Pamukkale Univ Muh Bilim Derg, 24(2), 283-291, 2018

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Twitter duygu analizinde terim ağırlıklandırma yönteminin etkisi

The impact of term weighting method on Twitter sentiment analysis

Önder ÇOBAN[1], Gülşah TÜMÜKLÜ-ÖZYER[2]*

[1]Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, Atatürk Üniversitesi, Erzurum, Türkiye.
onder.cbn@gmail.com
[2]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Atatürk Üniversitesi, Erzurum, Türkiye.
gulsah.ozyer@atauni.edu.tr

Geliş Tarihi/Received: 22.10.2016, Kabul Tarihi/Accepted: 16.01.2017
doi: 10.5505/pajes.2017.50480
* Yazışılan yazar/Corresponding author
Araştırma Makalesi/Research Article

Öz

Terim ağırlıklandırma, metin sınıflandırmada sonuçlar üzerinde doğrudan etkili olan önemli bir adımdır. Ancak, bir metin sınıflandırma problemi olarak ele alınan duygu analizinde farklı önişleme tekniklerine bağlı olarak ağırlıklandırma yönteminin davranışı değişebilmektedir. Bu çalışmada bilgi getirimi, metin sınıflandırma, doküman filtreleme gibi farklı çalışma alanları için yakın zamanda önerilen yöntemler Twitter duygu analizinde uygulanmış ve sonuçlar üzerindeki etkisi incelenmiştir. Öznitelikler çıkarılırken kelime torbası (BoW) ve karakter seviye N-gram olmak üzere iki farklı model kullanılmıştır. Deneyler Türkçe ve İngilizce Twitter mesajlarından oluşan veri kümeleri üzerinde uygulanmıştır. Twitter mesajlarının duygu sınıflandırması, Gizli Dirichlet Ataması (LDA) tabanlı konu modeli ile gerçekleştirilmiştir. Sınıflandırma aşamasında ise Destek Vektör Makinesi (SVM) algoritması kullanılmıştır. Deneysel sonuçlara göre, Twitter duygu analizi çalışmalarında kullanılabilecek en etkili terim ağırlıklandırma yöntemi önerilmiştir.

Anahtar kelimeler: Twitter, Duygu analizi, Terim ağırlıklandırma

Abstract

Term weighting is an important step which has direct impact on the result in classical text classification. However, the behavior of the term weighting method may vary depending on different preprocessing techniques in sentiment analysis which considered as a text classification task. In this study, term weighted methods which are newly proposed for various research areas such as information retrieval, text classification and document filtering, performed to investigate effect on results for Twitter sentiment analysis. In feature extraction phase, two different models are used including Bag of Words (BoW) and character level N-gram. The experiments conducted on data sets consist of Turkish and English Twitter feeds. Sentiment classification of Twitter feeds performed using topic model generated with Latent Dirichlet Allocation (LDA) method. The Support Vector Machine (SVM) algorithm is employed in the classification stage. According to the experimental results, the most effective term weighting method that can be used in Twitter sentiment analysis studies is suggested.

