. In ranking approach, the higher scores (lower ranks) mean the closer the candidate is to being a collocation.

Table 1. Examples of MWE studies

Study	*Method*	*Information Type*	*Language*	*Corpus*
Oflazer et al. [9]	Rule Based	Linguistic	Turkish	Two corpora of news text
Tsvetkov and Wintner[10]	Supervised	Linguistic	Hebrew	46M-token monolingual Hebrew corpus
Pecina [17]	Ranking and supervised	Statistical	German and Czech	German Adj-N & PP-Verb collocation candidates, Czech dependency bigrams from the Prague Dependency Treebank
Ramisch et.al [18]	Ranking and filtering	Statistical	English	Genia corpus
Kumova Metin [12]	Filtering	Statistical	English	Leipzig Corpora collection
Kumova Metin&Karaoğlan [8]	Ranking	Statistical	Turkish	Bilkent Corpus
Kim, et. al[13]	Filtering	Statistical	Korean	Yonsei corpus
Li et. al. [14]	Filtering	Statistical	Chinese	PolyU Treebank and Peking University Corpus
Piao, S et.al [15]	Filtering	Statistical	Chinese	Chinese corpus built at CCID tool

Several frequency metrics have been utilized in the literature such as point wise mutual information, joint probability and t-test ([5], [16]). A well-known work on ranking is presented in [17]. In [17], 55 association measures are combined by standard statistical classification methods, which are modified in order to provide scores for ranking [17]. It is reported that the methods that are the combinations of multiple frequency metrics result in performance improvement [17].

In earlier studies on MWE extraction, various methods are utilized English corpora because of the lack of tagged corpora in different languages [12]. However, recently, in a significant amount of studies, it is observed that non-English corpora; such as Turkish [9],[8], Korean [13] and Chinese [14] [15] are employed. For instance, in the study of Kim, et. al.[13], four statistical metrics have been utilized in order to deal with the flexible word order of the Korean collocations. Then they separated meaningful bigrams using an evaluation function [13]. Li et.al. [14] presented a corpus-driven framework which generate collocations for nouns and verbs phrase, then they combined them using statistical frequency metrics to extract noun/verb phrase collocations in Chinese. In the studyof Pia et. a [l5] an existing statistical tool made for English is used to test the automatic identification and extraction performance of Chinese MWEs.

3. Methodology

In MWE extraction, traditionally the association between words is measured by the co-occurrence frequencies of the words. It is simply accepted that as the co-occurrence amount of words increases, the ties between the words get stronger indicating the association between them. In statistical MWE detection, the metrics known as lexical association measures/metrics are employed to measure the strength of ties between words. Each lexical association measure presents a different way to utilize frequencies that belong to constituents and the word combination. For example, in the well-known metric, joint probability, only the frequency of word combination is considered. This metric enables to sort word combinations based on their frequency. If the co-occurrence frequency of the given combination in corpus is higher compared to other combinations in the same corpus, the regarding combination is accepted to be much more closer to form a MWE. On the other hand, in one other metric, point-wise mutual information, both the frequency of word combination and constituents' frequencies areconsidered. The ratio of constituents' frequenciesto the combination's frequency is accepted to indicate (non) existence of an MWE. Similar to joint probability, point-wise information metric also provides a list of sorted word combinations.

In this study, web is used to extract occurrence frequencies of MWE candidates together with their constituting words in lack of a reliable/large corpus. We believe that this new data source may provide more reliable occurrence frequencies since it has access to great numbers of documents when compared to a single traditional corpus. In other words, web containing the highest number of texts that are generated in language resembles the language better than a corpus of limited size. In our approach given in [4], we proposed to send the constituting words and the word combination independently to the search engine and retrieve the number of documents that they are observed, individually as presented in Fig. 1.In our experiments, web occurrence frequencies are obtained by querying the candidate MWEs (bigrams) and the constituting words from the popular search engine: Google.

To exemplify the results of this procedure, in Table 2, the retrieved number of documents (page counts) obtained for the bigram "Abidin Dino" (a famous artist in Turkey) and the constituting words "abidin" and "dino" are given. The retrieved number of documents; the frequency metrics; listed in Table 3, are used to evaluate the performance of the proposed notion of web-based frequency in our experiments. Thewell-defined frequency metrics in Table 3 are commonly used in the previous studies [11], [17], [19]. In Table 3, $f(w_1w_2)$ is the occurrence frequency (e.g. the number of retrieved documents from Google) of a bigram w_1w_2 $f(w_1)$ and $f(w_2)$ are the frequencies of constituents of the bigram w_1 and $w_2$respectively. $f(w_1\overline{w_2})$stands for a bigram that starts with word w_1 and the following word can be anything except w_2.$f(w_1w_2)$=261.000 is used as the occurrence frequency (web frequency) of the bigram "Abidin Dino". And the page counts 6.350.000 and 101.000.000 are accepted to be web-based frequencies of words "abidin" and "dino" respectively.

