International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Pronoun Resolution for Tweets in Turkish

Huseyin Abaci*[1], Mete Eminagaoglu[2], Iclal Gor[3], Yilmaz Kilicaslan[4]

Submitted: 15/02/2022 Accepted: 22/04/2022

Abstract: This paper aims to provide an analysis of anaphoric relations in tweets in Turkish language. The analysis offered rests on the results of a sequence of experiments conducted using a group of machine learning algorithms. The algorithms used in this study are J48, Voted Perceptron, SVM (support vector machine), Naive Bayes and k-nearest neighbours. These classifiers were experimented by parametric variations are scrutinized to elaborate on the problem of matching a model conveniently to the task available already. Another important contribution of the paper is the comparison offered between two genres of texts, namely tweets versus child stories. Our experimental results are compared with those of the previous work and, thereby, a comparison is offered between the anaphoric structure of tweets and that of child stories in Turkish.

Keywords: *Pronoun resolution, Turkish tweets, machine learning, anaphoric relation, learning curve*

1. Introduction

An anaphor is an expression that is referentially dependent upon another expression, which is said to be its antecedent. Anaphora resolution is the process by means of which the antecedent of an anaphor is discovered. Consider the following fragment of text.

Allen dropped the jar. **It** shattered loudly. (1)

The pronoun "it" in the second sentence is an anaphor. *Allen* and *the Jar* are the two candidate antecedents from which this anaphor can potentially get its reference. Considering the semantic properties of the candidates, it is obvious that *the jar* is the real antecedent. However, the resolution of a pronoun is not always so straight forward. Take, for instance, the Turkish tweet below (from the corpus used in this study).

Bizim ülke-nin kalkınmak için tek yol-u var.
our country-gen develop for single way-poss pred.exist
O da bilişim, teknoloji-ler-i-nden
she/he/it part informatics technology-plur-poss-abl (2)
en üst seviye-de yararlanmak.
most top level-loc benefit
'There is only one way for our country to develop. It is to benefit from informatics technologies at the top level.'

Ignoring semantic and/or pragmatic clues, the genderless Turkish pronoun 'o' (she/he/it) in this example can refer to any of the preceding third person nominals. In general, in natural language processing (NLP), determination of the antecedent of an anaphoric expression is accepted to be a difficult problem [1], [2]. Several types of research and studies have been conducted in order to find a solution to this problem. Different anaphora resolution methods were proposed previously such as knowledge-intensive techniques that rely on semantic, syntactic, and real-world information or knowledge-independent methods that are based on the information explicitly provided from the texts. Mitkov et al. point out that comparative assessment or analysis of models used in these areas is highly important due to the fact that a lot of diverse approaches or methods in anaphora resolution propagate rapidly with an increasing number of alternatives [1], [2].

Although an adequate number of researches have been conducted for English, there are only a handful of relevant studies offered for Turkish. Nonetheless, Turkish deserves much more attention as a language with its own peculiarities. Turkish is a pro-drop language. That is to say, pronouns can be left silent (i.e., non-pronounced) in this language. Such covert pronouns make both the annotation process and the resolution process more difficult. Several researchers have tried to make an explanation regarding the unique distribution of overt pronouns versus the covert ones in Turkish language. For instance, Kornfilt et al. point out that whenever possible, the pronoun is deleted or discarded in Turkish [3]. Yıldırım et al. put forth this particular feature in Turkish language by showing the difficulty in obtaining subcategorization frames automatically in Turkish is mainly caused by the fact that null pronouns are used much more often in Turkish when compared to other languages [4]. The same feature is known to be a vital obstacle in the analysis or resolution of anaphoric relations as well. In Turkish, the recovery of null pronouns will probably require extra effort and an extensive amount of error-prone work if a text is to be processed within a system for anaphora resolution. It should also be stressed that, if it is pointing its antecedent, then a third-person English pronoun can be more informative than a third-person Turkish pronoun, even when it is used overtly. Thus, such a pronoun may become ambiguous, since it does not possess or provide any gender information, which is used to distinguish between a male person, a female person, or an inanimate object. For instance, "o" in Turkish can be translated into "he", "she", or "it" in English depending on the context. Even though there is some work on anaphora resolution in Turkish, no work, to the best

[1] *Department of Computer Engineering, Adnan Menderes University, Aydın, 09010, Türkiye, ORCID ID:0000-0001-7433-540X*
[2] *Department of Computer Engineering, Dokuz Eylul University, İzmir, 35390, Türkiye, ORCID ID:0000-0003-2456-919X*
[3] *Department of Computer Engineering, Adnan Menderes University , Aydın, 09010, Türkiye, ORCID ID:0000-0002-1999-8283*
Department of Computer Engineering., Adnan Menderes University, Aydın, 09010, Türkiye, ORCID ID:0000-0002-5020-6547
** Corresponding Author Email: huseyin.abaci@adu.edu.tr*

of our knowledge, has been published as to how to carry out this task in Turkish tweets. In fact, even the English language has not been given satisfactory attention in this respect.

There are several basic differences between tweet texts and other genres (e.g., stories). For instance, tweeters usually think that sending tweets is a casual socializing activity (i.e., a written form of the spoken language). Hence, they usually do not care about grammatical or typological mistakes and spelling errors in their tweets. In addition to the spelling mistakes, also informal abbreviations of the words (e.g., "4 you" for "for you", "cu ltr" for "see you later", "how r u" for "how are you") are also frequently found in tweet texts. It is obvious that such abbreviations and grammatically incorrect or informal short sentences will bring about difficulties for the tasks of text annotation and anaphora resolution.

The aim of this paper is to provide an analysis of anaphoric relations in tweets in Turkish. The analysis rests on the results of a sequence of experiments conducted using a group of machine learning algorithms Different machine learning algorithms have been used by applying parametric variations to the selected algorithms and each of them is scrutinized to elaborate the problem of conveniently matching a model to the anaphora resolution of tweets in Turkish language. Another important contribution of the paper is the comparison offered between two genres of texts, namely tweets versus child stories. In a previous study, on which this work is based, the same learning algorithms have been applied to child stories in a similar experimental setting. We compare our experimental results with those of the previous work and, thereby, offer a comparison between the anaphoric structure of tweets and that of child stories in Turkish. The comparative results in this study provide assurance and validation of the theory that the anaphora resolution among less structured textual data, such as tweets, can be more successful than structured texts such as stories or novels. It should also be noted that the aim of this study is not to provide the highest accuracy levels with a machine learning algorithm or an automated system that could be as successful as humans could.

