Türkçe Twitter Mesajlarının Duygu Analizi
Sentiment Analysis for Turkish Twitter Feeds

Önder Çoban, Barış Özyer, Gülşah Tümüklü Özyer
Bilgisayar Mühendisliği Bölümü
Atatürk Üniversitesi, Erzurum, Türkiye
Email: onder.coban, baris.ozyer, gulsah.ozyer@atauni.edu.tr

***Özetçe* —Duygu analizi sosyal medya izleme çalışmaları için en kullanışlı yöntemlerden birisidir. Sosyal medya (Kişisel Blog, Twitter, Facebook) üzerinden elde edilen veri üzerinde duygu analizi uygulanarak, bir şirketin müşteri servisinin, müşterilerden gelen olumlu ve olumsuz geri bildirimlere göre müşteri memnuniyetini sağlaması ve maliyetleri düşürmesi sağlanabilir. Ayrıca ekonomik, ticari ve kullanıcılara yönelik fikir madenciliği gibi çeşitli alanlarda kullanılarak anlamlı bilgiler elde edilebilir. Bu çalışmada, Türkçe Twitter mesajlardan oluşturulan veri seti metin sınıflandırma yöntemleri ile analiz edilerek olumlu veya olumsuz olup olmadığı incelenmiştir. Deneysel sonuçlar SVM, Naive Bayes, Multinom Naive Bayes ve KNN algoritmalarıyla elde edilmiştir. Vector Space model ile temsil edilen öznitelikler, kelime torbası (Bag of Words, BoW) ve N-Gram model olmak üzere iki farklı şekilde elde edilmiş ve bu durumun sınıflandırma sonuçlarına olan etkisi incelenmiştir.**

Anahtar Kelimeler—twitter, duygu analizi, duygu sınıflandırması, makine öğrenmesi, metin sınıflandırma.

***Abstract*—Sentiment analysis is one of the most useful tools in social media monitoring. Implementing sentiment analysis on data gained from social media (Blogs, Twitter, and Facebook) can increase the customer satisfaction and decrease the costs for a company. Also sentiment analysis can be used in various domains, such as economic, commercial and opinion mining for the users to get meaningful information. In this study, Turkish Twitter feeds collected from Twitter API have been analyzed in terms of the sentiment context whether positive or negative using document classification methods. Experimental results have been conducted on machine learning algorithms such as SVM, Naive Bayes, Multinomial Naive Bayes and KNN. The features represented by vector space are extracted from two different models which are Bag of Words and N-Gram. The experimental results have been investigated on the effect of classification methods.**

Keywords—twitter, sentiment analysis, sentiment classification, machine learning, text classification.

I. GİRİŞ

Günümüzde internet, insanların görüşlerini ifade edebildikleri küresel bir foruma dönüşmüştür. Web ortamında sosyal medya içeriğinin artmasıyla birlikte, kullanıcılar herhangi bir konu ile ilgili düşüncelerini kişisel blog, Facebook ve Twitter gibi sosyal ağlarda ifade edebilmektedirler [1]. Bu kişisel kayıtlar, sosyal psikologlar, pazarlama zekâsı ve fikir madenciliği araştırmaları için zengin ve kullanışlı bir kaynak oluşturmaktadır [2],[3]. Bir ürün hakkında olumlu ve olumsuz yorumların, blog kullanıcılarının mutlu, kızgın gibi ruh hallerinin, toplumun güncel bir politik konuya verdiği tepkinin tespit edilmesi gibi konular bu tür araştırmalara örnek olarak verilebilir. Bu sebeple son yıllarda, özellikle Twitter mesajları olmak üzere çeşitli sosyal medya ortamlarından elde edilen veriler üzerinde, duygu analiz teknikleri kullanılarak yapılan çalışmalar araştırmacıların ilgisini çekmektedir [4].

