CHAT (SOHBET) ORTAMLARINDAN BİLGİ ÇIKARIM ÖRNEĞİ: CİNSİYET BELİRLEMEYE YÖNELİK İSTATİSTİKSEL VE ANLAMSAL YAKLAŞIM

Özcan ÖZYURT[1] **Cemal KÖSE[2]**

[1]Beşikdüzü Meslek Yüksekokulu, Karadeniz Teknik Üniversitesi
Beşikdüzü, Trabzon. *e-posta: oozyurt@ktu.edu.tr*
[2]Bilgisayar Mühendisliği Bölümü, Karadeniz Teknik Üniversitesi
Trabzon. *e-posta: ckose@ktu.edu.tr*

Anahtar sözcükler: Bilgi Çıkarımı, Cinsiyet Belirleme, Chat (Sohbet) Sistemleri, Makine Öğrenmesi

ABSTRACT

Chat mediums are becoming an important part of human life in societies and provide quite useful information about people such as current interests, habits, social behaviours and tendency of the people. In this study, we have presented an identification system that is designed to identify the sex of a person in a Turkish chat medium. Here, the sex identification is taken as a base study in the information mining in chat mediums. This identification system acquires data from a chat medium, and then automatically detects the chatter's sex from the information exchanged between chatters and compares them with the known identities of the chatters. To do this task a simple discrimination function is proposed. A semantic analysis method is also applied to determine the sex of the chatters. The system with the semantic analyser has achieved over accuracy 90% in the sex identification in the real chat medium.

1. GİRİŞ

Günümüzde internet kullanımının yaygınlaşmasıyla birlikte, chat (sohbet) ortamları insanların hayatlarında önemli yer tutmaya başlamıştır. Bu ortamlar çoğu zaman bilgi paylaşım ve haberleşme gibi amaçlar için kullanılmaktadır [1,2,3]. Chat ortamlarındaki konuşmalar, konuşmacıların alışkanlıkları, eğilimleri, ilgi alanları ve cinsiyetleri gibi birçok faktöre göre şekillenebilmektedir. Chat ortamlarından, bilgi çıkarımı sayesinde konuşmacılar hakkında birçok bilgi elde edilebilir [1,4,5,6]. Bu çalışmada, internet ortamlarından bilgi çıkarımına bir örnek uygulama olarak, chat ortamlarındaki konuşmacıların cinsiyetlerinin belirlenmesine yönelik bir cinsiyet ayırt etme fonksiyonu önerilmiştir. Bunun için, internette yaygın olarak kullanılan chat programı mIRC (Microsoft İnternet Relay Chat) ve yerel ağda chat yapmak amacıyla geliştirilen chat programı kullanılarak çok sayıda konuşma incelenmiş ve bu konuşmalardan çeşitli istatistiksel bilgiler çıkarılmıştır. Bu bilgiler yardımıyla, bayan konuşmacılar ve erkek konuşmacılara ait tanımlayıcı kelimeler belirlenmeye çalışılmış ve bu kelimeler gruplandırılmıştır. Gruplandırılmış kelimeler, konuşmalar içerisinde erkek ve bayan konuşmacıların kullanım durumları göz önünde bulundurularak ağırlıklandırılmış ve cinsiyet ayırt etme işlemi için kullanılmıştır.

Konuşmalardan elde edilen istatistiksel veriler kullanılarak, kelimeler "argo kelimeler", "nida-seslenme kelimeleri", "kısaltma grupları" gibi gruplar altında toplanmıştır. Her bir grup içindeki kelimelerin kullanım sıklıkları kullanılarak erkek-bayan konuşmacılar için belirleyicilik katsayıları belirlenmiştir. Cinsiyet belirlemede, istatistiksel veriler ve kelimelerin ağırlıklarının yanı sıra, anlamsal analize yönelik testler de yapılmış ve cinsiyet belirleme işleminde başarım artırılmıştır.

Geliştirilen yöntem yardımıyla, mIRC ve yerel ağ ortamlarında yapılan konuşmalar test edilmiş ve sistemin doğruluk derecesinin %90'lara ulaştığı görülmüştür.