2. Related Works

The study proposed by Yıldırım et al. has been considered as a pioneer work for our research, which is elaborated in this article [4]. In addition, our study can be thought of as an improvement of another study by Kılıçaslan et al., where the same algorithms as those used in this study were applied to a group of child stories in Turkish language in order to propose a unique approach for the analysis of Turkish anaphoric relations [5]. An adequate number of anaphora resolution researches available in English (selected list) [1], [2], [6-9], [44-45]. Cunnings et al. have studied anaphora resolution using the visual world paradigm in second language processing [10]. They compare native and non-native English speakers in terms of their anaphoric resolution capabilities. Stojnic et al. suggest a new approach to demonstrative and reflexive pronouns [11]. Anaphora resolution can also be utilized for different implementations such as document categorization [12]. Furthermore, Othman et al. empirically evaluate to decide on whether reply links promote product feature extraction in tweet texts or not [13]. Kornfilt et al. propose out that a pronoun is always deleted or discarded in Turkish language if it is possible to do so. They also point out null pronouns and the use of personal pronouns in Turkish is observed to be far less when compared to other languages such as English [3]. Some linguistic researches have focused on to find an explanation for the unique distribution observed among overt pronouns versus null pronouns in Turkish. Erguvanlı tried to formulate "sentence-bounded conditions" for Turkish such as when to use a null or overt pronoun in any sentence and where to use it in a sentence [14]. Turan et al. proposed one of the most elaborate studies regarding the discourse-level anaphoric relations in Turkish language [15]. B. J. Grosz et al. developed a framework based on Centering Theory [16-18]. One of the recent studies in anaphora resolution for Turkish is offered by Yüksek et al., who focus on reflexive pronouns [19].

Kılıçaslan et al. proposed that anaphoric relations in Turkish language cannot be explained entirely "in terms of linear precedence or dominance constraints within the boundaries of a sentence" [20]. Based on Banfield's study, Kılıçaslan et al. put forth a "distinction between the 'core' and 'periphery' of a sentence" [21].

Some studies have also focused on different languages for anaphora resolution. Tabrizi et al. proposed a novel pre-processing approach for pronoun extraction and pronoun mapping in the pronominal anaphora resolution system of English translations of the Quran from Arabic language [22]. Singla et al. suggest several approaches for resolution in Hindi [23]. Kawasaki et al. propose a new anaphora resolution method including cases in Japanese [24].

3. Materials and Methods

Twitter is an online social networking service and system that provides communications between people via short messages named as tweets. Tweets that are shared between socially connected users (followers) can contain text about news, advertisement, state, comments and response to someone's tweets (retweets). People send their tweets for all sorts of reasons, which they are mostly used for a recreational thing, attention to some serious subjects and reaching out to more followers. However, the size of a tweet is limited to 280 characters. This limitation simply changes the text structure since tweeters require expressing an idea or topic in a limited number of words and shortened sentences, which causes the deterioration of grammatical rules most of the time. The algorithms and their parameters in this study are the same as the ones that are used in our previous work with the child storybooks [16]. However, the methodologies used in the analysis are different in many aspects from our previous study due to grammatically unstructured tweet texts.

The details and the results of our unique approach are explained in this study, which is made up of two main phases. The first one is the classification of candidate antecedents of pronouns (derived from the annotation of tweet texts) with different machine learning algorithms, and the second phase is the final resolution of the anaphoric relations by choosing the closest candidate with the highest positive classification. This unique approach might derive valuable comprehensions related to issues such as how pronoun resolution in Turkish could be an intrinsically difficult task, and which machine learning algorithms or models should be adapted or implemented for such tasks. The extraction of the features, the annotation process for tweet texts, and the machine learning algorithms used in this study are discussed in the following sections.

3.1. Feature Selection

In this study, the data has to be represented as vectors of "feature-value pairs" [5] in order to be appropriate for machine learning classifiers, which is a common issue encountered with most of the similar approaches based on machine learning. The data should be annotated with the related features in order to overcome this issue. In anaphora resolution, most of the implementations based on machine learning use a combination of "syntactic", "lexical",

"semantic", and "positional" features. The features used for the annotation of the raw data are described shortly in Table 1.

The number of features that are used within several studies is known to be different values varying between eight [25] to sixty-six [26]. It should be noted that all of the features, which provide information for the classifier algorithms in this study have been used in several studies regarding the anaphora resolution. Grammatical role and case are known to be the two syntactic features, which have been mostly explored and observed in the previous works on anaphora [27-31], [42], and [43]. These syntactic features have also been used for Turkish language in this study.

3.2. Data Collection

We have collected live stream tweets including the keywords "social" and "media" for three days. More than a million tweets were downloaded during this period where we have eliminated re-tweets "RT" from the collection that are simply duplicates of original tweets and we did not want to add these tweets into the experiments. This should not be considered as a trivial process since it requires great processing capacity, which was established by using a cloud environment. The cloud environment enables seamless Twitter API integration and parallel processing capacity to process these large tweets data sets. The data sets consist of instances and the instances subsume pronouns and its candidate and real antecedents. These instances are used for both training and testing the classifier algorithms. Nine attributes were chosen during annotation for 830 instances that are extracted from tweets. The text data contain 1838 words, 143 of which are pronouns. The third-person pronouns are distributed among several types as follows: personal (76.9%), locative (13.2%), reflexive (6.2%), and reciprocal (3.4%). Moreover, 44.1% of total pronouns within the data set are covert and 55.9% are overt.

The comparison of the text formations between two data sets namely, "Dataset-1" which is gathered from 20 child stories reported in Kılıçaslan et al. and "Dataset-2" which is comprised of Turkish tweets text about "social media" is given in Table 2. In other words, the total number of words and pronouns, percent of personal, locative, reflexive types of all pronouns and overtness (covert, overt) of pronouns within two data sets are compared in Table 2. Due to the limitation of the number of words within the tweet messaging system, the word count and pronoun count are less than the child stories (Dataset-1). In addition, the rate of locative, reflexive, reciprocal, covert, and overt pronouns of tweets greatly differ from the child stories.