Twitter kullanıcıların, tweet olarak adlandırılan ve en fazla 140 karakterden oluşan, herhangi bir konu ile ilgili düşüncelerini paylaşabildikleri popüler sosyal ağlardan birisidir. Paylaşılan mesajlar, kullanıcıların faklı konular ile ilgili fikir ve duygularını içermektedir. Ayrıca Twitter farklı kültür ve seviyeden milyonlarca kullanıcı kitlesine sahip olduğu için farklı dillerde ve içeriklerde mesajlar toplamak mümkündür. Bu nedenle, bu çalışmada bir sosyal medya ortamı olan Twitter üzerinden elde edilen mesajlar duygu analizi sınıflandırılması probleminde kullanılmıştır [4], [8].

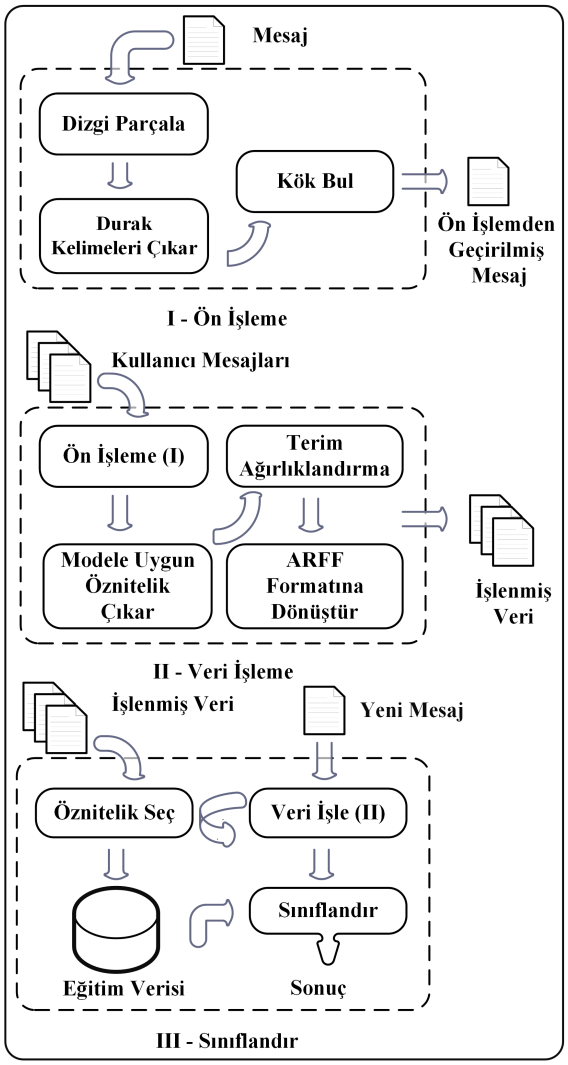
Literatürde duygu analizi alanında İngilizce için birçok istatistiksel ve dilbilimsel çalışma yapılmış olmakla beraber [6-9] Türkçe için henüz çok fazla çalışma yayınlanmamıştır [5], [14]. Bu bildiride öncelikli olarak Twitter üzerinden çekilen Türkçe bir veri seti oluşturulmuştur. Bu veri seti üzerinde kullanıcı mesajlarının olumlu veya olumsuz olup olmadığının tespit edilmesi amacıyla öznitelik çıkarım modeli geliştirilmiş ve bu modelinin sınıflandırma başarısına etkisi incelenmiştir.

Bildirinin geri kalan bölümleri şu şekilde düzenlenmiştir. Önerilen sistem modeli Bölüm II'de anlatılmıştır. Bölüm III'de veri setleri üzerinde duygu analizi için gerekli önişleme, öznitelik seçme ve sınıflandırma yöntemleri açıklanmıştır. Oluşturulan veri setine ait bilgiler Bölüm IV'de verilmiştir. Bölüm V'de gerçekleştirilen deneysel sonuçlar verilmiş ve karşılaştırmalar yapılmıştır. Son bölümde sonuçlar özetlenmiş ve gelecekte yapılacak çalışmalar belirtilmiştir.