2. KELİME GRUPLARI

Chat ortamlarından elde edilen konuşmalar değerlendirilmiş ve bu konuşmalarda sıkça kullanılan kelimelerin istatistikleri çıkarılmıştır. Bu konuşmalardan elde edilen verilerden, bazı kelimelerin erkek konuşmacılar tarafından, bazılarının da bayan konuşmacılar tarafından daha sık kullanıldıkları görülmüştür. Bu kelimelerin tespit edilmesinden sonra, kelimelerin kavramsal ilişkileri göz önünde bulundurulmuştur. Bu temel üzerine, konuşmalarda sık kullanılan kelimeler belirli gruplar altında toplanmıştır. Bu gruplar "argo kelimeler

grubu", "nida-seslenme kelimeleri grubu", "kısaltmalar grubu" gibi 8 başlık altında toplanmıştır. Bu gruplar oluşturulurken, kelimelerin birbirleriyle yakınlıkları ve ilişkileri göz önünde bulundurulmuş ve hangi gruba dahil edileceklerine karar verilmiştir. Aşağıdaki tabloda, oluşturulan gruplar ve bu gruplarda yer alan kelimelere ilişkin örnekler verilmiştir.

Tablo-1. Kelime Grupları ve Gruplara ait örnek kelimeler

No	Kısaltmalar Grubu	Argo Kelimeler Grubu	Nezaket Kelimeler Grubu	Nida ve Seslenme Kelimeler Grubu
1	Slm (Selam)	Oğlum!	Güzel	Yahu (ya,yav,yaw,yaws)
2	Cvp (Cevap)	Lan!	Tşk	Hımm (hımm, hımms)
3	Nbr (Naber)	Dayı!	Aferin	Ee (e,ee,eee)
4	U (Sen)	Defol!	Efendim	Aa (a,aa,aaa)
5	Tşk (Teşekkür)	Tövbe!	Siz	İi (iyi,ii,iii)
6	Msj (Mesaj)	Baba!	Canım	Alo (alo,aloo,alloo)

No	Yaş ve Cinsiyetle İlgili Kelimeler Grubu	Soru Kelimeleri Grubu	Edat ve Bağlaç Kelimeleri Grubu	Diğer Kelimeler Grubu
1	Yaş	Niye	Öyle	Sen
2	Cinsiyet	Neden	Yoksa	Ben
3	Aşkım	Hangi	Diye	Olsun
4	Bayanım	Nerde	Başka	Seni
5	Erkeğim	Nerdesin	Böyle	Bak
6	Manita	Kimsin	Şimdi	Yanlış

3.KELİME AĞIRLIKLANDIRMA VE CİNSİYET BELİRLEME

mIRC ve Yerel Chat ortamından elde edilen konuşmalardan, erkek ve bayan konuşmacıların hangi kelimeleri daha sık kullandıkları belirlenmiştir. Diğer bir deyişle, hangi kelimelerin erkekler tarafından, hangilerinin de bayanlar tarafından daha sık kullanıldığı bilgileri istatistikler sonucu çıkarılmıştır. Bu kelimeler, cinsiyet belirleme fonksiyonunun tanımlanmasında kullanılmıştır. Kelimeler, kullanım sıklıklarına göre ağırlıklandırılmıştır. Herhangi bir kelimenin ağırlığı

$$\alpha = x_m / (x_m + x_f) \quad (1)$$

şeklinde hesaplanır. Burada, x_m herhangi bir x kelimesinin erkekler tarafından kullanım sıklığı, x_f herhangi bir x kelimesinin bayanlar tarafından kullanım sıklığı ve α herhangi bir x kelimesinin belirleyicilik katsayısıdır. (1)'e göre, herhangi bir kelimenin ağırlık katsayısı 0-1 arasında normalize edilmiş olur. Buna göre, herhangi bir kelime erkekler için belirleyici ise α değeri 1'e yakın olacaktır. Diğer bir deyişle, herhangi bir kelimenin bayanlar tarafından kullanımı daha fazla ise, o kelimenin bayanlar için belirleyicilik katsayısı da 0'a yakın olacaktır.

