# MAKİNE ÖĞRENME ALGORİTMALARIYLA TÜRKÇE SÖZCÜK ANLAMI AÇIKLAŞTIRMA

**Zeynep ORHAN**[1] **Zeynep ALTAN**[2]

[1,2]Bilgisayar Mühendisliği Bölümü
Mühendislik Fakültesi
İstanbul Üniversitesi, 34850, Avcılar, İstanbul
[1]e-posta: zeyneporhan@yahoo.com [2] e-posta: zaltan@istanbul.edu.tr

*Anahtar sözcükler:Sözcük anlamı açıklaştırma, makine öğrenme algoritmaları*

## ABSTRACT

*In this paper we examine the word sense disambiguation problem and the difficulties of Turkish word sense disambiguation process. We present our research results obtained from Turkish texts by using machine learning approaches. METU-Sabancı Turkish Treebank that is provided in XML format is used as our text and sense tags are annotated manually. Then WEKA system which is a collection of machine learning algorithms for data mining tasks, is used for sense disambiguation. The selected suitable algorithms are: J48(C4.5) that is a decision tree algoritm, IBk and KStar that are exemplar-based algorithms and AODE that is an improved version of Naïve Bayes statistical algorithm.*

## 1. GİRİŞ

Sözcük anlamı açıklaştırma (SAA) doğal dil işleme çalışmalrına paralel olarak 1950'li yıllarda ilgi çeken ve araştırılmaya başlayan bir konudur. SAA, hesaplamalı dilbilim alanında oldukça ilgi görmüş ve hakkında pek çok araştırma yapılmış, literatürde yer alan önemli bir problemdir. Ancak SAA alanındaki çalışmalar henüz tam olarak bir olgunluk kazanmamıştır, hatta tanımında bile farklılıklar bulunmaktadır. SAA problemini genel olarak ele alacak olursak şunları söylemek mümkündür: Pek çok sözcüğün birden fazla anlamı vardır. Herhangi bir yerde kullanılmasına bakılmaksızın bu sözcüğün hangi anlamda olduğu konusunda genellikle bir belirsizlik bulunur. SAA araştırma çalışmalarındaki genel amaç normal bir sözlük, eşanlamlılar sözlüğü veya buna benzer bir kaynaktaki farklı anlamlar arasındaki belirsizliği çözümlemektir. Bir kişi, içinde anlam belirsizliği olan bir tümceyi anladığı zaman, belirsizliğe neden olan sözcüğün diğer anlamlarını elemiş ve o sözcüğün sadece bir anlamını göz önüne almıştır. İnsanoğlunun doğasında bu anlama işlemini gerçekleştiren bir sistem mevcuttur ve bu şekilde içinde anlam belirsizliği bulunan bir tümcenin anlaşılması demek "*insan - dil anlama işleminde muhtemel anlam kümesi içinden uygun anlamın seçilmesi"* demektir [1].

SAA , insan - dil anlama işleminin bir parçası olarak tanımlanır ve bu işlem bir SAA programı içerisinde modellenebilirse, pek çok uygulama için gerekli olan önemli bir fonksiyon yerine getirilmiş olur. Bu görüş Cottrell [2] tarafından şu tümce ile çok net bir şekilde ifade edilmiştir:

*"Sözcüksel belirsizlik Doğal Dil Anlama (DDA) sistemlerinin karşılaştığı en önemli problemlerdendir. Kullanılan sözcüklerin doğru anlamını anlama DDA'nın temelidir. İnsanların belirsizliği çözme mekanizmasını anlamak çok önemlidir; çünkü bu işi yapma yöntemleri ne olursa olsun gerçekten çok başarılı olmaktadır."*

Kısaca ifade etmek gerekirse SAA, belirli bir kullanım alanında anlam belirsizliği olan sözcüğün hangi anlamının seçildiğini, kullanıldığı tümcedeki diğer elemanları da göz önüne alarak tespit etmektir. Bir sözcüğün bir kaynakta belirtilmiş sonlu sayıda farklı anlamı vardır ve bir SAA programının işi kullanıldığı yere göre bu sözcüğün bir anlamını, daha doğrusu en uygun anlamını seçmektir.