II. SİSTEM MODELİ

Türkçe twitter mesajlarının duygu analizi için önerilen sistem modeli Şekil I'de gösterilmiştir. Sistem üç aşamadan oluşmaktadır. Sistemde kullanılacak olan eğitim ve test verisi ilk olarak ön işlemden geçirilmiştir. Ön işlemden geçirilen mesajlardan kelime torbası (Bag of Words, BoW) ve N-Gram model olmak üzere iki farklı yöntemle öznitelik çıkarılmaktadır. Daha sonra çıklarılan özniteliler ağırlıklandırılmıştır ve elde edilen veri arff (Attribute Relational File Format) formatına dönüştürülmüştür. Sistemin son aşamasında, veriyi temsil eden en iyi öznitelikler seçilerek sınıflandırma işlemi gerçekleştirilmiştir. Sistem modelinde kullanılan yöntemlere ait ayrıntılı bilgiler Bölüm III'de anlatılmıştır.

978-1-4673-7386-9/15/$31.00



Şekil I. Sistem Modeli

III. YÖNTEMLER

Veri setindeki her bir mesaj içeriği iki (pozitif, negatif) veya daha fazla kategoride (çok iyi, iyi, tatmin edici, kötü, çok kötü vb.) olmak üzere sınıflandırılabileceği için duygu analizi, her bir mesajın bir kategoriyi temsil ettiği bir doküman sınıflandırma işlemi olarak düşünülebilir [3], [10]. Bu nedenle çalışmamızda, Türkçe Twitter mesajları üzerinde temel doküman sınıflandırma yöntemleri uygulanmıştır.

A. Ön İşleme

Twitter mesajları belirgin bazı özelliklere sahiptir. Paylaşılan mesajlar genellikle "@" karakteriyle başlayan kullanıcı adı, "#" karakteriyle başlayan hashtag, his simgeleri ve URL içermektedir. Bu belirgin özelliklerin tespit edilmesi ve sınıflandırıcının eğitilebilmesi için verinin çeşitli ön işleme yöntemlerinden geçirilmesi gerekmektedir. Bu nedenle çalışmızda, ön işleme aşamasında öncelikle Retweet, Retweeted ve tekrar eden mesajlar elendikten sonra veri seti üzerinde aşağıda açıklanan işlemler uygulanmıştır.

Dizgi Parçalama : Veriyi parçalara ayırıp içerisinden anlamlı bilgi elde etmeyi amaçlayan işlemdir. Bu bağlamda dizgi parçalama aşamasında aşağıdaki işlemler uygulanmıştır.

- Pozitif ve negatif grupta bulunan his simgelerinden her ikisini de içeren mesajlar elenmiştir.
- Mesaj içeriği küçük harfe dönüştürülmüş ve harf olmayan her türlü karakter (noktalama işaretleri, rakamlar ve anlamsız karakterler) temizlenmiştir.
- Mesaj içeriğinde bulunan his simgeleri, URL, hashtag, kullanıcı adları ve uzunluğu üç karakterden az olan terimler çıkarılmış ve terim sayısı ikiden az olan mesajlar elenmiştir.

Son olarak BoW ve N-Gram model kullanılarak iki farklı şekilde öznitelik çıkarılmıştır [15], [16].

Durak Kelimeleri Çıkarma: Durak kelimeler bir dilde yaygın olarak kullanılan ve genellikle tek başına anlam ifade etmeyen kelime listesidir. Her dilin özel durak kelimeleri mevcut olmakla beraber, standart bir liste mevcut değildir. Bu çalışmada Lucene [1] uygulama geliştirme arayüzünde (API) Türkçe için mevcut durak kelime listesi kullanılmıştır.

Kök Bulma: Öznitelik uzay boyutunu düşürmeye yarayan ve kelimeleri köküne indirgeme amaçlı uygulanan bir kelime azaltma işlemidir. Bu çalışmada kelimelerin köke indirgenmesi için bir Türkçe doğal dil işleme kütüphanesi olan Zemberek[2] kullanılmıştır.