3.3. Annotation and Pair Generation

Turkish is a pro-drop language that causes strict difficulties with annotation phrases in Turkish texts. Hence, some annotators assisted us in this phase and they analysed the tweets to determine pronouns of tweets and possible candidates of pronouns of that tweet. For the annotation process, we distributed the refined tweets among ten annotators for annotating the raw tweet texts and eventually generating feature vectors. The essential benefit of using human resources for the annotation process instead of a specialized annotation tool is the complexity of the tweets as it contains casual grammar and it is difficult to deduct the grammatical relationship within the text.

The outcome of annotation process is an unbalanced set of the candidate and real antecedent set of instances. The antecedent-pronoun pairs were generated first. Then, the negative samples, which the candidate antecedent is in disagreement with the pronoun in number or person, were discarded. This filtering was done so in order to avoid the positive instances being outnumbered by negative ones. The positive instances consisted of only 22% of the whole data set, which is highly imbalanced and eventually not feasible for data mining operations. After the filtering operations, this ratio was updated to an acceptable balanced ratio, in other words, 50-50%.

3.4. Classification Algorithms Used in the Experiments

All of the experiments were carried out on Weka (version 3.9.1) using "stratified ten-fold cross-validation" methodology for validation of the models [32]. Weka is an open-source data mining and machine learning software that is developed in Java programming language [33]. Five different machine learning algorithms with different parameter settings implemented in Weka were applied to each of the two data sets for binary classification. One of these algorithms was a decision tree model named as J48 [34]. It should be noted that in this study both pruned and unpruned J48 models were tested. The pruning process is used in most of the decision trees as well as in J48, where the size of the tree is reduced in order to decrease the complexity and to mitigate overfitting problem [32]. Naïve Bayes and k-nearest neighbours (a type of

Table 2. Comparative analysis of text formation of the datasets: Dataset-1 (20 child stories in Turkish) and Dataset-2 (227 tweets in Turkish).

	Data Set 1 (Child Stories)	Data Set 2 (Tweets)
Word Count	10165	2969
Pronoun Count	1149	242
Personal (%)	82.3	74.3
Locative (%)	6.6	14.5
Reflexive (%)	10.7	5.9
Reciprocal (%)	0.4	3.8
Covert (% of total pronouns)	60.4	46.3
Overt (% of total pronouns)	36.4	53.9

Table 1. Feature set for the annotation of raw data.

Feature	***Explanation***
Case	The grammatical case which a candidate antecedent or a pronoun bears: nominative, accusative, locative, genitive, dative, ablative or instrumental
Grammatical role	Whether a candidate antecedent or a pronoun is a subject or an object
Overtness	Whether a pronoun is phonetically overt or not
Type	Whether a pronoun is personal, reflexive, locative or reciprocal
Semantic type	Whether the referent of a candidate antecedent is a human being, a place, an animal, an abstract object or a physical object
Person and number	The person and number information born by a candidate antecedent or a pronoun
Position	The position of a word within a text fragment that helps the calculation of distance information
Antecedent	The position of the true antecedent of a pronoun
Referential status	Whether a nominal is a non-pronominal expression that can function as an antecedent

instance-based learning algorithm) were also included in the experiments [35], [36]. A Voted Perceptron, which is a type of an artificial neural network, was also used in this study [37]. Support a vector machine with C-support vector classification was also used [38].

The parameter settings of the classifier algorithms are also denoted in Table 3 where some of them are used with the default parameters in the Weka software and some of these classifier algorithms' parameters were set to two alternative values according to the observations during the experiments. Hence, there were ten different classifier results for each of the data sets.

4. Results and Discussion

All of the experiments with two different data sets using several classifier algorithms were conducted on a hardware platform operating on 64-bit architecture and having an Intel Core i7 2.7 GHz central processing unit and 16 Gigabytes of random access memory. It should be noted that the performance results given in Table 4 are obtained by using stratified ten-fold cross-validation where the data set consists of a total of 358 instances in which 179 records labelled as "yes" and 179 records labelled as "no". On the other hand, the performance results given in Table 5 are obtained by using the methodology ten-fold cross-validation where the data set consists of a total of 830 instances in which 179 records labelled as "yes" and 651 records labelled as "no".

The performance metrics used to compare the classifiers were chosen as Accuracy, Precision, Recall, F-measure, and Kappa statistics [32]. The accuracy metric is usually used to measure the percentage of correct predictions among the entire data set. Accuracy can be simply described as the total number of correct predictions (between both types of classes) divided by the total number of instances. The recall value gives the fraction of the relevant documents that are successfully retrieved or the rate of the correctly classified positive instances, which is also named as true positive rate. On the other hand, the precision score indicates the fraction of retrieved documents that are relevant to the query, or it is the total number of true positives divided by true positives plus false positives. The F-measure can be obtained by calculating the harmonic mean of precision and recall.

Kappa statistic or Kappa coefficient is an alternative measurement that can be used to assess algorithms' classification performances. In machine learning, Kappa score is used as a measure to assess the improvement of a classifier's accuracy over a predictor employing chance as its guide [39]. Landis et al. suggest, "A Kappa score over 0.4 indicates a reasonable agreement beyond chance" [40]. As shown in Table 4, most of the algorithms' Kappa scores seem to be significantly better than a random classification. It could be seen from Table 4 that the best values in terms of all performance measures (Accuracy, Precision, F-measure, and Kappa statistic) were obtained by unpruned J48 decision tree algorithm.

Even though some of the machine learning algorithms that are given in Table 4 provided promising results, it can also be deduced that the performance results regarding "Precision", "Recall", and "F-measure" do not seem to be satisfying enough. Ho et al. attribute "three possible factors to the failure of popular classifiers to perform to perfect accuracy: (1) deficiencies in the algorithms, (2) intrinsic difficulties in the data and (3) a mismatch between problems and methods" [41]. All of the classifier algorithms seem to perform in favour of "no" cases, which is denoted in Table 5. The reason for these biased (overfitting) results is due to the fact that the number of instances classified as "no" is much more than the instances that belong to the "yes" class. It can be seen that there is a significant performance improvement after balancing the number of records classified as "no" and "yes". The balanced data set was achieved by random selection without replacement. We

Table 3. Parameter settings for each of the classifier algorithms.