SAA alanındaki çalışmalar özellikle İngilizce üzerinde yoğunlaşmıştır. Çince, İtalyanca, İspanyolca gibi bazı dillerde de çalışmalar yapılmaktadır. Ancak Türkçe için bu alanda şimdiye kadar çok fazla bir çalışma yapılmamıştır. Geleceğin dünyasının belirlenmesinde önemli bir yapı taşı olduğu kesin olan doğal dil işleme teknolojilerini, diğer teknik gelişmelerden ayıran önemli bir özellik vardır. Herhangi bir dilde yapılan araştırmalarda o dili anadili olarak kullananlar (veya çok iyi bilenler) tarafından yerel bir uyarlama yapılması gerekli olabilir. Yani Türkçe için Türk dilini konuşanların geliştireceği sistemler çok daha verimli olacaktır. Dünyada değişik lehçelerini de hesaba kattığımızda yaklaşık üçyüz milyon kişinin Türkçe konuştuğu düşünüldüğünde, Türkçe'nin işlenmesinde yapılacak işlerin yoğunluğu ve bu çalışmaların sonuçlarının getireceği olumlu gelişmeler heyecan vericidir.

Bu çalışmanın da amacı, Türkçe metinlerde anlam belirsizliği olan sözcüklerin anlamlarının bilgisayar yardımıyla açıklaştırılmasını sağlamadır. Bundan sonraki bölümlerde Türkçe SAA için yapılan çalışmanın adımları anlatılmıştır.

## 2. TÜRKÇE SÖZCÜK ANLAMI AÇIKLAŞTIRMA ÇALIŞMASI

Türkçe sözcük anlamı açıklaştırma (SAA) çalışması iki temel kısımdan oluşmaktadır. Birinci kısım çalışmalarda kullanılacak olan metinlerin bulunması, gerekli işaretlemelerin yapılması, ikinci kısım ise uygun algoritmaların işaretlenen metinler üzerinde denenmesidir. Bundan sonraki bölümlerde yukarıdaki iki temel işlem için çalışmada şimdiye kadar yapılan araştırmalar ile ilgili detaylı bilgi verilmektedir.

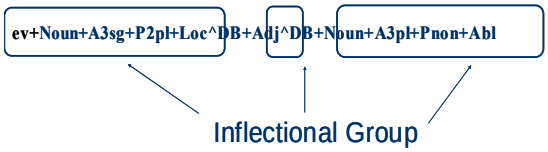
### 2.1. KULLANILACAK METİNLERİN BULUNMASI

ODTÜ-Sabancı[3] Türkçe ağaç bankası XML dosyaları biçimindedir (Şekil 1). Dosyalarda farklı dökümanlardan alınmış tümceler bulunmaktadır [4]. XML dosyaları SAA çalışmamızın girdi verilerini oluşturmaktadır. Bu dosyalarda bulunan tümce veya tümceler biçimbirimsel analizden geçirilmiş, biçimbirimsel analiz sonuçlarındaki belirsizlik çözümlenmiş ve tümcelerin öğeleri ilişkisel bir yapı ile gösterilmiştir.

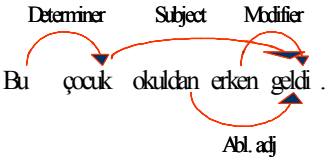
Sözcük düzeyinde işaretlemede sözcükler IG' lerden (inflectional groups- çekimli gruplar) oluşan birimler olarak kabul edilmiştir. Örneğin:

**Lemma+Infl1^DB+Infl2^DB+…^DB+Infln**

***evinizdekilerden***




Tümce düzeyinde işaretlemede tümce öğeleri arasındaki ilişkiler gösterilmiştir.




```
<?xml version="1.0" encoding="windows-1254" ?>
 - <Set sentences="1">
  - <S No="1">
    <W IX="1" LEM="" MORPH=""
       IG="[(1,"soğuk+Adj")(2,"Adv+Ly")]"
       REL="[2,1,(MODIFIER)]">Soğukça</W>
    <W IX="2" LEM="" MORPH=""
       IG="[(1,"yanıtla+Verb+Pos+Past+A1sg")]"
       REL="[3,1,(SENTENCE)]">yanıtladım</W>
    <W IX="3" LEM="" MORPH="" IG="[(1,".+Punc")]"
       REL="[,( )]">.</W>
  </S>
 </Set>
```