Terim Ağırlıklandırma: Terim ağırlıklandırma bir terimin doküman içerisinde ne kadar önemli olduğunu belirtmek amacıyla oluşturulan terim vektöründe her bir terim için ağırlık belirleme işlemidir [12]. Bu çalışmada özniteliklerin ağırlıklandırılması için aşağıda formülleri verilen TF (Terim Frekansı), Boolean, ve TF-IDF (Terim Frekansı-Ters Doküman Frekansı) üç farklı ağırlıklandırma yöntemi uygulanmıştır.

$$W_{TF}(c,d) = TF(c,d) \tag{1}$$

$$W_{boolean}(c,d) = \begin{cases} 1 & TF \geq 1 \\ 0 & \text{diğer durumlar} \end{cases} \tag{2}$$

$$DF(c,d) = |\{d \in D : c \in d\}| \tag{3}$$

$$W_{IDF}(c,d) = \log \frac{N}{DF(c,d)} \tag{4}$$

$$W_{TF-IDF}(c,d) = TF(c,d) * IDF(c,d) \tag{5}$$

W ağırlıklandırma fonksiyonunu, d bir dökümanı, c bir terimi, TF terim frekansını, DF doküman frekansını, D verisetindeki tüm dokümanları ve N toplam doküman sayısını temsil eder.

Tekrarlanan Harflerin Çıkartılması : Metin normalizasyonu, bir dizgi veya metin üzerinde çeşitli dönüşümler yaparak amaca uygunluğunu artırma işlemidir. Bu çalışmada normalizasyon kelime içerisinde tekrar eden harflerin bire indirgenmesi amacıyla kullanılmış böylece öznitelik uzayının boyutu azaltılmıştır. Örneğin kelime "günaaydıııınnn" ise bu kelime "günaydın" olacak şekilde dönüştürülmüştür.

[1] http://lucene.apache.org/
[2] https://code.google.com/p/zemberek/

B. Öznitelik Seçme

Öznitelik seçme, veri setinde ayırt edici özelliği yüksek olan özniteliklerin, başka bir ifadeyle veri setini en iyi temsil eden özniteliklerin seçilmesidir. Bu şekilde mevcut öznitelik uzayının boyutu düşürülmekte, zaman ve performans açısından avantaj sağlanmaktadır. Bu çalışmada, öznitelik seçme işlemi için kolerasyon tabanlı bir yöntem olan CfsSubset [13] algoritması kullanılmıştır.

C. Sınıflandırma

Sınıflandırma işlemi daha önce görülmemiş ve kategorisi bilinmeyen her bir örneğin, eğitim verisi kategorilerinden en uygun olan kategoriye atanması işlemidir. Bu çalışmada doküman sınıflandırma çalışmalarında yaygın olarak kullanılan k-en yakın komşu (k-NN), Destek Vektör makineleri (SVM), Naive Bayes (NB) ve Multinom Naive Bayes (MNB) kullanılmıştır [11].

IV. Veri Seti

Duygu analizi çalışmalarında kullanılabilecek Türkçe mesajlardan oluşmuş ve herkese açık bir Twitter veri seti mevcut değildir. Bu nedenle [6]'da uygulanan yöntemle Twitter API [3] kullanılarak Türkçe mesajlar içeren bir veri seti oluşturulmuştur. Twitter API, kendisine gönderilen her bir sorgu için maksimum 100 mesaj alınmasına izin verdiği için geliştirilen Java tabanlı uygulama ile daha fazla sayıda mesaj alınması sağlanmıştır. Mesajlar çekilirken hem gönderilen sorgularda hem de çekilen mesajın kategorisinin belirlenmesinde sadece his simgeleri kullanılmıştır. Mesajın içerdiği his simgesinin bulunduğu gruba göre, mesaj pozitif veya negatif olarak etiketlenmiştir [4], [8]. Anahtar kelime olarak sorgulama aşamasında ve mesajın etiketlenmesinde kullanılan iki gruba ayrılmış his simgeleri aşağıda verilmiştir.