Algorithm name	*Parameter settings*
J48 (pruned)	confidence factor used for pruning: 0.25, subtree raising on pruning, all the other parameters with default values
J48 (unpruned)	no pruning, all the other parameters with default values
SVM (C-SVC, RB kernel)	kernel function: Radial basis kernel, all the other parameters with default values
SVM (C-SVC, linear kernel)	kernel function: Linear kernel, all the other parameters with default values
Voted Perceptron (exp.=1)	exponent for polynomial kernel = 1, all the other parameters with default values
Voted Perceptron (exp.=2)	exponent for polynomial kernel = 2, all the other parameters with default values
Naive Bayes (kernel est.)	kernel estimator for numeric attributes, all the other parameters with default values
Naive Bayes (normal d.)	normal distribution for numeric attributes, all the other parameters with default values
k-NN (k=1)	k=1 (one nearest neighbor), distance metric: Euclidean distance, all the other parameters with default values
k-NN (k=11)	k=11 (eleven nearest neighbors), distance metric: Euclidean distance, all the other parameters with default values

Table 4. Classification results achieved by ten-fold cross-validation (balanced - weighted average of evenly distributed sample sizes) using the data set with 358 instances.

Algorithm name	*Accuracy*	*Precision*	*Recall*	*F-measure*	*Kappa statistic*
k-NN (k=1)	0.793	0.797	0.793	0.793	0.5866
C-SVC (Kernel function: Linear)	0.74	0.744	0.74	0.739	0.4804
Naive Bayes (Normal distribution for numeric attributes	0.718	0.72	0.718	0.717	0.4358
J48 (unpruned)	**0.813**	**0.813**	**0.813**	**0.813**	**0.6257**
Voted Perceptron (polynomial kernel exponent=1)	0.754	0.754	0.754	0.754	0.5084
Naive Bayes (Kernel estimator for numeric attributes)	0.718	0.72	0.718	0.717	0.4358
k-NN (k=11)	0.777	0.787	0.777	0.775	0.5531
J48 (pruned)	0.802	0.802	0.802	0.802	0.6034
Voted Perceptron (polynomial kernel exponent=2)	0.791	0.791	0.791	0.79	0.581
C-SVC (Kernel function: Radial basis)	0.76	0.767	0.76	0.758	0.5196

should not jump to any conclusions by referring to Table 5. It can be seen from this table that the unbalanced data set does not provide us realistic and unbiased information. If we look at the Kappa statistic results, they contain some misleading data. This is due to the fact that the majority of the instances are "no" cases, algorithms simply provide lucky "no" guesses for all of the instances. Therefore, algorithms provide high performance score because many instances within the test set are "no" cases. Relying on unbalanced data set to run learning algorithms would be a misleading or non-realistic approach. It is known that a small amount of data usually provides misleading outcomes. In order to determine the saturation level of algorithms (learning curve analysis), we have randomly selected 25 instances from the balanced data set, tested each set gradually increasing the train set size by 25, and optimized the algorithms within the entire data set. The highest F-measure and Kappa statistic values obtained in learning curve with balanced data sets are given in Table 6.

Figure 1(a) and 1(b) denote the F-measure and Kappa statistics of J48 algorithm with pruned and unpruned alternatives. The train data size has been increased by 25 instances for each experiment run to find out the saturation level of the algorithm for tweets analysis. The trend line of F-measure and Kappa results are also denoted in these figures. Figure 1(a) shows the saturation level (around 90%) has been reached at 175 instances both by the pruned and unpruned J48 decision trees. It can be seen from these figures that the J48 algorithm successfully classifies the tweet instances in terms of anaphoric resolution. In addition, F-measure reaches 96% by the unpruned J48, which is 3% more than the pruned J48 model. We can suggest that expressive power becomes stronger with the pruned J48 decision tree. Figure 2(a) and 2(b) show the k-NN algorithm with "k" parameter set to 1 and 11. Figure 2(b) shows that the saturation point (80%) is obtained first at train data size with 125 instances and followed by a decrease in the F-measure. The second saturation point is observed at training with 275 instances. This shows us the k-NN algorithm with k parameter set to 11 provides less accurate results and less expressive power than the k-NN with one nearest neighbour (k=1).

Two extreme cases for the k-NN algorithm are denoted in Figure 2(a) and 2(b), where the best case when parameter "k" is one and the worst case when "k" is 11. However, in order to observe the trend line of the parameter k's effect on the performance, we have also conducted additional experiments by setting the k value from two to 10. Figure 3 shows the performance score of F-Measure and Kappa Statistics of the k-NN algorithm in terms of different "k" values. The classification results are obtained by ten-fold cross-

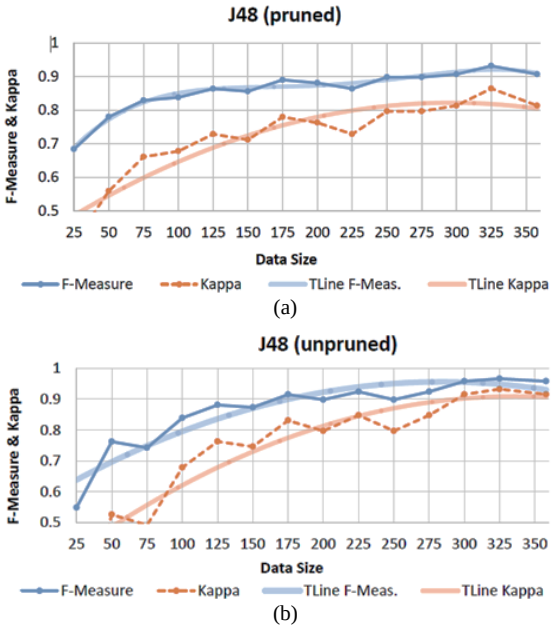

Fig. 1. F-measure and Kappa statistics of J48 algorithm and their tren lines (TLine) with respect to data size (training data). Figure a) shows pruned J48 and b) shows unpruned J48.

Table 5. Classification results achieved by ten-fold cross-validation using the unbalanced data set with 830 instances.

Algorithm name	*Accuracy*	*Precision*	*Recall*	*F-measure*	*Kappa statistic*
k-NN (k=1)	0.780	0.777	0.780	0.778	0.342
C-SVC (Kernel function: Linear)	0.787	0.762	0.787	0.700	0.030
Naive Bayes (Normal distribution for numeric attributes	0.790	0.761	0.790	0.766	0.261
J48 (unpruned)	**0.799**	**0.790**	**0.799**	**0.794**	**0.376**
Voted Perceptron (polynomial kernel exponent=1)	0.788	0.748	0.788	0.747	0.185
Naive Bayes (Kernel estimator for numeric attributes)	0.790	0.761	0.790	0.766	0.261
k-NN (k=11)	0.799	0.770	0.799	0.749	0.186
J48 (pruned)	0.795	0.763	0.795	0.761	0.233
Voted Perceptron (polynomial kernel exponent=2)	0.806	0.782	0.806	0.782	0.307
C-SVC (Kernel function: Radial basis)	0.784	0.615	0.784	0.690	0.000

Table 6. Highest F-measure and Kappa statistic values obtained in learning curve analysis with the balanced data sets.