**Şekil 1:ODTÜ-Sabancı Ağaç bankası XML dosya biçimi**

Bu veri Java programı ile alınıp tüm tümceler veya verilen kök sözcüğü içeren tümcelerle birlikte bu tümcelere ait bazı özellikler de elde edilebilmektedir. Sistemin kullandığı özellikler arasında tümce numarası, incelenen sözcükle ilişkili olan diğer sözcüklerin kökleri, ekleri, tipi, iki sözcük arasındaki ilişki, incelenen sözcüğün tipi, kendisinin ilişkide olduğu diğer sözcükle aralarındaki ilişki vs. bulunmaktadır. Bu özellikler kullanılarak WEKA [5] sisteminde kullanılabilecek girdi dosyaları ARFF biçimine göre hazırlanmaktadır (Tablo 1). Bu dosyadaki **?** ile gösterilen değerler metinden bu değerlerin (problemli durumlar veya bu değerlerin olmaması nedeniyle) alınamadığını göstermektedir. ARFF biçiminde incelenecek yapıya bir isim verilmekte (*@relation kat*: kat kökü inceleniyor), daha sonra seçilen bütün özelliklerin alabilecekleri değerler kümesi ve veri sıralanmaktadır.

**Tablo 1: WEKA sisteminde kullanılacak ARFF girdi biçimi**

```
@relation kat

@attribute vectors0 {1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20}
@attribute vectors1 {BİNA, YUKARI, BODRUM, İKİ,
DUYGU, SES, HAYAT, KAÇINCI, HAN, ÜST, ASMA,
ALTINCI, HARAM, SU, ÇİFT, ARA, BU, DEĞİL, BEŞ}
@attribute vectors2 {NOUN, ADJ, NUM, PRON, CONJ}
@attribute vectors3 {GEN, NOM, CARD, DAT, ORD, ACC}
@attribute vectors4 {POSSESSOR, MODIFIER, CLASSIFIER,
SUBJECT, DATIVE.ADJUNCT, OBJECT,
NEGATIVE.PARTICLE}
@attribute vectors5 {ADJ, NOUN, VERB}
@attribute vectors6 {MODIFIER, OBJECT, SUBJECT,
DATIVE.ADJUNCT, LOCATIVE.ADJUNCT, SENTENCE,
ETOL, CLASSIFIER}
@attribute vectors7 {GENELEV, KAPA, KAHVE, ÇIK, GİR,
TONLA, KADIN, ODA, VAR, DOLU, SEVGİLİ, BATTANİYE,
DERİ, ET, ŞEKİL, DEĞER, KADAR}
@attribute vectors8 {NOUN, ADJ, VERB, POSTP}
@attribute vectors9 {NOM, ZERO, DAT, LOC, A3SG, GEN,
A2SG, ABL, A1PL, PCDAT}
@attribute vectors10 {SENTENCE, MODIFIER, OBJECT,
LOCATIVE.ADJUNCT, SUBJECT, POSSESSOR, PROBLEM,
APPOSITION}

@data
1, BİNA, NOUN, GEN, POSSESSOR, ADJ, MODIFIER,
GENELEV, NOUN, NOM, SENTENCE, 1
1, YUKARI, ADJ, ?, MODIFIER, ADJ, MODIFIER, GENELEV,
NOUN, NOM, SENTENCE, 1
2, BODRUM, NOUN, NOM, CLASSIFIER, NOUN, OBJECT,
KAPA, ADJ, ZERO, MODIFIER, 1
.
.
.
```

### 2.2. SÖZCÜK ANLAMLARININ OLUŞTURULMASI

Sözcük anlamı açıklaştırma çalışmasında kullanılacak anlamların oluşturulmasında oldukça uzun bir çalışma yapılmıştır. Türkçe sözcüklerin ortalama anlam sayısı bu alanda üzerinde çalışılan İngilizce gibi dillere göre çok daha fazladır.