- Pozitif grup: ":-)", ":)", "=)", ":D"
- Negatif grup: ":-(", ":(", "=(", ";("

Bir mesajın etiketlenmesi için pozitif veya negatif grupta bulunan his simgelerinden birisini içermesi yeterlidir. Yukarıda açıklanan işlemler sonucu toplamda, 10000 pozitif ve 10000 negatif olmak üzere 20000 mesaj içeren bir Türkçe veri seti oluşturulmuştur.

V. Deneysel Sonuçlar

Oluşturulan Türkçe veri seti yöntemlerde açıklanan ön işlemden geçirildikten sonra 14777 adet mesajdan oluşan bir veri seti elde edilmiştir. Tablo I'de ön işlem önce ve sonrasına ait ortalama terim ve örnek sayıları verilmiştir. Tablo II'de ise çıkarılan özniteliklere ait istatistiksel bilgiler gösterilmiştir. Öznitelik çıkarma aşamasında BoW model için uygulanan kök bulma, durak kelimelerin ve tekrarlanan harflerin çıkarılması işlemleri N-Gram model için uygulanmamıştır. Bu nedenle ön işleme önce ve sonrasına ilişkin bilgiler ve öznitelik istatistikleri sadece BoW model için verilmiştir.

Sınıflandırmaya ait deneysel sonuçlar geliştirdiğimiz Weka [4] tabanlı Java uygulaması ile elde edilmiş, sınıflandırıcı olarak daha önce bahsedilen algoritmalar kullanılmıştır. Sınıflandırma yapılırken 10-kat çapraz geçerleme modeli kullanılmıştır. BoW ve N-Gram modellerinin her ikisi için de his simgeleri sadece mesajları etiketlemek için kullanılmış ön işleme aşamasında his simgeleri, hashtag, URL ve kullanıcı adları elenerek sınıflandırıcı sadece mesaj içeriğinden çıkarılan anlamlı kelimeler veya n-gram'lar ile eğitilmiştir. Bu çalışmada literatürde yapılan çalışmalardan farklı olarak N-Gram model için öznitelikler kelime seviyesinde değil karakter seviyesinde çıkarılmıştır.

Tablo I. Ortalama terim ve örnek sayıları

Özellik	Ön İşlem	
	Önce	Sonra
Ortalama Terim	7,5	5,1
Pozitif Etiketli Örnek	6887	6269
Negatif Etiketli Örnek	7890	7043
Toplam Örnek	14777	13312

Tablo II. Ön işlemden geçirilen öznitelik istatistikleri

Elenen öznitelik	Öznitelik Sayısı	Yüzdesi
Yok	111316	100%
Hashtag	889	0,79%
URL	1913	1,71%
Kullanıcı Adı	7026	6,31%
His Simgeleri	14778	13,27%
Çıkarılan Terim	18416	16,54%
Hepsi	68294	61,35%

Tablo III. Pozitif ve negatif kategorilerde ortak en çok geçen 30 kelime.

Ortak Kelimeler				
tatil	iyi	ol	gel	günaydın
yok	gün	et	gece	ilk
güzel	okul	yap	git	sev
bak	başla	al	bil	yaz
sabah	uyku	gör	yeni	sen
iste	iş	zaman	ulan	kal