Keywords: Twitter, Sentiment analysis, Term weighting

1 Giriş

Metin madenciliğinin önemli bir parçası olan metin sınıflandırma çalışmaları ile metin temsil kalitesi ve sınıflandırma başarısının artırılması amaçlanır. Birçok uygulama alanı bulunan temel metin sınıflandırma süreci ise veri toplama, önişleme, indeksleme, ağırlıklandırma ve sınıflandırma işlemlerinden oluşur. Bu süreçte bulunan indeksleme adımında metin içeriğinden çıkarılan çeşitli öznitelikler kullanılarak her bir doküman örneği temsil edilir. Terim olarak da adlandırılan bu öznitelikler: kelimeler, sözcük grupları veya metnin ayırt edilebilirliğini artırabilecek herhangi bir bilgi olabilir. Dokümanların genellikle bir öznitelik vektörü kullanılarak temsil edildiği metin sınıflandırmada her bir terim indekslenirken bir ağırlık değeri ile ilişkilendirilir. Bu ağırlık değeri, ilişkilendirildiği terimin önemini ve gözlendiği dokümanın sınıflandırılmasına yaptığı katkıyı ölçer [1]. Duygu analizinde ise kullanıcıların herhangi bir konu ile ilgili fikir, deneyim ve düşüncelerini ifade edebildikleri ortamlardan (Google+, Facebook, Twitter vb.) elde edilen içerikler analiz edilir. Elde edilen kısa metinlerin otomatik olarak analiz edilmesi ve içerdiği duygunun (pozitif, negatif, nötr vb. gibi) tespit edilmesi amaçlanır. Bu bağlamda, duygu analizi çalışmalarında kullanılan veriler de metin tabanlı olduğu için bir metin sınıflandırma problemi olarak ele alınır [2]. Ancak özellikle sosyal medya mesajları kendine özgü özelliklere sahip olduğundan duygu analizinde önişleme ve öznitelik çıkarma aşamasında farklı teknikler kullanılmaktadır. Bunun yanı sıra duygu analizi çalışmalarında, klasik metin sınıflandırma çalışmalarının aksine anlamlı bilgi elde etmek daha zordur. Bu durumun nedeni genellikle sosyal medya içeriklerinin kısa olması, resmi olmayan *informal* dil ile yazılması ve gramer kurallarına uyulmadığı için yazım yanlışları içermesidir.

Terim ağırlıklandırma, hem metin sınıflandırma hem de duygu analizi çalışmalarında başarı oranını etkileyen önemli bir aşamadır. Veri kümesinin boyutu ve kullanılan özniteliklere göre ağırlıklandırma yöntemleri farklı davranışlar göstermektedir. Çalışmamızda, özellikle bilgi getirimi alanında yaygın olarak kullanılan geleneksel terim ağırlıklandırma yöntemlerinin yanı sıra metin sınıflandırma, konu çıkarma vb. gibi çalışma alanları için önerilmiş yeni yaklaşımlar incelenmiştir. Bu yöntemler Twitter duygu analizinde uygulanmış ve sonuçlar üzerindeki etkisi incelenmiştir. Literatürde kısa metinlerin duygu analizinde ağırlıklandırma yönteminin etkisinin incelendiği az sayıda çalışma mevcuttur [3],[4]. Ancak bu çalışmalarda *genellikle* geleneksel yöntemler ve varyasyonları incelenmiş, incelenen yöntem bakımından çalışma dar kapsamlı tutulmuş ve Twitter mesajları kullanılmamıştır. Bu çalışmada, geleneksel yöntemlerin yanı sıra yakın zamanda diğer metin tabanlı çalışma alanları için önerilen yeni yaklaşımlar Twitter duygu analizinde uygulanmıştır. Bu yöntemlerin duygu analizinde etkisi incelenmiş ve geniş kapsamlı bir karşılaştırma yapılarak en başarılı yöntem önerilmiştir. Literatürde geniş kapsamlı ve

özellikle yeni yaklaşımların incelendiği benzer bir çalışma tespit edilemediğinden, bu inceleme Twitter mesajlarının duygu analizi bağlamında oldukça önemli ve değerlidir.

Çalışmanın geri kalan bölümleri şu şekilde düzenlenmiştir. İlgili çalışmalar ile materyal ve yöntem sırasıyla Bölüm 2 ve Bölüm 3'te anlatılmıştır. Bölüm 4'te deneysel sonuçlara yer verilmiş ve kıyaslamalar yapılmıştır. Ayrıca, sonuçlar özetlenmiş ve öneriler belirtilmiştir.