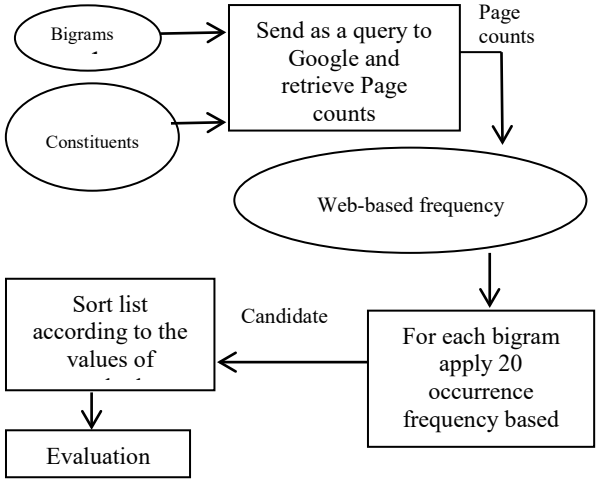

Fig. 1. Process flow chart of proposed method.

Table 2. Sample query results

Search term	Page count	Notation
"abidindino"	261.000	$f(w_1w_2)$
"abidin"	6.350.000	$f(w_1)$
"dino"	101.000.000	$f(w_2)$

$f(w_2|w_1)$is the conditional probability of w_2 given $w_1$and it is calculated as follows:

$$f(w_2|w_1) = \frac{f(w_1w_2)}{f(w_2)} \quad (1)$$

The metrics in Table 3are used in threegroups of experiments. Firstly, metrics are used tosort/rank MWE candidates according to their tendency to be a MWE. This experiment enables to compare the performance of different metrics when web and corpus frequencies are employed.

In second group of experiments, a well-known approach in feature selection, filtering, is employed to compare the performances of metrics (accepted as features) in MWE detection. Filtering is a method that assesses the MWE detection performance of features by the use of an attribute evaluator. The approach sorts the features based on decreasing order of evaluator scores enabling the comparison and determination of the best and worst performing features. In our experiments, we employed relief-F and information gain measures as attribute evaluators.

The relief-F (RelF) measure is an iterative evaluator, proposed in [20]. In this measure, a sample from data set is chosen in each iteration, two nearest samples in data set that belongs to the same and opposing class with the regarding sample is determined. The nearest sample in same class is named as “near-hit” and the other is “near-miss”. The distances to near-hit and near-miss are measured and the difference is calculated. If the distance difference for the whole data set for a feature is high, it is accepted that the feature is successful in classification. The attribute evaluator Relief–F in this study is the relief-F algorithm proposed by Kononenko[21] in which there exists improvements such as the replacement of Euclidean distance to Manhattan distance and the use of absolute distance.

Information gain is defined as a measure of reduction of disorder/uncertainty in data set based on a specific feature [22] Information gain (IG) is calculated as follows

$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$	(2)
$IG(S,A) = H(S) - H(S\|A)$	(3)

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 and $H(S|A)$ is the entropy of the class S given the feature A.

In third group of experiments, the metrics are accepted to be features/indicators in MWE identification. And MWE identification is assumed to be a binary classification problem where a given candidate is assigned as MWE or non-MWE by a supervised algorithm. In second group of experiments7 different supervised methods are utilized by WEKA [23] machine learning tool. The methods are (with their original names in WEKA tool):

1) *Naive Bayes (NB)*: Naive Bayes classifier is one of the simple and fast classifier that bases on Bayes theorem. In this classifier, it is assumed that each feature is conditionally independent of other features and the classification label (class) is conditionally dependent on all features. By these assumptions, for each class label, a conditional probability value is calculated and all values multiplied to generate a single value for each class. Following, the class that has the maximum probability value is assigned to the sample. The further information on the Naive Bayes classifier employed in our study may be found in [24].
2) *Sequential Minimal Optimization (SMO)*: SMO is a function-based classifier that implements John Platt's sequential minimal optimization algorithm [25] for training a support vector classifier.

Table 3. Occurrence frequency based metrics used for MWE detection

Frequency-Based Metrics

1. Jointprobability (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)- $\log\left(1+\frac{f(w_1w_2)}{(f(w_1w_2)+f(w_1\overline{w_2}))}\right) + \log(1+\frac{f(w_1w_2)}{(f(w_1w_2)+f(\overline{w_1}w_2))})$
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. Simpson (Simp) – $\frac{f(w_1w_2)}{\min(\ f(w_1w_2)+f(w_1\overline{w_2}),f(w_1w_2)+f(\overline{w_1}w_2)}$
15. Driver-Kroeber (DK) - $\frac{f(w_1w_2)}{\sqrt{(f(w_1w_2)+f(w_1\overline{w_2}).(f(w_1w_2)+f(\overline{w_1}w_2)}}$
16. Piatersky-Shapiro (PS) - $P(w_1w_2) - P(w_1)P(w_2)$
17. Jaccard (JC) - $\frac{f(w_1w_2)}{f(w_1w_2)+f(w_1\overline{w_2})+f(\overline{w_1}w_2)}$
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. Mountford (MF) - $\frac{2f(w_1w_2)}{2f(w_1\overline{w_2})f(\overline{w_1}w_2)+\ f(w_1w_2)f(w_1\overline{w_2})+f(w_1w_2)f(\overline{w_1}w_2)}$
20. Fager (F) - $\frac{f(w_1w_2)}{\sqrt{(f(w_1w_2)+f(w_1\overline{w_2}).(f(w_1w_2)+f(\overline{w_1}w_2)}} - \frac{1}{2}\max(f(w_1\overline{w_2})f(\overline{w_1}w_2))$