Algorithm name	*Train data size (instances)*	*F-measure*	*Kappa statistic*
J48 (pruned others: default parameters)	325	0.932	0.864
k-NN (k=1, others: default parameters)	358	1.000	1.000
NaiveBayes (Kernel estimator for numerical attributes)	325	0.864	0.723
NaiveBayes (Normal Distribution for numerical attributes)	325	0.864	0.723
k-NN (k=11, others: default parameters)	325	0.830	0.661
J48 (unpruned - others: default parameters)	325	0.966	0.932
VotedPerceptron (exponent=1 others: default parameters)	325	0.864	0.723
VotedPerceptron (exponent=2 others: default parameters)	300	0.898	0.797
C-SVC (Kernel function: Linear others: default parameters)	358	0.881	0.762
C-SVC (Kernel function: Radial basis others: default parameters)	358	0.863	0.723

validation using the balanced data set with 358 instances. In addition, similar experiments were also conducted with the weighted distance of the k-NN where weight is calculated as "1 / distance" and the closest instance has a greater contribution to determining the classification of the query instance. We have observed that when "k" is increased in ordinary the k-NN algorithm, F-Measure decreases gradually from 0.8 reaching 0.7 and Kappa score changes from 0.58 to 0.55. We have observed a similar trend line for other values of "k" (4, 5 …, 10) which are not shown in the figure since performance scores consistently decrease while "k" is incremented. However, we did not observe the same trend line with the weighted distance of the k-NN analysis.

The performance scores of the F-measure and Kappa statistics interestingly peaked at k = 11, it changed from 0.8 (k = 1 with weighted distance) to 0.81 (k = 11) for the first case and from 0.58 to 0.61 for the latter. Using weighted distance, k-NN's Kappa score increases by 4% when k is set to 11. Within the weighted distance k-NN algorithm, the distance becomes a denominator of value "1" and a smaller distance implies that the denominator value eventually gets closer to zero. Therefore, nearer neighbours' contribution to the classification performance is higher and this is best observed with the higher "k" values.

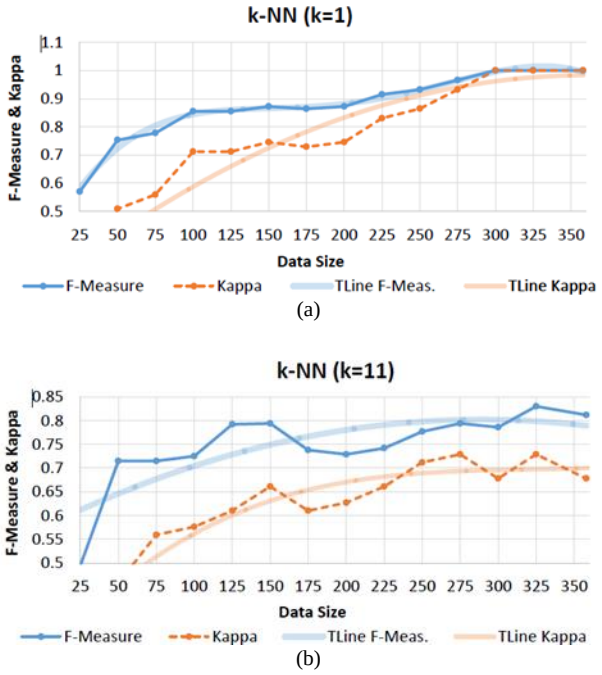

Fig. 2. F-measure and Kappa statistics of k-NN algorithm and their trend lines (TLine) with respect to data size (training data). Figure a) shows parameter k is set to 1 and b) is set to 11 for k-NN.

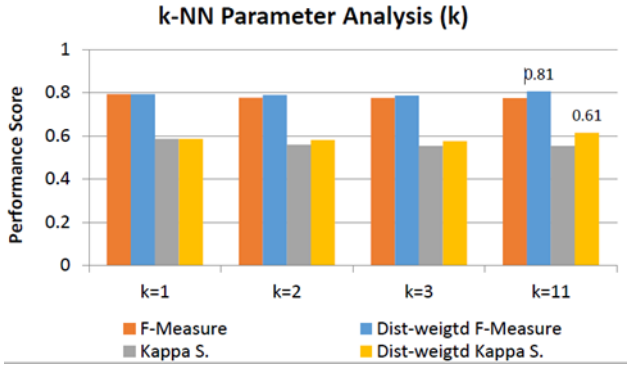

Fig. 3. Performance score of F-Measure and Kappa Statistics of the kNN algorithm in terms of parameter "k" change and distance weighted k-NN.

F-measure and Kappa statistics of the Naïve Bayes algorithm with kernel estimator and normal distribution parameters are shown in Figure 4(a) and 4(b). Trend lines imply that the saturation level is reached at 150 for kernel estimator and normal distribution cases. These figures also show us the Naïve Bayes' learning curve F-score is around 85%, which is less than the other algorithms' learning curve performance scores. The reason for this weak performance is based on the fact that the Naïve Bayes assumes there does not exist any relationship between features or attributes. This very nature of the algorithm produces a false assumption because relationships exist actually between the features that we have selected for anaphoric resolution analysis of twitter texts.