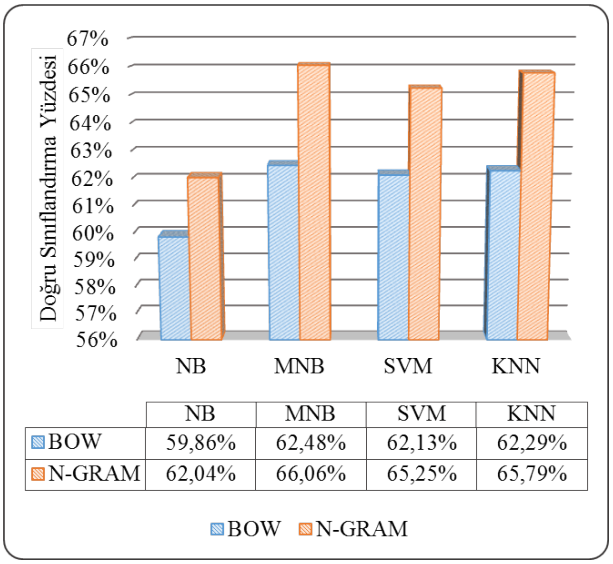
Tablo IV. Pozitif ve negatif kategorilerde en çok geçen 20 kelime.

En Çok Geçen Kelimeler					
Pozitif	mutlu	bul	süper	hadi	um
	hey	mükemmel	herkes	gül	kardeş
	komik	sevgi	film	yıl	teşekkür
	takip	doğ	hile	konu	az
Negatif	lütfen	of	kalk	keşke	çocuk
	öl	sık	kötü	falan	gir
	çek	önce	dön	hasta	lise
	san	ağla	unut	lütfen	acı

Deneysel sonuçlara göre, oluşturduğumuz veri üzerinde karakter seviye N-Gram model BoW modelden tüm sınıflandırıcılar için daha iyi sonuç vermiştir. Özniteliklerin ağırlıklandırılması aşamasında Boolean, Terim Frekansı (TF) ve Terim Frekansı-Ters Doküman Frekansı (TF-IDF) yöntemleri kullanılmış en iyi sonucu Boolean yöntemi vermiştir. Öznitelik seçme aşamasında, ön işlemden sonra N-Gram model için elde edilen toplam 17720 benzersiz öznitelikten 144 en iyi

[3] https://apiwiki.twitter.com/

[4] http://www.cs.waikato.ac.nz/ml/weka/



	NB	MNB	SVM	KNN
BOW	59,86%	62,48%	62,13%	62,29%
N-GRAM	62,04%	66,06%	65,25%	65,79%

Şekil II. NB, MNB, SVM ve k-NN algoritmalarının farklı öznitelik çıkarım modelleri için sınıflandırma başarıları

öznitelik seçilmiştir. BoW model için ise Tablo II'de belirtilen toplam öznitelik sayısından elde edilen 7569 benzersiz öznitelikten 122 en iyi öznitelik seçilmiştir. Her iki modelde çıkarılan öznitelikler kelime ve karakter tabanlı olmak üzere iki farklı yapıda olduğundan, toplam öznitelik sayısı ve bu özniteliklerin ağırlıkları değişiklik göstermiştir. Bu nedenle seçilen en iyi öznitelik sayılarında farklılık oluşmuştur.

Sınıflandırma aşamasında öznitelikler N-Gram model için 2, 3 ve 4-gram olmak üzere üç farklı şekilde elde edilmiş ve en iyi sonucu 3-gram vermiştir. Ayrıca kullanılan makine öğrenmesi yöntemlerinden her iki model için de en iyi sonucu Multinom Naive Bayes vermiştir. SVM sınıflandırmasında doğrusal çekirdek kullanılmış, k-NN sınıflandırmasında en yakın komşu sayısı k=1 alınmıştır. Sonuçlara ait özet bilgiler Şekil II'de verilmiştir. Ayrıca BoW model için pozitif ve negatif kategorili mesajlarda ayrı ayrı ve ortak geçen frekansı en yüksek köküne indirgenmiş kelimeler tespit edilmiştir. Ortak kelimelerden 30 tanesi Tablo III'de ve her iki kategoride ayrı ayrı geçen kelimelerden 20 tanesi de Tablo IV'de verilmiştir.