3) *K-nearest neighbor (IBk):* In k-nearest neighbor algorithm, known as a lazy algorithm, the samples are labeled with the class of the majority class of its k number of neighbors. For example, if k=3, then the 3 closest labeled neighbors of the regarding sample is determined. If most of the neighbors are MWE then the sample is labeled with MWE, vice versa. In our experiments, we set k=5.
4) *One Rule (OneR):*OneR classifier generates a rule for each predictor/feature in the data set and specifies the rule with the smallest total error as its "one rule". A frequency table for each feature against the target is formed in order to create a rule for a feature [26].
5) J*48*: J48 is the WEKA implementation of pruned or un-pruned C4.5 decision tree. C4.5 tree can be thought as an improved version of ID3 tree [27]. While using the concept of information entropy, C4.5 forms decision trees from a set of training data with the same way as ID3. According to their frequency of access, sub-trees can be moved to the different levels. Unlike ID3 trees, pruning can be done in C4.5 trees. In every node of the tree, C4.5 selects the data attribute that splits its instances into subsets ideally. While splitting the criteria is the normalized information gain, the attribute that has the highest normalized information gain is selected to make the decision. Following, sub-decision trees can be constructed by creating a sub-list under the new decision node [27]
6) *Adaptive Boosting (AdaBoostM1)*:Adaptive boostingis a machine learning meta-algorithm introduced in [28]. This algorithm can be used with other learning algorithm in order to improve their performance. If there is a learning algorithm that generates classifiers whose performance is a little better than random guessing, AdaBoosting can be used to diminish the error [28]. *AdaBoostM1* is one of the versions of AdaBoosting and it is used in binary classification problems[28]. *AdaBoostM1* can be used when there is a multiclass classification problem and multiclass base classifier needs to be boosted [29]
7) *RandomForest(RF)*:Random forest algorithm, based on decision tree method, requires to merge the trees that are trained by a different training subset[30] In this algorithm, multiple classifiers are trained and the samples are classified according to the votes that come from these classifiers. In order to generate classifiers independently, the features that are employed in each tree are chosen randomly. The trees in random forests are not pruned ([30],[31]). The algorithm is faster compared to the similar algorithms and vulnerable to over-fitting [32].

4. Experimental Setup

In this section, the experimental setup procedure is given in detail. In the following subsections, base sets employed in experiments are presented; the evaluation metrics and the experimental steps followed in study are defined.

4.1. Base Sets

In this study, three base sets, BS1, BS2 and BS3 are employed in experiments. The first base set, BS1, is built by frequency-based methods. Normalized frequency, point-wise mutual information, chi-square test and t-score methods are applied and all bigrams in corpus are sorted in decreasing order of their corresponding values. The first 200 bigrams that have occurrence frequency more than or equal to 5 in sorted lists are merged to build BS1, similar to the procedure in [8] as given in Fig 2.
Bilkent[33], Leipzig[34], Egecorpus, BilCol[35], Muder[36] and Metu [37] corpora are used to construct BS1. The second base set, BS2, is a set of idioms and bigrams that mimic the features of idioms. BS2 is prepared to assess the proposed method on MWE candidates that are not occurring frequently in language. The procedure that is followed to build BS2 is presented in Fig 3.
Base sets, BS1 and BS2, are annotated by 3-4 native speakers based on a guideline provided by researchers. Inter-rater agreement among the annotators is measured by Fleiss Kappa [38]. The resulting Fleiss Kappa values are calculated as ~0.728 and ~0.767, respectively for BS1 and BS2.

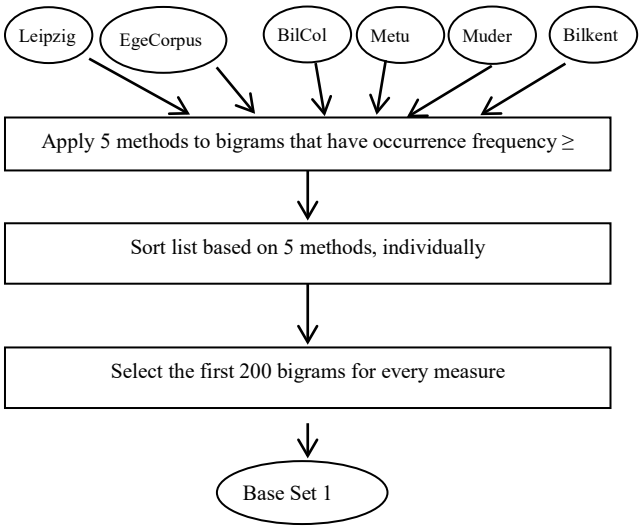

Fig. 2. The procedure followed in construction of base set 1 (BS1).