F-measure and Kappa statistics obtained by the Voted Perceptron (VP) algorithm with polynomial kernel exponent parameter set to 1 and 2 are given in Figure 5(a) and 5(b) respectively. The figures show us the expressive power of the VP is found to be less than the J48 and the k-NN algorithms. Our initial observation showed that the value of polynomial kernel exponent changes the overall performance of the VP. In order to improve the performance of the VP, we also observed the performance where the polynomial kernel exponent was set to three and four. Figure 6 shows performance score of the F-Measure, and Kappa Statistics of the Voted Perceptron algorithm with four different polynomial kernel exponent values. It can be seen from Figure 6 that although performance scores are very close to each other, the performance is improved the most when exponent parameter is set to 3. The F-measure and Kappa statistics for the C-SVC algorithm with linear kernel function and radial basis kernel function are shown in Figures 7a and 7b respectively. It can be noticed from the figures that linear kernel function performs better than radial basis kernel function because characteristics of the tweets are not appropriate for radial basis kernel function. The C-SVC algorithm with linear kernel function provides an F-measure score around 90%, which is denoted in Figure 7a. From the Figures 1(a) to 7(b), two conclusions can draw: (i) Overall performance of the algorithms regarding the expressive power is observed to be robust. If expressive power gets higher, saturation level can be delayed with limited number of instances. Hence, more data performs reverse

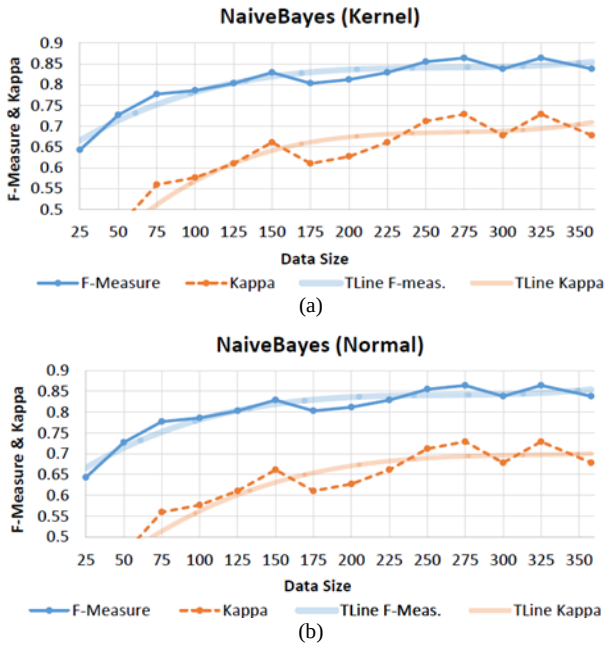

Fig. 4. F-measure and Kappa statistics of Naive Bayes algorithm and their trend lines (TLine) with respect to data size (training data). Figure a) shows results for Kernel estimator, and b) for normal distribution.

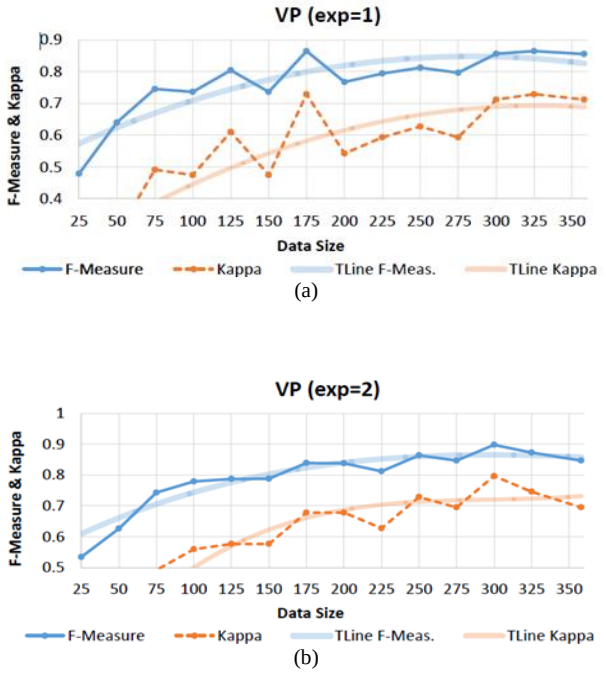

Fig. 5. F-measure and Kappa statistics of Voted Perceptron (VP) algorithm and their trend lines (TLine) with respect to data size (training data). Figure a) shows results for exponent=1 and b) for exponent=2.

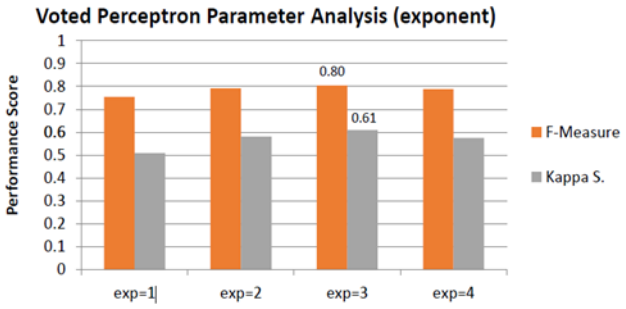

Fig. 6. Performance score of F-Measure and Kappa Statistics of the Voted Perceptron algorithm in terms of parameter "exponent" change. Exp = 3 gives better result.

effect on both. (ii) We have reached the happy graph (saturation level) within each algorithm's learning rate performance.

The results obtained from the balanced data set in this study are also compared with the previous study using the same machine learning algorithms with the same parameters and they are given in Table 7. The highest values in terms of all performance measures (Accuracy, F-measure, and Kappa statistic) were obtained by support vector machine with a radial basis kernel function (C-SVC) in previous study, but all the best performances were observed with the unpruned J48 decision tree algorithm in this study. As shown in Table 7, support vector machine's accuracy value seems to be slightly better than the J48 decision tree.

On the other hand, F-measure and Kappa statistics obtained by the J48 algorithm outperforms the support vector machine's results. When the two data sets are compared, it can be seen that the accuracy and Kappa scores are similar. However, there are noticeable differences between F-measures obtained with child stories and tweet texts. This is induced by the different linguistic structures of the child stories and tweet texts. The real antecedent of a pronoun is usually located on one or more sentences previously in the child stories. This distance even increases within a text written for adults. However, the real antecedent of a pronoun

