

Sosyal Ağımdaki Duygusal Uyum

Emotional Harmony in My Social Network

Kaan Ant¹, Banu Diri²

^{1,2}Bilgisayar Mühendisliği Bölümü, Yıldız Teknik Üniversitesi, İSTANBUL/TÜRKİYE

Email: ¹kaanantt@gmail.com, ²banu@ce.yildiz.edu.tr

Özetçe—Mikroblog günümüzde İnternet kullanıcıları arasında çok popüler bir iletişim aracı haline gelmiştir. Mikroblog sitelerinde yazılanlar doğal dil işleme çalışmaları için büyük bir veri madeni niteliği taşımaktadır. Bu platformlar üzerinde yapılan çalışmalar genellikle görüş analizi (olumlu, olumsuz, nötr) üzerine olup, metinlerdeki duyguyu tespit etmekte yetersiz kalmaktadır. Bu çalışmada doğal dil işlemenin metinlerde geçen duyguyu tanımda nasıl kullanılabileceği gösterilmiş ve Türkçe diline ait bir duygu tanıma aracı geliştirilmiştir. Twitter kullanıcılarının yazmış olduğu metinler alıntılanarak “mutluluk, korku, öfke, üzgün” duygularını barındıran tweetler bir vektör olarak ifade edilirken, “soldan – k” ve “ikili vektör” adı verilen yöntemler kullanılmıştır. İki yöntemde de “mutluluk” ve “korku” sınıfları yüksek başarımlarla tespit edilirken, “kızgın” ve “üzgün” sınıflarının başarımları bu iki duygunun yazım dilinde birbirine yakınlığı sebebiyle daha düşük olmuştur.

Anahtar Kelimeler—Doğal dil işleme; duygu analizi; twitter

Abstract— Microblog has become a very popular communication tool among internet users these days. Microblog sites are a great data mine for natural language processing. Studies on these platforms are insufficient to determine the emotion in texts based on opinion analysis (positive, negative, neutral). In this study, it was shown how natural language processing can be used in sentence recognition in texts and a tool for recognizing emotion in Turkish language has been developed. The texts written by Twitter users are quoted and tagged with tweets that contain feelings of "happiness, fear, anger, sadness". Labeling was done using "left -k" and "binary vector" methods. In both methods, the "happiness" and "fear"

classes are labeled with high achievements, but the achievements of the "anger" and "sad" classes are lower because of the closeness of the two senses.

Keywords—Natural language processing; emotion analysis, twitter

I. GİRİŞ

Mikroblog, günümüzde internet kullanıcıları arasında en fazla tercih edilen iletişim araçlarından biridir. Popüler web sitelerinde her gün milyonlarca mesaj görünmektedir. Twitter, Tumblr ve Facebook bu mikrobloglara verilebilecek örneklerdendir. Mikroblog yazarları, düşünceleri ile ilgili çeşitli konularda yorum yazmak ve güncel konuları tartışmak için yazılar paylaşmaktadır.

Ücretsiz mesaj biçimi ve kolay erişilebilirlik nedeniyle, mikroblog platformlarının geleneksel iletişim araçlarına (bloglar veya e-posta listeleri) kıyasla daha fazla kullanıcıya sahip olduğu bilinmektedir. Kullanıcıların her konuda bu platformlarda fikirlerini paylaşıyor olmaları, doğal dil işleme çalışmaları için büyük bir veri kaynağı oluşturmaktadır.

Bugüne kadar Twitter’den toplanan veriler ile doğal dil işleme ve veri madenciliği alanında pek çok çalışma yapılmıştır. Çalışmalara örnek olarak salgın hastalıkları önceden tahmin eden [1], ilaçların bilinmeyen yan etkilerini keşfeden [2], algının zamanla değişimini tahminleyen [3], turistik bir beldeye gelen turistlerin tweetleri üzerinden algı analizi yapan [4], içerisinde geçen konunun olumlu veya olumsuz düşünceden hangisini yansıttığını bulan [5], film yorumlarının değerlendirilmesini yapan [6] çalışmalar örnek olarak verilebilir.

Pak [7] çalışmasında Twitter üzerinden duygu analizi ve fikir madenciliği işlemlerinin gerçekleştirilmesi için bir derlem oluşturmuş ve bu derlem üzerinden bir duygu sınıflandırıcısı gerçekleştirmiştir. Cümleler

pozitif, negatif ve nötr olarak etiketlenmiştir. Bu çalışmada oluşturulan duygu sınıflandırıcısının, daha önceden kullanılan anlamsal analiz araçlarına göre daha verimli olduğu ve daha yüksek performans gösterdiği tespit edilmiştir. Çalışma İngilizce dili için yapılmış olsa da dilden bağımsız çalışabileceği belirtilmiştir.