Kelimelerin ağırlıkları belirlendikten sonra, konuşmalar içerisinde bu kelimelerin geçip geçmediği kontrol edilmiştir. Eğer herhangi bir kelime konuşma içinde geçmişse bu kelime için belirleyicilik katsayısı dikkate alınır ve o konuşma için hesaba katılır. Konuşma içerisinde aynı kelimenin birden fazla geçmesi durumunda, kelimenin erkek ya da bayan konuşmacıdan hangisi için belirleyicilik özelliği varsa, o değer daha da güçlendirilir. Örneğin, herhangi bir x_m kelimesinin erkekler için belirleyicilik katsayısı 0.7 olsun. Bu kelimenin konuşma içerisinde geçme sayısı birden fazla ise, bu kelime için ağırlık katsayısı her bir fazla kullanım için 0,025 oranında artırılmıştır. Bu artırım 3 kez yapılmış ve bundan fazla kullanım göz ardı edilmiştir. Eğer kelimenin erkekler için belirleyicilik katsayısı düşükse, diğer bir deyişle bayanlar için belirleyicilik katsayısı yüksekse (α değeri < 0,5 ise), bu sefer aynı oranda katsayıda azaltma yapılmış ve bayanlar için belirleyicilik katsayısı artırılmıştır.

Herhangi bir konuşma için konuşmacıların cinsiyetlerinin belirlenmesinde, fonksiyon aşağıdaki gibi oluşturulmuştur:

Herhangi bir konuşma içerisinde, belirleyici kelimelerden a_1, a_2 ... a_n kelimelerinin var olduğu kabul edilsin. Buna göre, herhangi bir konuşma için

$$\alpha_{Konuşma} = (\alpha_{(a1)} + \alpha_{(a2)} + \ldots + \alpha_{(an)})/ n \quad (2)$$

değeri hesaplanır. Buna göre, konuşma için çıkan $\alpha_{Konuşma}$ katsayısı, aynı şekilde 0-1 arasına düşürülmüş olur. $\alpha_{Konuşma}$ katsayısı 1'e yakınsa bu konuşma erkek ağırlıklı; 0'a yakınsa bayan ağırlıklı olarak değerlendirilir.

4. ANLAMSAL YAKLAŞIMLAR

Cinsiyet belirlemede, istatistiksel bilgilerin yanında anlamsal analiz de kullanılarak sonuçlar güçlendirilebilir [3,6,7]. Konuşmalarda cinsiyet

sorma, seslenme deyim gibi kalıpların kullanılması, o konuşmadaki konuşmacıların cinsiyetleri hakkında bilgi verebilir. Selamlama ve ad-adres sorma cümlelerinden "ben ali", "bende yavuz", "adım hasan" gibi diyaloglardan karşı tarafın veya konuşmacının cinsiyeti belirlenmektedir. Bunların yanı sıra, "nbr ahmet", "iyilik/iyidir hasan", "merhaba Mehmet" gibi seslenme ve selamlama deyimlerinden bu bilgiler çıkarılır. Burada, isimlerin erkek-bayan ismi olduğu bilinmekte ve buna göre davranılmaktadır. Örneğin, "nbr Murat" ifadesinde, "Murat" kelimesinin erkekler için kullanılan bir isim olduğu ve karşı taraftaki konuşmacının cinsiyetinin erkek olduğu bilgisine ulaşılır. "ismin neydi?", "sen kimsin?", "ismin ne?", "adın ne?", "isim" vb. gibi sorular ve bunlara verilen cevaplardan faydalanılır. Tablo-2 de, konuşma örneklerinden bir parça verilmiş ve bu konuşmada geçen yapılar analiz edilmiştir.

Tablo-2. mIRC konuşma örneği–1

	Konuşmacılar	**Cümleler**
...		
(1)	Vah	İst
(2)	Vah	Sen?
(3)	GeceMavisi	Trabzon
(4)	GeceMavisi	İsim
(5)	Vah	Güneş
(6)	GeceMavisi	Memnun oldum
(7)	GeceMavisi	Bende Yavuz
(8)	Vah	ii
...		

Tablo-3. mIRC konuşma örneği–2

	Konuşmacılar	**Cümleler**
...		
(1)	GencPrens	:)
(2)	GencPrens	Yoksa kafana göre birini bulamadın mı?
(3)	Merix	Kafama göre birini bulamadım
(4)	GencPrens	O zaman sen evde kalırsın bu gidişle
(5)	Merix	Zaten olmasında
…		

Selamlama, isim-adres sorma gibi ifadelerinin değerlendirilmesinden sonra, konuşma içinde geçen erkek-bayan karakteristik gösteren deyimlerin bulunması ve değerlendirilmesi işlemi yapılmıştır. Türkçe'de, erkek ve bayanlar için kullanılan deyim ve ifadelerin varlığı bilinmekte ve bunlar göz önüne alınmaktadır. Bu deyim ve ifadelere örnek olarak, "ev kızıyım", "istediğimi giyiyorum", "çok tatlısın", "yakışıklıyım" gibi cümleler gösterilebilir. Bu ve benzeri cümle/deyimlerin değerlendirilmesi yapılmıştır.