Sözlüklerde bulunan anlamlar daha çok kullanımları sıralamakta ancak bir anlam sınıflandırması yapmamaktadır. Sözcüklerin asıl kullanımları yanında deyimsel kullanımlar, bileşik sözcükler, atasözleri vs. de ele alınınca anlam sayısı da paralel olarak artmaktadır. Bu yan kullanımlar da düşünülecek olursa bazı sözcüklerin anlam sayısı SAA çalışmasında incelenemeyecek kadar fazla olmaktadır. Buna ek olarak elimizdeki işlenmiş metinlerde bu

anlamların bazıları ya hiç geçmemekte, ya da çok az geçtiği için daha sonraki sınıflandırma aşamaları için yeterli veriyi sağlayamamaktadır. SAA alanında diğer dillerde yapılan çalışmalara bakıldığında da özellikle bu konudaki en kapsamlı değerlendirme projesi olan SENSEVAL projesinde [6] incelen sözcüklerin veya diğer dillerdeki sözcüklerin ortalama anlamının sekiz[1] civarında olduğu, deyimsel kullanışların bu çalışmalarda atıldığı ve kullanılan eğitim ve test verilerinin Türkçe için kullanılabilecek verilerden çok büyük olduğu düşünülecek olursa Türkçe için SAA'da kullanılacak sözcük anlamlarını oluşturma aşamasının zorluğu daha iyi anlaşılabilecektir.

Bu anlamları sınıflandırmak için çalışmamızda farklı yöntemler denemeye çalıştık. Bunlardan birincisinde, sözlüklerdeki anlamları baz alarak elimizde bulunan metinlerde geçen anlamlar için tüm anlamların bir alt kümesini oluşturduk. Bu işlem sırasında ilgili bazı anlamları genel bir başlık altında topladık, bazı anlamları da hiç kullanılmadığı için göz ardı ettik. Ancak sonuçta elde ettiğimiz anlamlar alt kümesi de bazı sözcükler için çok fazla sayıda anlam içerebiliyordu.

İkinci bir yol olarak çevirileri kullandık. SAA'da kullanılmak üzere seçilen sözcüklerin anlam sınıflandırmasını diğer dillerdeki farklı karşılıklarına göre sınıflandırma yoluna gittik. Ancak burada da bazı problemler karşımıza çıktı. İlk olarak bu yaklaşım SAA'nın kullanılabileceği diğer alanlara uymayabilirdi. Sözcüklerin farklı dillerde farklı karşılıkları bulunabiliyordu ve farklı sınıflandırmalar ortaya çıkabiliyordu. Ayrıca bazı sözcüklerin kullanımı veya anlamı sadece dilimize özgü olabiliyor ve diğer dillere çevirilse bile tam olarak karşılığı bulunmayabiliyordu.

Bütün bu zorluklar göz önünde bulundurularak daha önce yapılan anlam sınıflandırma işlemi tekrar ele alınıp yeni bir anlam işaretleme yöntemi seçilmiştir. Bu yeni yöntemde seçilen sözcük için aşamalı bir anlam sınıflandırılması yapılmıştır. Buna göre önce sözcüğün asıl anlamı ele alınıp asıl anlamdaki kullanımlara birinci anlam, bunun dışındaki diğer kullanımlara ise ikinci anlam denmiştir. Bu şekildeki sınıflandırma ile bazı algoritmalar denenerek sonuçları daha sonraki aşamalarda sınıflandırma için kullanılmıştır. Sonraki aşamalarda diğer kullanımlar kendi arasında tekrar sınıflanmış ve bu şekilde anlam sınıflaması bölünerek devam etmiştir. Aşamalı anlam sınıflandırması anlama etki eden faktörleri ayırt edebilmek için daha faydalı olacaktır.

WEKA sisteminde işaretlenen dosyaların dağılımları gözlenmiş ve anlam sınıflandırması çeşitli makine öğrenme algoritmaları ile denenmiş; elde edilen sonuçların hata analizi yapılmıştır. Çalışmamızda kullanılmak üzere makine öğrenme algoritmalarından karar ağacı (J48), örnek tabanlı yöntemler (IBk ve KStar) ve AODE(Aggregating one-dependence estimators) istatistiksel yöntemleri seçilmiştir [7]. Çalışmamızda bu yöntemler Türkçe SAA çalışması için işaretlenen metinler üzerinde denenmiş ve bu yaklaşımlardan hangisinin daha uygun olduğu araştırılmıştır.