2 İlgili çalışmalar

Temeli özellikle bilgi getirimi alanında 1970'li yıllara dayanan terim ağırlıklandırma, duygu analizi ve diğer metin madenciliği çalışmalarında (bilgi getirimi, metin sınıflandırma, metin kümeleme vb.) performansı artırmak için terimlere ağırlık atanmasıdır. Bu ağırlıklandırma işlemi ile bir terimin bir dokümandaki önemi belirtilmiş olur [5]. Bu nedenle, doküman temsilinin doğru ve etkili olmasında terim ağırlıklandırma önemli bir role sahiptir. Salton ağırlıklandırma yönteminin: terim frekansı, ters doküman frekansı ve normalizasyon olmak üzere üç faktör ile ele alınmasını önermiştir [6]. Bu değerlendirmeden hareketle geleneksel olarak nitelendirilen birçok ağırlıklandırma yöntemi önerilmiştir. Metin madenciliğinde kullanılan bu geleneksel yöntemler (TF, BINARY, TFIDF, ANTF, LTF vb.) ise çoğunlukla bilgi getirimi alanından uyarlanmıştır. Bu yöntemlerden en basiti sadece terimin bir dokümanda gözlenip gözlenmediğiyle ilgilenen ikili (BINARY) yöntemdir. TFIDF ise en çok kullanılan geleneksel yöntemdir. Ayrıca Okapi BM25 [7] ile birlikte TFIDF varyasyonu olan çeşitli yöntemler de bulunmaktadır [8],[9]. Geleneksel yöntemleri yerel ve global kapsamlı olarak kategorize etmek de mümkündür. Yerel olarak nitelendirilen BINARY, TF, ANTF ve LTF terimin ilgili dokümandaki frekansına bağlıdır [10]. Global olarak nitelendirilen IDF, GFIDF ve Entropy gibi yöntemler ise ilgili terimin *doküman frekansı* kullanılarak hesaplanan TF varyasyonlarıdır.

Geleneksel yöntemlerde, genellikle terimler ile sınıflar arasındaki dağılım dikkate alınmaz. Bu nedenle indeksleme, getirme, derecelendirme, sınıflandırma vb. gibi çalışmalarda daha etkin ve doğru ağırlıklandırma için yeni yaklaşımlar önerilmiştir [11]-[16]. Bu yeni yaklaşımlar incelendiğinde genellikle olasılık tabanlı olduğu ve ağırlıklı olarak sınıf bilgisinin kullanıldığı görülmektedir. Ayrıca klasik TFIDF yöntemindeki IDF parametresinin öznitelik seçme, sınıflandırma ve benzerlik ölçüm yöntemleri ile değiştirildiği yaklaşımlar da mevcuttur [17]-[19]. Bu yönüyle, yeni yaklaşımları genellikle sınıf bilgisi de kullanıldığından *denetimli*, geleneksel yöntemleri ise *denetimsiz* olarak nitelendirmek mümkündür [8],[20].

3 Materyal ve yöntem

3.1 Veri kümesi

Çalışmamızda Türkçe ve İngilizce Twitter mesajlarından oluşan iki farklı veri kümesi kullanılmıştır. Böylece metnin diline bağlı olarak değişebilen öznitelik uzay boyutu ve özniteliklerin de ağırlıklandırma yöntemleri üzerindeki etkisinin incelenmesi amaçlanmıştır. Duygu analizi çalışmalarında kullanılabilecek herkese açık standart bir Twitter veri kümesi mevcut değildir. Bu nedenle, araştırmacılar Twitter tarafından sunulan *Twitter API* aracılığıyla oluşturdukları veri kümeleri üzerinde çalışmalarını gerçekleştirmektedir. Bu çalışmada, İngilizce için Sentiment140 projesi kapsamında oluşturulan veri kümesinin bir alt seti kullanılmıştır [21]. Sentiment140 Twitter veri kümesi (S140) pozitif ve negatif kategoride 800000 adet olmak üzere toplam 1600000 mesaj içeren eğitim seti ve 177 negatif, 182 pozitif olmak üzere toplam 359 mesaj içeren test setinden oluşmaktadır. S140 büyük boyutlu bir veri kümesidir ve veri kümesinin boyutu bu çalışmanın kapsamı dışındadır. Bu nedenle, çalışmamızda S140 veri kümesine ait eğitim setinin bir alt seti (AS140) kullanılmıştır. Türkçe için ise S140 veri için kullanılan yöntemle (his simgelerinin etiketleme amacıyla kullanıldığı bir yöntem, distant supervision) Twitter API kullanılarak oluşturulan Türkçe Twitter Mesajları (TTM) veri kümesi kullanılmıştır. Her iki veri kümesi de mesajlar toplanırken sorgu anahtarı olarak his simgelerinin kullanılmasıyla oluşturulmuştur. Veri kümelerindeki mesajlar yine mesajlarda gözlenen his simgelerine bakılarak önceden pozitif veya negatif olmak üzere etiketlenmiş durumdadır. Yani his simgeleri hem sorgu anahtarı, hem de mesajları etiketlemek için gerekli bir nitelik olarak kullanılmıştır. *TTM* ve *AS140* veri kümeleri pozitif ve negatif kategorilerde eşit sayıda olmak üzere toplam 20000 Twitter mesajı içermektedir.