Bu cümle örneklerinden birini içeren konuşmanın bir örneği Tablo-3'te verilmiştir.

Bu cümleler değerlendirilirken, cümlelerin morfolojik (biçimbirimsel) analizleri yapılmış ve kelimelerin aldıkları ekler değerlendirilerek hangi şahsa ait oldukları tespit edilmiştir. "yok ev kızıyım" cümlesinde, "ev kızı" deyiminin bayanlar tarafından kullanıldığı bilinmektedir. Bu deyim belirlendikten sonra, "kızıyım" kelimesinin morfolojik analizi sonucu -(y)ım ekinin 1. Tekil Şahıs sahiplik eki olduğu görülür. Böylelikle konuşmada "ev kızıyım" ifadesinin söyleyen kişiye ait olduğu anlaşılır. Buradan bu cümleyi kullanan kişinin bayan olduğu kanısına varılır. Benzer şekilde, "yakışıklısın" ifadesi için karşılaşıldığında, bu ifade "yakışıklı-sın" olarak ayrıştırılır [8]. "-sın" eki, 2. Tekil Şahıs eki olduğu için yakışıklı ifadesinin karşı taraftaki konuşmacı için söylendiği anlaşılmaktadır. "yakışıklı" ifadesi erkekler için kullanılan bir ifade olduğu için karşı taraftaki konuşmacının erkek olduğu kanısına varılır.

Anlamsal analizler yapılırken, konuşma içerisinde erkek ya da bayanlar tatarından sıkça kullanılan deyimler/ifadeler ve konuşmacıların birbirlerine seslenme-hitap ifadelerinden faydalanılmıştır. Anlamsal analiz sonuçlarının eklenmesiyle, cinsiyet ayırt etme fonksiyonu güçlendirilebilir. Diğer bir ifadeyle, anlamsal analizin sonuçları cinsiyet ayırt etme fonksiyonunun geliştirilmesinde kullanılır. Herhangi bir konuşma içerisinde, erkekler için belirleyicilik özelliği tespit edildiğinde, bunun için $\alpha_{anlamsal}$ değeri 1 olarak alınır. Bu durumda, bu ifadenin bayanlar için belirleyicilik katsayısı $\alpha_{anlamsal}$ 0 olacaktır. Aynı şekilde, konuşma içerisinde bayanlar için belirleyicilik özelliği içeren bir ifade tespit edildiği zaman, $\alpha_{anlamsal}$ değeri 0 olarak alınır.

Herhangi bir konuşma için $\alpha_{anlamsal}$ değeri,

$$\alpha_{anlamsal} = (\alpha_{anlamsal\text{-}1} + \alpha_{anlamsal\text{-}2} + \ldots + \alpha_{anlamsal\text{-}n})/n \quad (3)$$

olarak hesaplanır. Burada n değeri, konuşma içerisindeki anlamsal analiz sonucu bulunan belirleyicilik özellik sayısıdır.

Bu durumda, konuşma için belirleyicilik katsayısı olan $\alpha_{Konuşma}$ aşağıdaki gibi yeniden hesaplanır:

$$\alpha_{Konuşma} = (\alpha_{Konuşma} + \alpha_{anlamsal}) / 2 \quad (4)$$

şeklinde yeniden tanımlanır.