Bilkent[33], Leipzig[34], Egecorpus, BilCol[35], Muder[36] and Metu [37] corpora are used to construct BS1. The second base set, BS2, is a set of idioms and bigrams that mimic the features of idioms. BS2 is prepared to assess the proposed method on MWE candidates that are not occurring frequently in language. The procedure that is followed to build BS2 is presented in Fig 3.
Base sets, BS1 and BS2, are annotated by 3-4 native speakers based on a guideline provided by researchers. Inter-rater agreement among the annotators is measured by Fleiss Kappa [38]. The resulting Fleiss Kappa values are calculated as ~0.728 and ~0.767, respectively for BS1 and BS2.
The last base set, BS3, is a subset of BS1 that is built to compare the performance when web-based frequency is used instead of corpus-based frequency. The corpus-based frequencies in BS3 are obtained from Leipzig corpus [34]. It is observed that Leipzig corpus include a subset of 1245 (~55.85%) candidates of BS1. Table 4 gives the statistics of BS1, BS2 and BS3

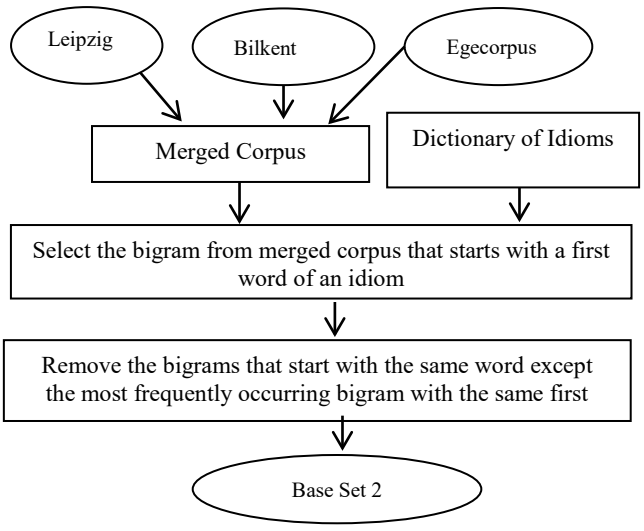

Fig. 3. The procedure followed in construction of base set 2 (BS2).

Table 4. Annotated base sets

Base Set	Number of Bigrams annotated as MWE	Number of Bigrams annotated as non-MWE	Total
BS1	1194(~53.56%)	1035(~46.43%)	2229(100%)
BS2	891(~63.14%)	520(~36.85%)	1411(100%)
BS3	733(~58.87%)	512(~41.124%)	1245(100%)

4.2. Evaluation

Experimental results are evaluated by well-known measure of classification; F1-measure which is harmonic mean of precision and recall. In MWE identification, precision can be considered as the fraction of correctly MWE assigned candidates. The recall is the fraction of correctly MWE assigned candidates to the all MWEs in the set. And F1 measure is represented as follows:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

4.3. Experimenting Procedure

In order to analyze the performance of web-based frequency in MWE detection, firstly, web-based frequencies are employed to calculate the values of frequency-based metrics mentioned in section 3. The whole set of the candidates and their constituents are sent as queries to Google individually in a period of 2 months. The retrieved page counts are used whenever required. Since the total size of the web or in other words the total size of the words in web is not known, instead of calculating probabilities matching frequencies are utilized in frequency metrics.

The study covers three groups of experiments. In first group, each frequency-based metric is utilized to sort the MWE candidates individually. The metrics succeeding in MWE detection; in other words, the metrics that tend to assign MWEs to lower ranks and vice versa; when web-based frequencies are employed are examined. In addition, the performance change when web-based frequency is used instead of corpus-based frequency is also investigated. In this experiment, the MWE candidates in BS1 that are also observed in Leipzig corpus [34] are used. The corpus based frequencies for the regarding candidates are obtained from Leipzig corpus. The data set (candidates) in this experiment includes 733 (~58.87%) MWEs and 512 (~41.124%) non-MWEs.

In second group of experiments, the MWE detection is accepted as a binary classification problem rather than a sorting task. The frequency-based metrics are used as features distinguishing MWEs and non-MWEs. The filtering method with two different attribute evaluators is applied to sort the metrics in descending order of classification performance. In the last group of experiments, different classification methods with different number of frequency-based metrics are run to analyze the success in classification.

5. Results

In first group of experiments, after utilizing each frequency metric listed in Table 3, the base sets (BS1 and BS2) are sorted according to the metric values of the candidates. F-measure is measured for first *N* candidates of the sorted base sets where *N* is varied from 1 to total number of candidates in set to obtain the curves. Then, the average value of F-measure (F_{avg}) value, the area under F-curve (F_{area}) of every metricis calculated. In Table 5, sorted lists according to F_{avg} of 20 frequency-based metrics can be seen for BS1 and BS2.