3.2 Önişleme

Önişleme adımının temel amacı, metinsel verilerden örnek kategorilerinin birbirinden ayırt edilmesini sağlayabilecek önemli özniteliklerin ortaya çıkarılması ve uygun formata dönüştürülmesidir [22]. Buradaki formattan kasıt verinin makine öğrenmesi algoritmalarının işleyip yorumlayabileceği şekilde sayısallaştırılmasıdır [23]. Sosyal medya ortamlarında paylaşılan veriler de yapılandırılmadığı ve doğal dil ile yazıldığı için çeşitli önişleme adımlarının uygulanması gerekmektedir. Bu nedenle, AS140 ve TTM verileri üzerinde önceki çalışmamızda uygulanan önişleme adımları uygulanmıştır [24]. Önişlemden sonra, daha önce his simgeleri kullanılarak etiketlenmiş mesajlar duygu sınıflandırmasından geçirilerek yeniden etiketlenmiştir. Böylece his simgelerinin etiketleme üzerinde bir etkisi kalmamış sadece sorgu anahtarı olarak kullanılması sağlanmıştır. Ancak, duygu sınıflandırmasından sonra his simgeleri ve Twitter'a özgü terimler (kullanıcı adları, URL ve hashtag) mesaj içeriklerinden çıkarılmış ve tamamen anlamlı içerikten öznitelik elde edilmesi sağlanmıştır.

3.3 Duygu sınıflandırması

Eğitilmiş modelin başarısı, modele verilen eğitim verisi örneklerinin doğru etiketlenme başarısıyla doğru orantılıdır. Bu nedenle, örnek kategorilerinin önceden bilinmediği durumlarda yüksek sınıflandırma başarısı elde edebilmek için öncelikle veri kümesindeki örneklere doğru kategorilerin atanması (annotation) gerekmektedir. Bu amaçla, literatürde his simgeleri, sözcük türü etiketleri, hashtag vb. gibi mesaj içeriğinde gözlenen nitelikler kullanılmaktadır [21],[25],[26]. Ayrıca, ön tanımlı sözlükler ve semantik işlemlere ihtiyaç duyulan doğrusal sınıflandırıcılar etiketleme amacıyla kullanılmaktadır [27],[28]. Bu yöntemlerle duygu analizinde elde edilen başarı oranının ise %59-87 arasında olduğu görülmektedir [30]. Bu yönüyle, Twitter duygu analizinde en büyük zorluklardan birisi mesajların doğru bir şekilde etiketlenmesi yani kategori atanmasıdır. Bu çalışmada kullanılan AS140 ve TTM veri kümeleri ise his simgeleri kullanılarak önceden etiketlenmiş durumdadır. Ancak bu yöntem ile çok başarılı sonuçlar elde edilememektedir. Bu nedenle, çalışmamızda Gizli Dirichlet Ataması (*GDA*) tabanlı konu modelleme yaklaşımı duygu sınıflandırması için kullanılmıştır [29],[30]. GDA tabanlı yarı otomatik bu yöntemde his simgeleri ve Twitter'a özgü terimler özel terimlerle kodlanmış ve iki konudan oluşan bir konu modeli oluşturulmuştur. Daha sonra elde edilen konularda en sık