# 3. SONUÇLAR

Bu bölümde gel sözcüğünün anlam incelemesi detaylı olarak anlatılacaktır. Diğer sözcüklere de aynı işlemler yapılmaktadır. *gel* sözcüğünün sırasıyla 2, 3 ve 4 anlam sınıfı kullanıldığı zaman elde edilen sonuçları çıkarılmıştır. Tablolarda birinci sütün kullanılan yöntemin adıdır. Test için *k–katlı çapraz doğrulama* (*k-fold cross validation-CV*) yöntemi kullanılmıştır. k-CV ile elde edilen sonuçlar hata içermektedir. Bunun nedeni ise örnek verinin az olması ve eğitimde kullanılan veri ile testte kullanılan verinin farklı anlamlar içerebilmesidir. Bölme işlemi farklı *k* değerleri için denenmiş ancak sonuçlar çok farklılık göstermemiştir.

Gel sözcüğü için iki, üç, dört anlam kullanıldığı zaman anlamlara göre dağılımlar Tablo 2'de verilmiştir.

**Tablo 2:Gel sözcüğünün anlamlara göre dağılımları**

| Anl | Oran | Anl | Oran | Anl | Oran |
|---|---|---|---|---|---|
| 1 | 0.59 | 1 | 0.59 | 1 | 0.59 |
| 2 | 0.41 | 2 | 0.14 | 2 | 0.10 |
| | | 11 | 0.27 | 5 | 0.04 |
| | | | | 11 | 0.27 |

Bu dağılımlarda birinci anlam olarak gösterilen anlam, gel sözcüğünün *gitmek, ulaşmak, varmak,* vs genel olarak kullanılan anlamıdır. İlk anlam sınıflamasında ikinci anlam ise bu anlam dışında kalan anlamları kapsamaktadır. Bu ikinci anlam daha sonraki aşamada kendi arasında tekrar gruplanmıştır. Deyimsel kullanışlara on birinci anlam denmiş diğerleri ikinci anlam olarak bırakılmıştır. Bir sonraki aşamada ise ikinci anlam içerisinden *bir şeyin zamanı, yeri sırası vs. gelmek* anlamı yeni bir anlam sınıfı olarak belirlenmiş ve ikinci anlam kendi arasında tekrar bölünmüştür. Bu analizler sırasında kullanılan ***"Confusion Matrix" (CM)*** yapılan tahminlerden elde edilen olası tüm sonuçları göstermektedir. CM bir sınıflandırıcının ne kadar iyi tahmin edebileceğini gösteren bir tablodur. Gel sözcüğünün sırasıyla iki, üç, dört anlamı için elde edilen CM değerleri Tablo 3, Tablo 4, Tablo 5'te verilmiştir. Elde edilebilecek olası sonuçlar Doğru pozitif (**TP**), Doğru Negatif (**TN**), Yanlış Pozitif (**FP**) ve Yanlış Negatif (**FN**) elemanlarından oluşmaktadır.

[1] SENSEVAL2 Lexical Sample verisinde (isimler için kullanılan) test ve eğitim için ortalama anlam sayısı 7.93'tür. Eğitim verisinde seçilen sözcüğün farklı anlamlarının geçtiği 3587 tümce, test verisinde ise 1773 tümce bulunmaktadır.

**Tablo 3:Gel sözcüğünün iki anlamlı sınıflandırmasında elde edilen Confusion Matrix değerleri**

| Algoritma | Anlam sayısı=2 |
|---|---|
| J48 | a b <-- classified as<br>288 25 \| a = 1<br>146 68 \| b = 2 |
| IBK | a b <-- classified as<br>275 38 \| a = 1<br>99 115 \| b = 2 |
| Kstar | a b <-- classified as<br>272 41 \| a = 1<br>94 120 \| b = 2 |
| AODE | a b <-- classified as<br>265 48 \| a = 1<br>92 122 \| b = 2 |