In Tables 5, it is clearly seen that LFMD metricprovides the maximum F_{avg} and F_{area} for both base sets. In addition, the scores of the second maximum values for all measures are significantly lower than the scores of LFMD. LFMD, MD and CP measures that involve the operands $f(w_1w_2)$, $f(w_1)$ and $f(w_2)$ in common are in the set of 5 best performing metrics for both sets. It is also observed that in most of the metrics in best performing metrics set the term $f(w_1w_2)$ is divided by the multiplication off(w_1) and $f(w_2)$.

The F-measure curves of 3 best and worst performing measures are presented in Figures 3 and 4 for BS1 and BS2 respectively. In Figure 3 and 4, horizontal axis represents the number of MWE candidates in base sets and vertical axis represents the F-value.

Table 5. Test results of the frequency metrics for BS1 and BS2 sorted according to F_{avg}

	Base Set 1			Base Set 2		
	Measure	**F_{area}**	**F_{avg}**	**Measure**	**F_{area}**	**F_{avg}**
1	LFMD	1271,98	0,571	LFMD	825,42	0,585
2	Rcost	1266,67	0,568	JP	820,61	0,582
3	CP	1266,53	0,568	CP	820,12	0,581
4	MD	1266,13	0,568	JC	814,05	0,577
5	DK	1266,13	0,568	MD	813,31	0,576
6	SSS	1265,39	0,568	DK	813,31	0,576
7	JC	1264,80	0,567	Rcost	812,83	0,576
8	SK	1261,96	0,566	NE	812,79	0,576
9	NE	1256,18	0,564	BB	811,31	0,575
10	Simp	1256,07	0,564	SSS	810,87	0,575
11	Scost	1255,79	0,563	SK	805,61	0,571
12	FK	1253,92	0,563	FK	803,74	0,570
13	BB	1253,31	0,562	Simp	801,59	0,568
14	RCP	1214,43	0,545	Scost	801,59	0,568
15	JP	1207,94	0,542	PMI	784,33	0,556
16	PMI	1182,46	0,530	RCP	772,39	0,547
17	MF	1159,13	0,520	MF	752,32	0,533
18	Ucost	1093,43	0,491	Fager	745,74	0,529
19	Fager	1041,69	0,467	Ucost	740,69	0,525
20	PS	1010,50	0,453	PS	728,70	0,516

Observing the F measure curves for BS1 and BS2, it can be stated that for both data sets Fager, PS and U cost metrics fail in ranking the candidates considering the whole range of N. Overall, it can be stated that LFMD and DK measures generate higher F-values for BS1 and LFMD and CP measures perform better for BS2.

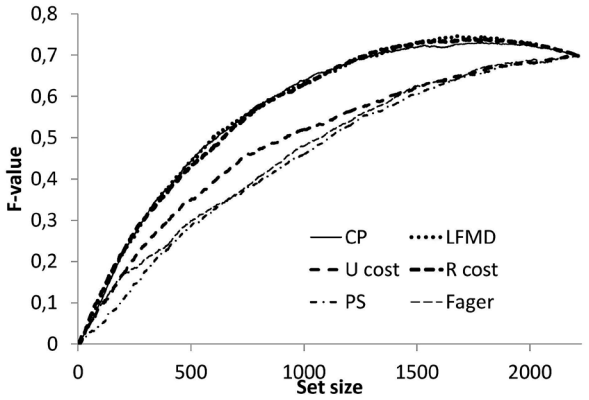

Fig. 3.F-measure curves of BS1.

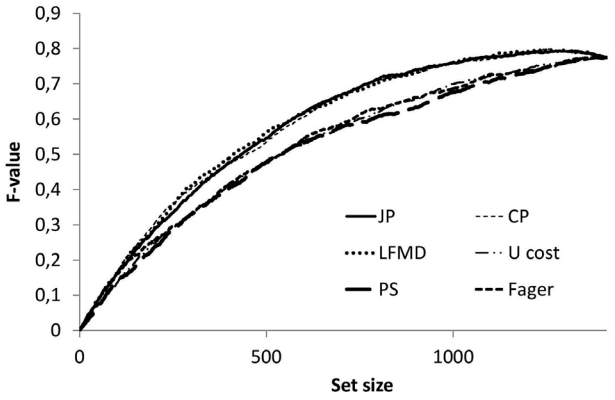

Fig. 4.F-measure curves of BS2.

Table 6 shows the results where the use of web and corpus-based frequencies are to be compared on BS3. As given in Table 6, LFMD is examined to be the best performing measure when the frequencies are obtained from Leipzig corpus and CP is the best performing measure when the frequencies are obtained from web. F_{avg} and F_{area} value are approximately dame for best performing metrics for two different types of frequency.MD and Rcost methods are commonly observed in the first 5 best performing measures for both sources.