[8] Metindeki duyguları tespit ederek bireyin anksiyete, depresyon veya refahının ölçülmesi için yapılmış bir çalışma olup, emojiler, noktalama işaretleri ve geniş duygu sözlüğünden yararlanılmış ve makine öğrenmesi yöntemleri kullanılarak sınıflandırılmış ve % 90'dan fazla başarı elde edilmiştir. Akgül [9], sözlük ve n-gram yöntemlerini kullanarak her tweeti olumlu, olumsuz ve nötr olarak sınıflandırmış ve sırasıyla %70 ve %69 başarı elde etmiştir.

Bu çalışmada insanların Twitter üzerinden yaptığı paylaşımların içerdiği duyguların tespiti ve tespit edilen duygular doğrultusunda kullanıcıya geri bildirim verilmesi hedeflenmiştir.

Makalenin ikinci bölümünde geliştirilen sistemin tasarımı, üçüncü bölümde deneysel çalışmalar ve son bölümde de sonuç yer almaktadır.

II. SİSTEM TASARIMI

Çalışma kapsamında ilk aşamada Twitter API ile veri kümesi oluşturulmuştur. İkinci aşamada oluşturulan veri kümesi üzerinden duygu analizi gerçekleştirilmiştir. Son olarak da duygu analizi sonuçlarına göre kullanıcılar hakkında bilgi çıkarımı yapılmış ve graf yardımıyla kullanıcıya geri bildirim verilmiştir.

A. Twitter API Modülü

Twitter API, Twitter üzerinde yapılan çalışmalar için Twitter tarafından oluşturulmuş bir uygulama geliştirme aracıdır. Bu aracın kullanımı için Python diline ait "Twitter" kütüphanesi kullanılmıştır. Twitter API üzerinde çalışabilmek için "uygulama anahtarı" ve "tüketici anahtarı" gerekmektedir. Çalışma kapsamında 'kaanantt' kullanıcısı merkezi kullanıcı olarak tanımlanmıştır. Veri setinin oluşturulabilmesi için bu kullanıcının arkadaş listesi ve bu listede bulunan kişilerin paylaşımlarına erişilmiştir. Veri kümesi oluştururken, içerisinde *Hashtag*, *URL* bulunan ve *Retweet* edilen mesajlar veri setine dahil edilmemiştir. Oluşturulan veri kümesi, JSON formatında içerisine kullanıcı bilgileri de eklenerek MongoDB'ye kaydedilmiştir.

B. Duygu sınıflarına göre etiketlenmiş tweet veri setinin oluşturulması

Duygu barındıran kelimelerin tespiti için yapılan kapsamlı araştırmalardan sonra, Kanada Ulusal Araştırma Konseyi üyesi Saif Muhammed [13] tarafından oluşturulan sınıflandırılmış İngilizce kelime veri seti Türkçe'ye tercüme edilerek her bir duygu sınıfı için sözlükler oluşturulmuştur.

Kelimelerin sıklık ve uyum dağılımlarını hesaplayabilmek için de iki milyon tweet kullanılmıştır [10]. Her iki veri kümesi de ön işleme (Türkçedeki zamir, edat gibi gereksiz kelimelerin temizlenmesi, kelimelerin ilk 5 harfinin kök olarak kabul edilmesi ve gereksiz boşlukların temizlenmesi) tabi tutulmuş ve "Utf-8" formatına uyarlanmıştır. Son olarak, başarıyı hesaplayabilmek için 800 tweet iki kişi tarafından elle etiketlenmiş, farklı etiket verilme durumunda yeniden kontrol edilerek homojen bir veri kümesi oluşturulmaya çalışılmıştır.

Kullanıcı	Tweet
ProfDemirtas	Bir projeye yatırım yaparken hangi methodları kullanıyorsunuz
Hakan_Gunday	Kinyas mı döver Kayra mı
EmrahSerbes	Çok üzgünüm Hepimizin başı sağolsun bütün insanlığın
busranze	Geçen yıl eylül ayından beri 3 haftada bir komiteye giriyorum yeminle bunaldım artık

Tablo I. Veri seti örnekleri

Bir tweetin duygu sınıfını belirlemeden önce, tweet içerisindeki kelimelerin sınıfları tespit edilmiştir. Bu aşamada, Zemberek kullanılarak kelimelerin kökleri tespit edilmeye çalışılmış ancak, her kelime için birden fazla kök değeri döndürüldüğü için "soldan-k" yöntemi kullanılmıştır. Türkçe ile ilgili çalışmalar incelendiğinde, bir kelimenin ilk 5 harfinin o kelimenin kökü olarak kabul edildiği çalışmalara rastlandığından [11, 12] *k* değeri 5 alınarak bir kelimenin kökü olarak belirlenmiştir.