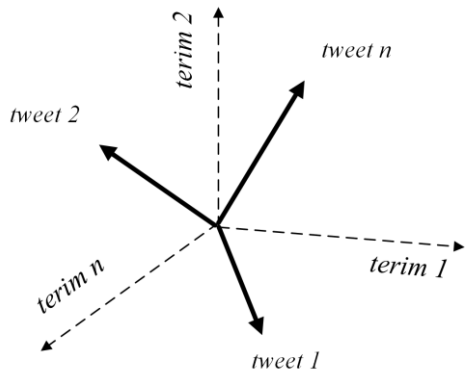
gözlenen ilk 50 kelime incelenerek pozitif kelimelerin yoğun olduğu konu pozitif, negatif kelimelerin yoğun olduğu konu negatif olarak değerlendirilmiştir. Son aşamada ise veri kümesindeki her bir mesaj otomatik olarak konu modeline verilmiş ve hangi konuya daha yakın olduğunu gösteren olasılık bilgisi (0-1 aralığında) elde edilmiştir. Yöntemin son adımında ise elde edilen bu orana göre mesajlar yakın olduğu konunun etiketiyle etiketlenmiştir. Örneğin, mesaj 0.8 oranında negatif 0.2 oranında pozitif konu yakınlığına sahip ise daha yakın olduğu konunun etiketi olan negatif ile etiketlenmiştir. Duygu sınıflandırması AS140 ve TTM veri kümelerinin her ikisi için de uygulanmıştır. Böylece, etiketleme işleminin his simgelerinden bağımsız olması ve başarının artırılması hedeflenmiştir.

3.4 Öznitelik elde etme

Öznitelik elde edilirken, *N-gram* ve kelime torbası (*BoW*, Bag of Words) model olmak üzere iki farklı model kullanılmıştır. Elde edilen öznitelikler ise Şekil 1'de verilen Vektör Uzay Modeli kullanılarak indekslenmiştir [31].

3.4.1 Kelime torbası model

Kelime torbası modelde öznitelik vektörü olarak temsil edilen her bir örnek kelime-frekans ilişkilendirmesi yapılarak indekslenir. Metin içeriğinden çıkarılan her bir kelime öznitelik, kelimenin gözlenme sıklığı ise öznitelik değeri olmaktadır [32]. Çalışmamızda, kelime torbası modelde öznitelik uzay boyutunun artmasını engellemek amacıyla durak kelimeler çıkarılmış ve kelimeler köküne indirgenmiştir. Duygu analizi çalışmalarında kullanılabilecek standart bir durak kelime listesi mevcut değildir. Bu nedenle, çalışmamızda *Lucene* kütüphanesinde bulunan durak kelime listeleri kullanılmıştır. Kelimelerin köküne indirgenmesi (gövdeleme) aşamasında ise İngilizce ve Türkçe için sırasıyla *Porter Stemmer* algoritması ve Türkçe doğal dil işleme kütüphanesi olan *Zemberek* kullanılmıştır [33],[34].



Şekil 1: Vektör uzay modeli.

3.4.2 N-gram model

Çalışmamızda, genellikle kelime düzey modelden daha başarılı olan karakter düzey N-gram model kullanılmıştır [35]. Karakter seviye N-gram modelde öznitelikler n karakter uzunluğunda karakter katarlarından meydana gelmektedir. Böylece karakter düzey N-gram öznitelikleri dilden bağımsız olmakta, yazım yanlışları ve kısaltma kullanımı gibi durumlara karşı güçlü olmaktadır. Ayrıca, farklı seviyelerde (bigram, trigram, fourgram) bilgi çıkarılmasına da imkân vermektedir [36].

3.5 Terim ağırlıklandırma

Makine öğrenmesinin kullanıldığı süreçlerde metinsel veri, terim ağırlıklarının hesaplanması ve terim-doküman matrisinin oluşturulması için öznitelik vektörleri ile temsil edilir. Vektör Uzay Modelinde terim-doküman matrisi ve bir doküman ise aşağıdaki gibi temsil edilir. Denklem (1) ve Denklem (2)'de d_i bir dokümanı ve t_i ilgili dokümanda gözlenen bir terimi temsil eder [14].

$$A = [d_1, d_2, \dots, d_n] \quad (1)$$

$$d_i = (t_{1i}, t_{2i}, \dots, t_{mi})^T \quad (2)$$

Terim ağırlıklandırma ise bir terimin bir dokümandaki önemini belirtme amaçlıdır [5]. Bu nedenle, doküman temsilinin doğru ve etkili olmasında terim ağırlıklandırma önemli bir role sahiptir.

3.5.1 Geleneksel yöntemler

Çalışmamızda yeni yaklaşımlar ile kıyaslama yapabilmek amacıyla geleneksel yöntemlerden olan ikili (BINARY), Terim Frekansı (TF) ve klasik Terim Frekansı-Ters Doküman Frekansı (TFIDF) ile birlikte TF varyasyonu olan aşağıdaki yöntemler de kullanılmıştır [6],[37].