Table 6 Sorted lists of frequency metricsfor BS3 (based on F_{avg})

Web-based frequency			Corpus-based frequency		
Measure	**F_{area}**	**F_{avg}**	**Measure**	**F_{area}**	**F_{avg}**
CP	705,45	0,57	LFMD	705,04	0,57
JC	705,29	0,57	MD	676,59	0,54
SSS	704,08	0,57	DK	676,58	0,54
Rcost	703,88	0,57	Rcost	676,50	0,54
MD	703,44	0,57	BB	675,64	0,54
DK	703,44	0,57	NE	675,62	0,54
LFMD	701,54	0,56	FK	675,62	0,54
CP	705,45	0,57	LFMD	705,04	0,57
NE	700,32	0,56	JC	675,62	0,54
BB	699,93	0,56	SSS	675,62	0,54
FK	699,67	0,56	CP	668,79	0,54
SK	698,61	0,56	SK	668,06	0,54
Simp	693,41	0,56	Simp	663,42	0,53
Scost	693,32	0,56	Scost	663,03	0,53
PMI	680,93	0,55	MF	660,95	0,53
RCP	670,33	0,54	PMI	656,43	0,53
MF	667,39	0,54	RCP	654,47	0,53
Ucost	650,45	0,52	Ucost	653,68	0,53
Fager	650,19	0,52	Fager	649,95	0,52

Table 7 gives the results on BS1 and BS2 of the filtering experiments. For each data set, two evaluators (IG and RelF) are used to evaluate the performance of frequency-based metrics in MWE detection. The column *Av* represents the average of IG and RelF ranks. For example, PMI is the first and the third best metric in IG and RelFevaluatorsrespectively on BS1 and on average it is also best performing frequency metric for BS1.In Table 7, the frequency metrics are given in sorted order of the average rank (*Av*). The shaded regions contain 5 best performing metrics for BS1 and BS2. It is examined that PMI, LFMD, MD and UCost are commonly reside in this best performing group. When result of first experiment and filtering are considered together, it may be stated that there exists some metrics such as LFMD and MD that may succeed both in ranking (first group experiments) andclassifying (second group of experiments) the MWE candidates.

Table 7. Filtering results on BS1 and BS2 methods

BS 1				BS 2			
Metric	***IG***	***RelF***	***Av***	***Metric***	***IG***	***RelF***	***Av***
PMI	1	3	2	PMI	1	1	1
LFMD	6	1	3,5	LFMD	6	3	4,5
MD	5	4	4,5	MD	5	4	4,5
Ucost	9	2	5,5	RCP	4	6	5
Scost	8	5	6,5	Ucost	9	2	5,5
JP	2	12	7	CP	3	10	6,5
RCP	4	14	9	Scost	8	5	6,5
Rcost	10	9	9,5	NE	7	8	7,5
CP	3	18	10,5	BB	13	7	10
FK	11	10	10,5	JP	2	18	10
NE	7	15	11	Rcost	10	12	11
JC	17	8	12,5	SK	12	11	11,5
SSS	18	7	12,5	DK	15	9	12
Fager	20	6	13	FK	11	16	13,5
PS	16	11	13,5	Simp	14	13	13,5
BB	13	16	14,5	JC	17	14	15,5
SK	12	19	15,5	SSS	18	15	16,5
DK	15	17	16	PS	16	19	17,5
MF	19	13	16	Fager	20	17	18,5
Simp	14	20	17	MF	19	20	19,5

Table 8 presents the filtering results on BS3 in a similar fashion to Table 7. This experiment has an important outcome. It isthat the 5 best performing frequency metrics (given in shaded regions) are almost same (e.g. LFMD, PMI, MD, NE) either web-based or corpus-based frequency is employed. In addition, failing metrics such as PS, Simp, SK are also same for both frequency sources.

Table 8. Filtering results on BS3

Web-based frequency				Corpus –based frequency			
Measure	**IG**	**RF**	**Av**	**Measure**	**IG**	**RF**	**Av**
PMI	1	3	2	Fager	5	1	3
LFMD	6	2	4	NE	1	10	5,5
Ucost	9	1	5	PMI	9	3	6
MD	5	7	6	MD	6	8	7
NE	7	6	6,5	LFMD	12	5	8,5
RCP	4	13	8,5	DK	7	11	9
Scost	8	9	8,5	CP	13	6	9,5
BB	13	5	9	JC	3	16	9,5
JP	2	16	9	Ucost	19	2	10,5
Rcost	10	8	9	FK	2	20	11
CP	3	19	11	RCP	18	4	11
Fager	20	4	12	BB	11	12	11,5
DK	15	10	12,5	Rcost	8	15	11,5
FK	11	15	13	SSS	4	19	11,5
JC	17	11	14	MF	10	14	12
SK	12	18	15	Simp	17	7	12
SSS	18	12	15	Scost	16	9	12,5
MF	19	14	16,5	SK	15	13	14
PS	16	17	16,5	PS	14	18	16
Simp	14	20	17	JP	20	17	18,5