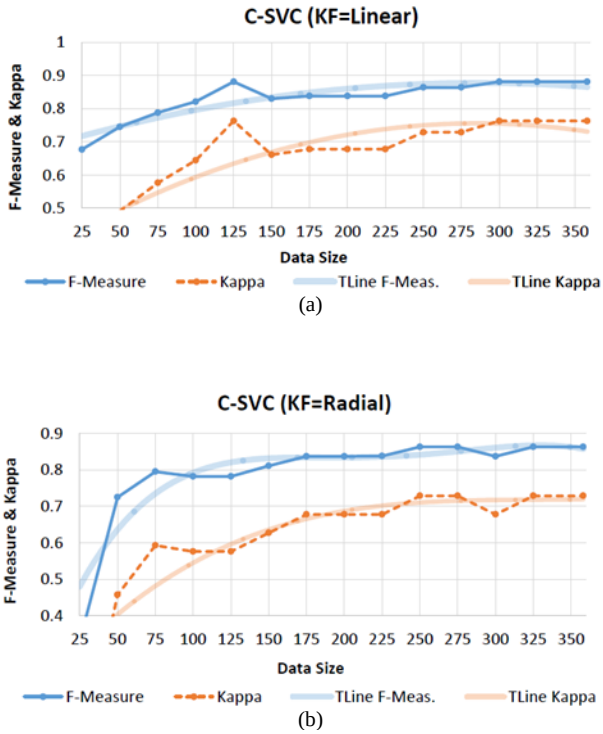

Fig. 7. F-measure and Kappa statistics of C-SVC algorithm and their trend lines (TLine) with respect to data size (training data). Figure a) shows results for kernel function=linear and b) for kernel function=radial basis.

is usually positioned within the same sentence in tweet texts. This greatly affects the "distance" feature and the algorithms' classification performance as well.

Moreover, the number of candidate antecedents of a pronoun also varies greatly within these two data sets. The number of candidate antecedents is observed between 8 and 10 in the child stories, however, this number is observed as five or less in tweet texts. This means that the number of instances in the child stories is higher than the latter case, and machine learning algorithms hovering over larger solution space to find the real antecedent. This decreases the performance score of the classifier algorithms (i.e. Accuracy, or F-measure) because algorithms process the wrong outcome such as finding non-real antecedents, which eventually lowers the scores.

5. Conclusion and Future Work

In this study, the anaphora resolution for tweets that are written in Turkish was analysed by the means of several machine learning algorithms. The machine learning algorithms that were used in this study were a decision tree (J48), an artificial neural network (voted perceptron), a support vector machine, a Naive Bayes, and a k-NN with different parameters. A large number of tweets were collected via Internet cloud environment where re-tweets and advertisement tweets were discarded. The tweets were analysed by annotators in order to obtain the relevant data sets. The data sets consist of instances and the instances contain pronouns and its candidate and real antecedents. These instances are used for both training and testing the machine learning algorithms for binary classification.

The results obtained in this study show that the saturation point is reached between 150 and 300 instances within the algorithms' learning curve analysis. The highest F-measure performance score was obtained as 83%, which can be considered as a promising result because obtaining such classification rates for Turkish tweet data is very difficult. Kappa statistics show that classification

Table 7. Comparative analysis of classification performances of several algorithms among the data sets: Dataset-1 (20 child stories in Turkish) and Dataset-2 (227 tweets in Turkish).

Algorithm name	*Accuracy*	*Dataset-1 F-Measure*	*Kappa statistic*	*Accuracy*	*Dataset-2 F-Measure*	*Kappa statistic*
k-NN (k=1)	0.750	0.670	0.479	0.793	0.793	0.587
C-SVC (Kernel function: Linear)	0.770	0.640	0.474	0.740	0.739	0.480
Naïve Bayes (Normal distribution)	0.780	0.680	0.520	0.718	0.717	0.436
J48 (unpruned)	0.770	0.670	0.497	**0.813**	**0.813**	**0.626**
Voted Perceptron (polynomial kernel exponent=1)	0.790	0.680	0.515	0.754	0.754	0.508
Naïve Bayes (Kernel estimator)	0.800	0.700	0.540	0.718	0.717	0.436
k-NN (k=11)	0.800	0.720	0.553	0.777	0.775	0.553
J48 (pruned)	0.810	0.710	0.571	0.802	0.802	0.603
Voted Perceptron (polynomial kernel exponent=2)	0.810	0.710	0.574	0.791	0.790	0.581
C-SVC (Kernel function: Radial basis)	**0.820**	**0.740**	**0.602**	0.760	0.758	0.519

performance is based on a reliable classification success beyond chance due to the high Kappa scores where the highest was observed as 72%.

This paper also elaborates on a comparative discussion concerning anaphoric resolution on Turkish child story texts and tweet texts. Tweets are short, composed of a few sentences and they are grammatically weak or incorrect when compared to child story texts. One of the remarkable points in our research is, comparing these two extensively different types of texts where the first one is well-structured and grammatically proven long texts and the second one is short, grammatically disproven, and consists of some misspelled texts. Contrary to our prior knowledge and experience, the results intriguingly showed that classification performance with the tweet data could be better than the child stories in Turkish. This can be concluded as one of the most important findings and contributions to this study. The results show that although tweets are short and grammatically weak, the overall expressive power of the several machine learning algorithms is strong and successfully finds the real antecedents of a pronoun within tweet texts. It should be noted that our findings in this study are within the scope of Turkish language and similar studies might be carried out among different languages in the future.

One of our plans is to observe the changes in classifier algorithms' performance by using alternative features obtained by several feature selection methodologies and to compare the performance of some other types of machine learning algorithms such as rule-based classifiers, and ensemble learners with the algorithms used in this work by using a larger data set with more instances.

References

[1] Mitkov R. "Evaluating anaphora resolution approaches," In: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2), Lancaster, UK, August 1998, pp. 164-172.

[2] Mitkov R. "Anaphora Resolution: The State of the Art," Technical Report, University of Wolverhampton, UK, 1999.

[3] Kornfilt J. "Turkish," New York: Routledge, 1997.

[4] Yıldırım S and Kılıçaslan Y. "A machine learning approach to personal pronoun resolution in Turkish," In: Proceedings of 20th International FLAIRS Conference, FLAIRS-20, Key West, Florida, USA, 7-9 May 2007, pp. 269-270.

[5] Kılıçaslan Y et al. "Learning-based pronoun resolution for Turkish with a comparative evaluation," Computer Speech and Language 2009; 23(3): 311-331.

[6] Quasim I et al. "Concept map construction from text documents using affinity propagation," Journal of Information Science 2013; 39(6): 719-736.

[7] Lubani M et al. "Ontology population: Approaches and design aspects," Journal of Information Science 2018; 45(4): 502-515.

[8] Hoste V and Daelemans W. "Comparing learning approaches to coreference resolution: There is more to it than 'bias'," In: Proceedings of the ICML-2005 Workshop on Meta-Learning, Bonn, Germany, 2005, pp. 20–27.