III. DENEYSEL ÇALIŞMALAR

Sistem tasarımı bölümünde belirtildiği gibi, tweet'in içerdiği duyguyu tespit etmek için içerisinde yer alan kelimelerin *soldan-k* yöntemi ile kökleri tespit edilmiş ve kökler çoğunluk olarak hangi duygu sözlüğünde yer alıyorsa tweetin duygu sınıfı olarak etiket verilmiştir. Öncelikli olarak her sınıf etiketi için

yaklaşık 200 tweet alınarak veri seti oluşturulmuştur. *Soldan -k* yönteminin başarımını test etmek için bu veri seti elle dört sınıf için etiketlenmiştir. Veri seti oluşturulurken 170 *Kızgın*, 225 *Mutlu*, 148 *Üzgün*, 108 *Korku* sınıfına ait tweetler elle oluşturulmuş ve etiketlenmiştir. Etiketlenmiş bu veri seti sözlük kullanılarak duygusal olarak sınıflandırıldığında Tablo II'deki hata matrisi değerleri, Tablo III'te de her sınıf için F-Ölçüm değerleri verilmiştir.

Gerçek	Tahmin				
		Kızgın	Mutlu	Üzgün	Korku
	Kızgın	131	10	11	18
	Mutlu	20	184	5	16
	Üzgün	42	9	83	14
	Korku	8	4	7	89

Tablo II. 800 Tweet için hata matrisi

Tablo III sınıflandırma sonuçlarına göre Tutturma, Anma ve F-Ölçüm' değerlerini göstermektedir.

Sınıflar	Tutturma	Anma	F-Ölçüm
Kızgın	0.77	0.65	0.70
Mutlu	0.81	0.88	0.85
Üzgün	0.56	0.88	0.54
Korku	0.82	0.64	0.72

Tablo III. 800 Tweet için tutturma, anma, f- ölçüm değerleri

Kelimelerin duygu sözlükleri içerisinde geçme sıklıkları dikkate alınarak yapılan etiketleme de alınan ortalama başarı %73 olarak tespit edilmiştir. *Üzgün* ile *Kızgın* duygu sınıfları birbiri ile karışabilmektedir. *Üzgün* ve *kızgın* olduğumuz zamanlar kullandığımız kelimelerden daha çok ruh halimiz ve yüz ifademiz karşımızdaki kişiye bizim duygu durumumuz hakkında daha fazla bilgi vermektedir.

Başarıyı attırabilmek adına dört farklı duygu sözlüğünde yer alan kelimeleri tek bir sözlük altında topladık. Ancak, bu duygu ifadesini veren kelimeler birden fazla duygu sınıfında yer alabildiği için oluşturulan sözlük tekil kelimelerden oluşturuldu. Bu sözlük referans alınarak her bir tweet, 540 özellikten oluşacak şekilde bir ikili vektör ile temsil edildi. Tweet içerisinde ilgili duygu kelimesi yer alıyorsa "1", almiyorsa "0" ile gösterildi. Tweet içerisindeki her kelime yine *soldan-k* yöntemi ile alındı ve sözlük içerisinde en uzun kapsamanın olduğu duygu kelimesi ile eşleştirildi.

Sistemin başarımında bir artış olup olmadığını gözlemlemek adına Weka içerisinde yer alan Destek Vektör Makinesi (DVM) ve Rastgele Orman (R0) yöntemleri 10-katlı çapraz geçişleme ile çalıştırılmıştır.

Tablo IV'te DVM için alınan hata matrisini ve Tablo V'te de sınıf bazında başarı gösterilmektedir.

Gerçek	Tahmin				
		Kızgın	Mutlu	Üzgün	Korku
	Kızgın	106	12	57	2
	Mutlu	23	182	29	1
	Üzgün	25	15	119	1
	Korku	10	6	20	73

Tablo IV. 800 Tweet için DVM hata matrisi

	Tutturma	Anma	F-Ölçüm
Kızgın	0.59	0.64	0.62
Mutlu	0.77	0.84	0.81
Üzgün	0.74	0.52	0.62
Korku	0.67	0.94	0.78

Tablo V. 800 tweet için DVM tutturma, anma, F-ölçüm değerleri.