The third group of experiments to evaluate the web-based frequency in MWE detection includes utilization of 7 supervised methods. The meta-algorithm mentioned in this group of supervised methods, *AdaboostM1* algorithm, is used to improve *J48* algorithm and named as *AdaJ48*. In this study, the tests are performed by 5 fold-cross validation using WEKA machine learning tool. The average weighted F-values of 5 folds are calculated for two different sets of frequency metrics. The first set includes the whole set of metrics (totally 20 metrics) mentioned in Table 3. The second set includes the metrics that are examined to be in 5 best performing metrics during filtering experiments (given in shaded regions of Table 7).Table 9 gives the experimental results for BS1 and BS2 data set. The columns *All* and *Best5* refer to the weighted F-values that are obtained when the whole set of metrics and best 5 metrics (determined by filtering) are employed, respectively. The shaded regions in Table 9 show the highest F value in regarding column. The results in Table 9 show that supervised methods provide F-values in range [0.622 0.675] except *NB*. Though the method*NB* is the most improving method when the metrics are filtered and it generates the highest F value in BS2, when all metrics are employed it fails in classification. Examining the results in Table 3, it may be stated that *AdaJ48* where *J48* is boosted generates highest F-values in BS1. In addition, it is observed that *AdaJ48*together with *NB* succeed most in BS2 when best metrics are considered.

Table 9. Machine learning results on BS1 and BS2

	BS1		BS2	
Method	***All***	***Best5***	***All***	***Best 5***
NB	0,415	0,586	0,581	0,658
SMO	0,663	0,656	0,622	0,612
IBk	0,646	0,638	0,635	0,638
OneR	0,638	0,639	0,648	0,648
J48	0,673	0,652	0,632	0,649

AdaJ48	0,675	0,657	0,638	0,658
RF	0,661	0,653	0,638	0,648

Table 10 presents the experimental results of BS3 (weighted F values) that are obtained in a same manner to BS1 and BS2 data sets. The row *Average* in Table 10 refers to the average F value of the regarding column. For example, when best 5 metrics are used the supervised machine learning methods generates F=0.615 on average. Examining the averages, it is seen that when best performing metrics are used the performance of supervised methods improves on average for both frequency types. Similar to Table 9, in Table 10 the highest F values are shaded for each column. When the highest F values are considered, the best performing method is *NB* except the case where web-based frequency is used with whole set of metrics. One other important outcome that may be extracted from Table 10 is that there exists no significant difference between the performances of web-based frequency and corpus-based frequency, and the F values vary in range [0,591 0.645] excluding NB with all metrics and [0,577 0.644] for web-based and corpus-based frequencies respectively.

Table 10. Machine learning results on BS3

	Web-based frequency		**Corpus-based frequency**	
Method	***All***	***Best5***	***All***	***Best 5***
NB	0,302	0,645	0,644	0,642
SMO	0,600	0,594	0,622	0,609
IBk	0,604	0,595	0,577	0,606
OneR	0,591	0,604	0,589	0,589
J48	0,638	0,618	0,586	0,605
AdaJ48	0,643	0,641	0,611	0,607
RF	0,625	0,611	0,615	0,620
Average	0,572	0,615	0,606	0,611

6. Conclusion

In this study, we analysed the performance change in MWE detection when web-based frequency is used instead of corpus-based frequency. The main aim in use of the web-based frequency is that the occurrence frequencies or any other information based on frequencies obtained from a static corpus may not be realistic since the static corpora include limited and static number of texts that may not represent the whole language. Obtaining the frequency from search engine, we performed three different experiments by employing 20 frequency-based metrics: sorting the MWE candidates based on their tendency to be a MWE, selecting the succeeding metrics by feature selection, supervised learning. The MWE detection performances are obtained both for web and corpus-based frequencies in all experiments. As a conclusion, it is examined that the use of web-based frequency in MWE detection is an alternative solution to corpus-based studies.

Acknowledgements

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

References

[1] I. A. Sag, T. Baldwin , F. Bond , A. Copestake , D.Flickinger, "Multiword Expressions: A Pain in the Neck for NLP", In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, 2001 (CICLing-2002)

[2] C.D. Manning and H. Schutze, "*Foundations of Statistical Natural LanguageProcessing"*, MIT Press. England, 1999.

[3] J.R. Firth,"Modes of Meaning",.Papers in Linguistic 1934-51, Oxford UniversityPress ,1967.

[4] H.Aka-Uymaz, S.Kumova-Metin "Using web data in identification of multiword expressions in Turkish" in 4th International Conference on Advanced Technology& Sciences (ICAT'Rome) Rome, Italy, November 23-25, 2016

[5] R. K Bisht, H.S.Dhami, and N.Tiwari,"An evaluation of differentstatistical techniques of collocation extraction using a probability measureto word combinations", Journal of Quantitative Linguistics, Vol.13,161-175, 2006.

[6] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics",1990, Vol. 16 No.1, 22-29 .

[7] F.A. Smadja,"Retrieving Collocations from Text:Xtract",Computational Linguistics,Vol. 19 No. 1, 143-177 ,1993.

[8] S.Kumova-Metin andB.Karaoğlan,, "Collocation Extraction in Turkish Texts Using Statistical Methods", 7th International Conference on Natural LanguageProcessing (LNCS-ISI) IceTAL, Reykjavik, Iceland,2010.