VI. SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada metin sınıflandırma alanında kullanılan temel yöntemler kullanılarak Twitter ortamından elde edilen mesajlar üzerinde duygu analizi yapılmış ve literatürde temel yöntemler kullanılarak yapılan çalışmalarda elde edilen başarı oranı Türkçe için yakalanmıştır. Sonuçlar duygu analizi çalışmalarının bir metin sınıflandırma problemi olarak ele alınabileceğini göstermekle beraber başarı oranının yükseltilebilmesi için harici yöntemler uygulanması gerektiği açıktır. Bu araştırmadan elde edilen bulgulara göre, deneysel sonuçlar Twitter mesajlarının makine öğrenmesi yöntemleriyle sınıflandırılabileceği tezini doğrulamıştır.

Veri setinin rastgele ve sadece his simgeleri kullanılarak oluşturulması ve eğitim aşamasında bu simgelerin elenmesi sınıflandırma başarısını düşürmüştür. Ayrıca mesajların belirli bir konu gözetilmeksizin rastgele çekilmesi, sınıflandırıcının genelleme yapma yeteneğini düşürmüştür. Başka bir deyişle Tablo III'de verilen kelimelerden de anlaşılacağı gibi pozitif kategoride ayırt edici öznitelik olması beklenen "iyi, güzel, günaydın, yeni" kelimelerinin negatif kategorili mesajlarda da çok yüksek frekansta geçtiği tespit edilmiştir. Bu nedenle sınıflandırma başarısı BoW model için daha düşük elde edilmiştir. Sınıflandırma başarısına kullanılan eğitim verisinin ve veri setinden çıkarılan özniteliklerin doğrudan etkisi olduğu bilinmektedir. Bu nedenle gelecekte yapılacak çalışmalarda, ayırt edilebilirliği çok daha iyi olan örneklerden oluşan bir veri setinin oluşturulması ve sınıf sayısının artırılması (2'den fazla) düşünülmektedir. Ayrıca Twitter mesajları 140 karakterle sınırlı olması verinin boyutunu düşürdüğünden, daha fazla mesaj çekilerek veri boyutunun artırılması ve böylece sınıflandırıcıların genelleme yapma yeteneğinin artırılması ve öznitelik çıkarım aşamasında ayırt edici özelliği daha iyi olan özniteliklerin çıkarılabilmesi için ek olarak semantik (kelime tabanlı model için kelime önerme, eş ve zıt anlamlı kelimelerin ağırlıklarını değiştirme, sıklıkla kullanılan çeşitli kısaltmalar için elenmesini önleyecek yöntem geliştirme vs.) ve matematiksel yöntemlerin uygulanması düşünülmektedir.

KAYNAKÇA

[1] Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis." Entropy 17 (2009).

[2] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.

[3] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." Journal of Informetrics 3.2 (2009): 143-157.

[4] Tang, Huifeng, Songbo Tan, and Xueqi Cheng. "A survey on sentiment detection of reviews." Expert Systems with Applications 36.7 (2009): 10760-10773.

[5] Kaya, Mesut, Guven Fidan, and Ismail H. Toroslu. "Sentiment analysis of turkish political news." Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2012.

[6] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford (2009): 1-12.

[7] Mishne, Gilad, and Natalie S. Glance. "Predicting Movie Sales from Blogger Sentiment." AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[8] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[9] Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[10] Sebastiani, Fabrizio. "A tutorial on automated text categorisation." Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. Buenos Aires, AR, 1999.

[11] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[12] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing and management 24.5 (1988): 513-523.

[13] Hall, Mark A. Correlation-based feature selection for machine learning. Diss. The University of Waikato, 1999.

[14] Dilek, K., and Ralf Steinberger. "Experiments to Improve Named Entity Recognition on Turkish Tweets." Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL. 2014.

[15] Scott, Sam, and Stan Matwin. "Feature engineering for text classification." ICML. Vol. 99. 1999.

[16] Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Ann Arbor MI 48113.2 (1994): 161-175.