**Tablo 4:Gel sözcüğünün üç anlamlı sınıflandırmasında elde edilen Confusion Matrix değerleri**

| Algoritma | Anlam sayısı=3 |
|---|---|
| J48 | a b c <-- classified as<br>286 6 18 \| a = 1<br>49 22 4 \| b = 2<br>97 3 42 \| c = 11 |
| IBK | a b c <-- classified as<br>278 7 25 \| a = 1<br>36 26 13 \| b = 2<br>80 7 55 \| c = 11 |
| Kstar | a b c <-- classified as<br>278 6 26 \| a = 1<br>36 27 12 \| b = 2<br>76 6 60 \| c = 11 |
| AODE | a b c <-- classified as<br>281 4 25 \| a = 1<br>39 18 18 \| b = 2<br>72 3 67 \| c = 11 |

**Tablo 5:Gel sözcüğünün dört anlamlı sınıflandırmasında elde edilen Confusion Matrix değerleri**

| Algoritma | Anlam sayısı=3 |
|---|---|
| J48 | a b c d <-- classified as<br>287 4 1 18 \| a = 1<br>32 17 1 4 \| b = 2<br>15 1 5 0 \| c = 5<br>98 3 0 41 \| d = 11 |
| IBK | a b c d <-- classified as<br>277 6 1 26 \| a = 1<br>23 24 1 6 \| b = 2<br>11 1 5 4 \| c = 5<br>80 3 3 56 \| d = 11 |
| Kstar | a b c d <-- classified as<br>278 5 0 27 \| a = 1<br>22 25 1 6 \| b = 2<br>14 1 0 6 \| c = 5<br>76 3 2 61 \| d = 11 |
| AODE | a b c d <-- classified as<br>279 3 1 27 \| a = 1<br>30 13 1 10 \| b = 2<br>13 0 0 8 \| c = 5<br>73 2 0 67 \| d = 11 |

**Tablo 6:Algoritmaların doğru/yanlış sınıflandırma oranları**

| Algoritma | Anlam sayısı=2 | Anlam sayısı=3 | Anlam sayısı=4 |
|---|---|---|---|
| J48 | 67.5 | 65.0 | 66.0 |
| IBK | 74.0 | 69.4 | 68.7 |
| Kstar | 74.3 | 68.1 | 69.0 |
| AODE | 73.4 | 69.2 | 68.1 |

Tablo 6 algoritmaların iki, üç ve dört anlam sınıfı için elde ettikleri başarı yüzdelerini göstermektedir. Algoritmalardan IBk ve KStar genelde birbirine yakın sonuçlar vermektedir. Bu durum da her ikisinin de örnek tabanlı yöntemler olmasından ve temelde aynı esasları kullanmalarından kaynaklanmaktadır. Karar ağacı algoritması olan J48 diğerlerinden daha kötü sonuçlar vermiştir. İstatistiksel ve örnek tabanlı yöntemler Türkçe SAA çalışmaları için daha uygun görünmektedir.

## KAYNAKLAR


[1] Kilgarriff, A., 1999, I don't believe in word senses, Computers and the Humanities. B. (Eds.), English Corpus Linguistics.,Longman, London, 8-29.

[2] Cottrell, G. W.,1989, A Connectionist Approach to Word Sense Disambiguation. Research Notes in Artificial Intelligence. London: Pitman.

[3]ODTÜ Türkçe Derlem Projesi, http://www.ii.metu.edu.tr/~corpus/indextr.html

[4] Nart B. Atalay, Kemal Oflazer, Bilge Say, The Annotation Process in the Turkish treebank, in Proceedings of the EACL Workshop on Linguistically Interpreted Corpora-LINC, April 13-14, 2003, Budapest.

[5] WEKA system, http://www.cs.waikato.ac.nz/ml/weka

[6] Senseval Project, http://www.senseval.org

[7] Paliouras, G., Karkaletsis, V., Androutsopoulos, I., D. Spyropoulos, C., Learning Rules for Large-Vocabulary Word Sense Disambiguation: a comparison of various classifiers, Proceedings of the 2nd International Conference on Natural Language Processing (NLP 2000), Patra, Greece, D.N. Christodoulakis (Ed.). Lecture Notes in Artificial Intelligence, 1835, pp. 383 – 394, Springer, 2000.