[9] K.Oflazer,O.Çetinoğlu and B. Say," Integrating morphology with multi-wordexpression processing in Turkish", Proceedings of theWorkshop on MultiwordExpressions: Integrating Processing,.p. 64-71, 2004.

[10] Y. Tsvetkov and S. Wintner, "Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources", Proceedings of the 2011Conference on Empirical Methods in Natural Language Processing, pages:836-845, Edinburgh, Scotland, UK, July 2731,2011.

[11] G. Bouma, "Collocation Extraction beyond the Independence Assumption"Proc. ACL 2010 Conf. Short Pap.;10914,2010.

[12] S.Kumova-Metin, "Neighbour Unpredictability Measure in Multiword Expression Extraction", International Journal of Computer Systems Science andEngineering:31-3, 2016.

[13] S. Kim, J. Yoon and M. Song,"Automatic Extraction of Collocations From KoreanText",Computers and the Humanities 35: 273–297, 2001.

[14] W. Li, Q. Lu and J. Liu,"Chinese typed collocation extraction using corpus based syntactic collocation patterns", IEEE NLP-KE 2007 - Proceedings ofInternational Conference on Natural Language Processing and KnowledgeEngineering,2007.

[15] Piao, S, Sun, G, Rayson, P and Yuan, Q "Automatic extraction of Chinese multiword expressions with a statistical tool" Paper presented at Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, 2006, .

[16] P. Pecina, "Lexical association measures and collocation extraction." Language Resources Evaluation. 2010;44(1-2).

[17] P.Pecina, "A Machine Learning Approach to Multiword Expression Extraction", Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions,2008.

[18] Ramisch, C., Villavicencio, A., Boitet, C.: mwetoolkit: a Framework for Multiword Expression Identification, LREC, 2010.

[19] S. Kumova-Metin, T.Kışla and B.Karaoğlan, "Named Entity Recognition in Turkish Using Association Measures", Advanced Computing: An International Journal, Vol.3, No.4, 2012

[20] K. Kira, and L. A. Rendell. "A Practical Approach to Feature Selection", Proceedings of the ninth international workshop on Machine learning,1992.

[21] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF." Machine Learning: ECML-94 784: 171–82.,1994.

[22] T. Mitchell "Machine Learning" WCB. Boston: McGraw-Hill, 1997.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten"The WEKA data mining software: an update", SIGKDD Explorations Newsletter 11(1): 10,2009.

[24] G. H. John. and P. Langley. "Estimating Continuous Distributions in Bayesian Classifiers" In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, 338-345.

[25] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines", Microsoft, 1998.

[26] R.C. Holte "Very simple classification rules perform well on most commonly used datasets", Machine Learning. 11:63-91, 1993.

[27] J.R. Quinlan,"C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[28] Y. Freund and R. E. Schapire,"Experiments with a new boosting algorithm,. In: Thirteenth International Conference on Machine Learning, San Francisco, 148-156, 1996.

[29] G. Eibl and K. P. Pfeiffer, "How to Make AdaBoost.M1 Work for Weak Base Classifiers by Changing Only One Line of the Code" In: Elomaa T., Mannila H., Toivonen H. (eds) Machine Learning: ECML 2002. Lecture Notes in Computer Science, Vol.2430. Springer, Berlin, Heidelberg, 2002.

[30] L. Breiman, “Random Forests”. Machine Learning. 45(1):5-32, 2001.
[31] K.J. Archer and R. V. Kives,“Empirical characterization of random forest variable ımportance measures”, computational statistical data analysis, Computational Statistics & Data Analysis, 52(4), 2249-2260, 2008.
[32] L. Breiman and A.Cutler, Random forest, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm , 2005, (Accessed 3/3/2017).
[33] G.Tür, D. Hakkani-Tür and K.Oflazer, “A statistical Information ExtractionSystem for Turkish” Natural Language Engineering”, Vol 9 No.2, 181-210,2003.
[34] U. Quasthoff, M. Richter and C.Biemann,“Corpus portal for search in monolingual corpora”, Proceedings of the Fifth International Conference on Language Resources and Evaluation, 2006
[35] F. Can, S. Kocberber, O. Baglıoglu, S.Kardas, H. C. Ocalanand E. Uyar,“New event detection and topic tracking in Turkish”, Journal of the American Society for Information Science and Technology, Vol. 61, no. 4, pp. 802-819,2010.
[36] T. Dinçer, “Türkçeiçinistatistikselbirbilgigeri-getirimsistemi”, Phd Dissertation,U.B.E.,EgeUniversitesi ,2004.
[37] B. Say, D.Zeyrek, K.Oflazer and U.Ozge, "Development of a Corpus and a Treebank for Present-day Written Turkish", Proceedings of the Eleventh International Conference of Turkish Linguistics, 2002.
[38] J.L Fleiss, "Measuring nominal scale agreement among many raters", PsychologicalBulletin 378-382,1971.