Herhangi bir konuşma için $\alpha_{Konuşma}$ değeri belirlendikten sonra, bu değere göre konuşmacının cinsiyetine karar verilir. Bu değer, 0 – 1 arasına normalize edilmiş bir değerdir. Bu değer 1' e ne kadar yakınsa o konuşmacının erkek olduğu kanısına varılır. Diğer bir ifadeyle, herhangi bir konuşma için elde edilen $\alpha_{Konuşma}$ değeri, bayan konuşmacı için 0'a yakın, erkek konuşmacı için 1'e yakın çıkacaktır. Aradaki değerler için aşağıdaki kararlar verilmektedir:

$\alpha_{Konuşma}$ >= 0.75 ise, “Konuşmacı Erkektir”;
0.75 > $\alpha_{Konuşma}$ >= 0.60 ise, “Konuşmacı Muhtemelen Erkektir”;
0.60 > $\alpha_{Konuşma}$ > 0.52 ise, “Konuşmacı Erkek olabilir”;
0.52 >= $\alpha_{Konuşma}$ >=0.48 ise, “Konuşmacının cinsiyeti hakkında karar verilemiyor”;
0.48 > $\alpha_{Konuşma}$ > 0.40 ise, “Konuşmacı Muhtemelen Bayandır”;
0.40 >= $\alpha_{Konuşma}$ > 0.25 ise, “Konuşmacı Bayan olabilir”;
0.25 >= $\alpha_{Konuşma}$ ise, “Konuşmacı Bayandır”

kararları verilir.

5. SONUÇ

Bu çalışmada, internet ortamlarından bilgi çıkarımının bir örneği olarak, bu ortamlardaki konuşmacıların cinsiyetlerinin belirlenmesine yönelik bir fonksiyon önerilmiştir. Bu fonksiyon yardımıyla, yerel ağda chat yapmak amacıyla geliştirilen chat programı yardımıyla elde edilen konuşmalar ve internet ortamında yaygın olarak kullanılan mIRC’den elde edilen konuşmalar teste tabi tutulmuş ve fonksiyonun sonuçları incelenmiştir. Tablo–4 ve Tablo-5’de yerel chat ve mIRC ortamlarından alınan konuşmalara ait test sonuçları verilmiştir. Önerilen fonksiyon ile cinsiyet ayırt etmede %90’lara varan başarının olduğu gözlenmiştir.

Tablo–4. Yerel Chat üzerindeki konuşmalar için cinsiyet ayırt etme fonksiyonunun sonuçları

	Erkek	Bayan
Konuşmacı Sayısı	54	44
Doğru Karar Sayısı	47	40
Yanlış Karar Sayısı	4	4
Karar Yok Sayısı	3	0
Doğru Karar Oranı	87.0%	90.9%
Yanlış Karar Oranı	7.5%	9.1%
Kararsız Oranı	5.5%	0.0%

Tablo–5. mIRC üzerindeki konuşmacılar için cinsiyet ayırt etme fonksiyonunun sonuçları

	Erkek	Bayan
Konuşmacı Sayısı	19	8
Doğru Karar Sayısı	17	7
Yanlış Karar Sayısı	2	1
Doğru Karar Oranı	%89.5	%87.5
Yanlış Karar Oranı	%10.5	%12.5

KAYNAKLAR

[1] Khan Faisal M., Fisher Todd A., Shuler Lori, Wu Tianhao and Pottenger William M., Mining Chat-room Conversations for Social and Semantic Interactions., Lehigh University Technical Report LU-CSE-02-011, 2002.

[2] Elnahrawy, E.: Log-Based Chat Room Monitoring Using Text Categorization: A Comparative Study. The International Conference on Information and Knowledge Sharing, US Virgin Islands (2002)

[3] Storey V. C., Chiang R., Chen G. L., Ontology creation: Extraction of domain knowledge from web documents., Lecture Notes in Computer Science – Conceptual Modeling– ER 2005, Vol. 3716, pp.256-269, 2005.

[4] Kaban A., Wang Xin., Context based identification of user communities from Internet chat., IEEE International Joint Conference on Neural Networks, Vol. 4, pp. 3287 – 3292, 2004.

[5] Iiritano S., Ruffolo M., Managing the knowledge contained in electronic documents: a clustering method for text mining., 12th International Workshop on Database and Expert Systems Applications, pp.454 – 458, 2001.

[6] Gao Xiaoying, Zhang Mengjie, Learning knowledge bases for information extraction from multiple text based Web sites, IEEE/WIC International Conference on Intelligent Agent Technology, pp.119 – 125, 2003.

[7] Lee M.C. and Leong R.V., NLUS - A Prolog Based Natural Language Understanding, System, In Proceeding of the 4th International Conference on Computing and Information, pp. 204-207, May 1992.

[8] Oflazer, K.: Two-level Description of Turkish Morphology. Literary and Linguistic Computing, Vol. 9. pp. 137-148 1994.