Destek Vektör Makinesinde de gerçekte *Kızgın* olan durum *Üzgün* olarak etiketlenmiştir. Hata oranı en düşük olan sınıf *Mutlu* olup, yöntemin başarı ortalaması %70'dir.

Aynı denemeleri Rastgele Orman ve Naive Bayes ile tekrarladığımızda Tablo VI ve VII'da Rastgele Orman için hata matrisi ve başarımların değerleri verilmiş iken, Tablo VIII ve IX'de de Naive Bayes için verilmiştir.

Gerçek	Tahmin				
		Kızgın	Mutlu	Üzgün	Korku
	Kızgın	149	11	11	6
	Mutlu	61	164	6	4
	Üzgün	69	12	77	2
	Korku	24	5	5	75

Tablo VI. 800 tweet için rastgele orman hata matrisi

	Tutturma	Anma	F-Ölçüm
Kızgın	0.84	0.49	0.62
Mutlu	0.70	0.85	0.76
Üzgün	0.49	0.78	0.60
Korku	0.69	0.86	0.76

Tablo VII. 800 tweet için rastgele orman tutturma, anma, f-ölçüm değerleri

Rastgele Ormanda sınıfların karıştırılma durumları daha fazladır. *Mutlu*, *Kızgın* ile karışırken *Üzgün* yine *Kızgın* ile karışmıştır. Bunun sebebi ağaçların derinliğinin fazla olmasıdır. Hiçbir budama işlemi yapılmadan ilk değer ile çalıştırılmıştır. Yöntemin ortalama başarısı %69'dur.

Gerçek	Tahmin				
		Kızgın	Mutlu	Üzgün	Korku
	Kızgın	13	8	42	4
	Mutlu	26	186	17	3
	Üzgün	37	11	110	2
	Korku	16	6	13	74

Tablo VIII. 800 tweet için Naive Bayes hata matrisi

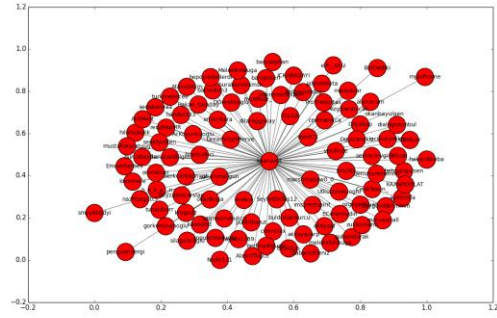
	Tutturma	Anma	F-Ölçüm
Kızgın	0.69	0.61	0.65
Mutlu	0.80	0.88	0.84
Üzgün	0.69	0.60	0.64
Korku	0.68	0.89	0.77

Tablo IX. 800 tweet için Naive Bayes tutturma, anma, f-ölçüm değerleri

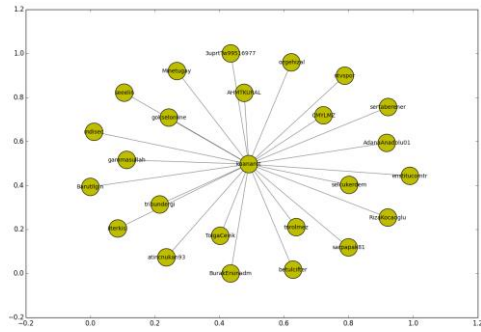
Naive Bayes yönteminde de *Kızgın* duygusu *Üzgün* duygusu ile karıştırılmıştır. Yöntemin ortalama başarısı %72'dir. Duygu sınıflandırmada Naive Bayes yöntemi en başarılı sınıflandırıcı olmuştur.

Geliştirilen bu yöntem ile bizler %72 doğruluk ile gelen Tweetleri 4 farklı duygu durumuna göre etiketleyebiliyorsak, bir kullanıcının baskın olan duygusal durumunu, takip ettiği kişilerin duygusal durumlarını tahmin edebiliriz. Hatta çalışmanın ileri aşamalarında takip ettiğimiz ve aynı duygusal durumlara sahip olan arkadaşlarımızı birbirlerine takip etmeleri için önerebiliriz. Elimizde benzer çalışmaların yapıldığı hiçbir çalışmaya ait etiketli veri olmadığından başarımlar ölçütleri ve veri seti karşılaştırılması yapılamamıştır.