[9] Bickel B. "Referential density in discourse and syntactic typology," Language 2003; 79: 708–736.

[10] Cunnings I et al. "Anaphora resolution and reanalysis during L2 sentence processing: evidence from the visual world paradigm," Studies in Second Language Acquisition 2017; 39(4): 621-652.

[11] Stonic U et al. "Discourse and logical form: pronouns, attention and coherence," Linguistics and Philosophy 2017; 40(5): 519–547.

[12] Dhole K and Kohli H. "Document Categorization Using Semantic Relatedness & Anaphora Resolution: A Discussion," In: Proceedings of IEEE International Conference on Research in Computational Intelligence and Communication Networks, Kolkata, West Bengal, India, 20-22 November 2015, pp. 439-443.

[13] Othman R et al. "Towards Using Public Conversations to Mine Product Features in Twitter," In: Proceedings of IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, Tunisia, 2017, pp. 966-972.

[14] Erguvanlı T. E. "Pronominal vs. zero representation of anaphora in Turkish," In: Slobin, D. I. and Zimmer, K (eds) Studies in Turkish Linguistics. Amsterdam, Holland, 1986, pp. 206–233.

[15] Turan U. D. "Null vs. Overt Subjects in Turkish Discourse: A Centering Analysis," Ph.D. Thesis, University of Pennsylvania, USA, 1996.

[16] Grosz B et al. "Providing a unified account of definite noun phrases in discourse," In: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, USA, 15-17 June 1983, pp. 44-50.

[17] Grosz B and Ziv Y. "Centering, global focus, and right dislocation," Harvard University, USA, 1998, pp. 1-14.

[18] Grosz B et al. "Centering: a framework for modeling the local coherence of discourse," Computational Linguistics 1995; 2(2): 203–225.

[19] Gracanin-Yüksek M et al. "The Interaction of Contextual and Syntactic Information in the Processing of Turkish Anaphors," Journal of Psycholinguist Res 2017; 46(6): 1397–1425.

[20] Kılıçaslan Y. "Syntax of information structure in Turkish," Linguistics 2004; 42(4): 717–765.

[21] Banfield A. "Unspeakable Sentences: Narration and Representation in the Language of Fiction," London: Routledge & Kegan Paul, 1982.

[22] Tabrizi A. A. et al. “A Rule-Based Approach for Pronoun Extraction and Pronoun Mapping in Pronominal Anaphora Resolution of Quran English Translations,” Malay. Journ. of Comp. Sci. 2016; 29(3): 207-224.

[23] Singla D and Kumar P. “Rule Based Anaphora Resolution in Hindi,” In: Proceedings of IEEE International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2-3 June 2017, pp: 1-5.

[24] Kawasaki T and Kimura M. “A Novel Japanese Anaphora Resolution Method Using Deep Cases,” In: Proceedings of IEEE International Symposium on Computer Science and Intelligent Controls, Budapest, 2017, pp. 129-134.

[25] McCarthy J. F. and Lehnert W. G. “Using decision trees for coreference resolution,” In: Proceedings of International Joint Conference on AI, Palais de Congres Montreal, Quebec, Canada, 1995, pp. 1050–1055.

[26] Aone C and Bennet S. W. “Evaluating automated and manual acquisition of anaphora resolution rules,” In: Proceedings of 33rd Meeting of the Association for Computational Linguistics, USA, 1995, pp. 122–129.

[27] Yang X et al. “Coreference resolution using competition learning approach,” In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 7-12 July 2003, pp. 176–183.

[28] McCarthy J. A. “Trainable Approach to Coreference Resolution for Information Extraction,” Ph.D. Thesis, Dept. of Comp. Sci., Univ. of Massachusetts, USA, 1996.

[29] Cardie C and Wagstaff K. “Noun phrase coreference as clustering,” In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. University of Maryland, College Park, MD, USA, 21-22 June 1999, pp. 82–89.

[30] Ng V and Cardie C. “Improving machine learning approaches to coreference resolution,” In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Univ. of Pennsylvania, Philadelphia, USA, 7-12 July 2002, pp. 104–111.

[31] Trouilleux F. “A rule-based pronoun resolution system for French,” In: Proceedings of the Fourth Discourse Anaphora and Anaphora Resolution Colloquium (DAARC’02). Lisbon, Portugal, 18-20 September 2002.

[32] Witten I. H. and Frank E. “Data Mining: Practical Machine Learning Tools and Techniques,” 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[33] Weka 3: Data Mining Software in Java. Machine Learning Group at the University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/, 2018.

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

[35] John G. H. and Langley P. “Estimating Continuous Distributions in Bayesian Classifiers,” In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, CA, USA, 1995, pp. 338-345.

[36] Aha D. W. et al. “Instance-based learning algorithms,” Machine Learning 1991; 6(1): 37-66.

[37] Freund Y and Schapire R. E. “Large margin classification using the Perceptron algorithm,” Machine Learning 1999; 37(3): 277-296.

[38] Chang C. C. and Lin C. J. “LIBSVM – A Library for Support Vector Machines,” http://www. csie.ntu.edu.tw/ ˜cjlin/libsvm, 2004.

[39] Cohen J. A. “Coefficient of agreement for nominal scales,” Educational and Psychological Measurement 1960; 20(1): 37-46.

[40] Landis J. R. and Koch G. G. “The measurement of observer agreement for categorical data,” Biometrics, 1977; 33(1): 159–174.

[41] Ho T. K. et al. “Measures of Geometrical Complexity in Classification Problems,” In: Basu M and Ho T. K. (eds) Data Complexity in Pattern Recognition. London: Springer-Verlag, 2006.

[42] Gračanin-Yuksek, Martina, et al. "The interpretation of syntactically unconstrained anaphors in Turkish heritage speakers." Second Language Research 36.4 (2020): 475-501.

[43] Özge, Duygu, and Ebru Evcen. "Referential form, word order and emotional valence in Turkish pronoun resolution in physical contact events." Discourse Meaning: The View from Turkish 341 (2020): 165.

[44] Atkinson J. and Escudero A. "Evolutionary Discovery of Natural-Language Coreference Chains for Social Media Analysis." (2021).

[45] Wongkoblap A., Vadillo M. A., and Curcin A. "Deep Learning With Anaphora Resolution for the Detection of Tweeters With Depression: Algorithm Development and Validation Study." JMIR Mental Health 8.8 (2021): e19824.