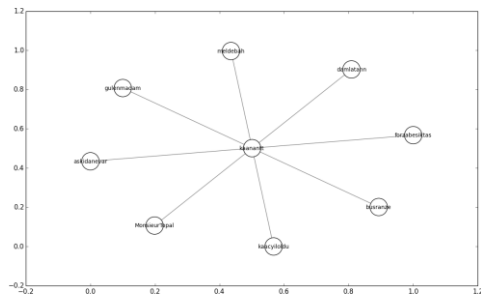
Şekil 1'de *kaanantt* isimli kullanıcının attığı tweetlerden genelde duygusal durumunun kızgın olduğu ve takip ettiği aynı zamanda kendisi gibi kızgın bir ruh haline sahip kullanıcılar gösterilmiştir. Uygulamada kızgın-kırmızı, mutlu-yeşil, korku-beyaz ve üzgün-mavi ile renklendirilmiştir.

**Şekil 1.** Atılan Tweetlerde genelde kızgınlık duygusu barındıran kullanıcılar

Şekil 2'de ise yine merkez noktasında seçilen kullanıcı ve takip ettiği kişilerden mutluluk duygusu daha baskın olan kullanıcılar gösterilmektedir.

**Şekil 2.** Atılan Tweetlerde genelde mutluluk duygusu barındıran kullanıcılar

Şekil 3'te aynı şekilde attığı Tweetlerde korku duygusu taşıyan kullanıcıları göstermektedir.

**Şekil 3.** Atılan Tweetlerde genelde korku duygusu barındıran kullanıcılar

IV. SONUÇ

Bu çalışmada Tweetler seçilen kızgın, üzgün, mutlu ve korku gibi dört farklı duygu durumuna göre etiketlendirilmesi için bir yöntem önerilmiştir. Makine öğrenmesi yöntemleri arasında; Naïve Bayes yöntemi ile %72 ile en başarılı sonuç elde edilmiştir. Amacımız, çalışmanın sonraki aşamalarında veri setindeki etiketli örnek sayısını arttırmak ve başarıyı yükseltmektir. Bu çalışmada kişinin ve takip ettiği kişilerin duygu durumu tahmin edilirken ileride takip ettiği kişiler ile ne kadar oranda duygusal olarak benzerlik gösteriyor, kendi mizacının dışındaki kişileri takip etme oranı nedir? sorularına cevap veriyor iken daha kapsamlı bir duygusal sosyal ağ analizi çıkarabilmektir.

REFERANSLAR

- [1] M. Szomszor, P. Kostkova, De Quincey, E. Swineflu, "Twitter predicts swine flu outbreak in 2009", *In Proceeding of 3rd International ICST Conference on Electronic Healthcare for the 21st century*, 2009.
- [2] J. Bian, U. Topaloglu, F. Yu, "Towards Large-scale Twitter Mining for Drug-related Adverse Events", *SHB'12*, Maui, Hawaii, USA, 2012.
- [3] L.E. Nguyen, P. Wu, W. Chan, W. Peng, "Predicting Collective Sentiment Dynamics from Time-series Social Media", *WISDOM '12*, Beijing, China, 2012.
- [4] W.B. Claster, H. Dinh, M. Cooper, "Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism - Social Media Data Analysis", *In 2010 Second World Congress on Nature and Biologically Inspired Computing*, Fukuoka, Japan, 2010.
- [5] U. Eroğul, "Sentiment analysis in Turkish", Middle East Technical University, Ms Thesis, Computer Engineering, 2009.
- [6] A.G. Vural, B.B. Cambazoğlu, P. Şenkul, ve Z.O. Tokgoz, "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in turkish", *Computer and Information Sciences III*, s 437-445, Springer, 2013.
- [7] P. P. Alexander Pak, "Twitter as a corpus for sentiment analysis and opinion mining", *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010.
- [8] M. Hasan, E. Rundensteiner, E. Agu, "Emotex: Detecting emotions in twitter messages", *Computer Science Department, Worcester Polytechnic Institute*, 2013.
- [9] E.S. Akgül, E. Ertano, B. Diri, "Twitter verileri ile duygu analizi", *Pamukkale Üniversitesi Mühendislik Bilim Dergisi*, 22(2), s. 106-110, 2016.
- [10] M.F. Amasyalı, A. Bolat, *2 milyon tweet veri kümesi*, YTÜ-Doğal Dil İşleme Araştırma Grubu-Kemik, Nisan 2013.
- [11] F. Can, S. Koçberber, "Türkçe Metinlerde Bilgi Erişimi", *Bilkent Bilgi Erişim Grubu*, 12 Haziran 2017.
- [12] Ş. Öztürkmenoğlu, *Matrislerler Şifreleme Üzerine*, Yüksek Lisans Tezi, Muğla Sıtkı Koçman Üniversitesi, Şubat 2016.
- [13] Saif Mohammad, Pierre Charron, "The Sentiment and Emotion Lexicons", *National Research Council Canada*