- ✓ Genişletilmiş Normalize Terim Frekansı (ANTF),
- ✓ Ölçeklenmiş Terim Frekansı (STF),
- ✓ Logaritmik Normalize TF (LTF),
- ✓ Okapi BM25.

3.5.2 Yeni yaklaşımlar

Bu bölümde, çalışmamızda uygulanan yeni yaklaşımlar kısaca açıklanmış ilgili yöntemlere ait matematiksel notasyonlar ise Tablo 1'de verilmiştir. Denetimli öğrenme modelinin kullanıldığı metin sınıflandırma ve duygu analizi çalışmalarında sınıf bilgisi önceden bilinmektedir. Bu nedenle, belirtilen alanlarda yeni yaklaşımlar etiketlenmiş eğitim verisinden öğrenme odaklı olmaktadır [12]. Bu yöntemlerden olan *BTWS* (Balanced Term Weighting Scheme), diğer yaklaşımların aksine dokümanda gözlenmeyen terimleri de dikkate alarak dokümanda gözlenen ve gözlenmeyen terimler için ayrı stratejiler kullanır [14]. *CDW* (Categorical Difference Weights) yöntemi sınıf bilgisine dayalıdır ve özellikle iki sınıflı duygu sınıflandırması çalışmaları için önerilmiştir [11]. *TM2* bir terimin doküman frekansının yanı sıra terimin farklı kategorilerdeki dokümanlarda gözlenme frekansını da değerlendirir [38]. *STWS* (Supervised Term Weighting Scheme) duygu analizi için önerilen, ITD (Importance of a term in a document) ve ITS (Importance of a term for expressing sentiment) faktörlerine bağlı olarak terimin ağırlığını belirleyen bir yöntemdir [12]. *ITW* (Improved Term Weighting) yöntemi öznitelikler arası ve öznitelik ile sınıflar arası dağılımı da dikkate alan geliştirilmiş bir TFIDF varyasyonudur [15].

PTW (Probability based Term Weighting) yöntemi klasik TFIDF yöntemindeki IDF yerine özniteliğin bir sınıftaki etkisini gösteren değerle değiştirilmesini öneren olasılık tabanlı bir yöntemdir [16]. *NTW* (Novel Term Weighting) metodu öznitelik seçme yaklaşımı kullanılarak önerilmiş ve duygu analizinde TFIDF yönteminden daha başarılı olduğu gösterilmiştir bir yöntemdir [13]. *TFRF* (Term Frequency-Relevance Frequency) klasik TFIDF yöntemindeki IDF yerine RF (Relevance Frequency) parametresinin kullanıldığı yöntemdir [8]. *RelDF* (Relative Document Frequency) doküman filtreleme için önerilen ve belirli bir konudaki terimlerin ilgili dokümanlarda daha fazla gözlenmesi gerektiğini varsayan bir yöntemdir [39]. *TFICF* (Term Frequency-Inverse Corpus Frequency) yöntemi terimin global frekansından bağımsızdır ve dinamik doküman kümeleme için önerilmiştir [9].

Tablo 1: Denetimli terim ağırlıklandırma yaklaşımları ve matematiksel notasyonları.

Yöntem	Matematiksel Notasyon	Yöntem	Matematiksel Notasyon
BTWS	$W_{BTWS_{t,d}} = \frac{f_i . log_2(\frac{n}{n_i}+1)}{\sqrt{\sum_j^t f_j^2 (log_2(\frac{n}{n_j}+1))^2}}$	IQIW	$W_{IQIW_{t,d}} = log\left(\frac{N}{tp+fn}\right).\log(tp+1).log\left(\frac{\lvert C\rvert}{cf}+1\right)$
CW	$W_{CW_{t,d}} = \log(tf_{t,d}+1).maxstr(t)$	STWS	$W_{STWS_{t,d}} = \left(0.5+\frac{0.5.f_{ij}}{\max_k f_{kj}}\right).max\{OR(f_i,D^1), OR(f_i,D^2)\}$
TM2	$W_{TM2_{t,d}} = \sum_{j=1}^{\lvert C\rvert}\left(\hat{P}(c_j\lvert t_i) - \hat{P}(c_j)\right)^2$	TWIE	$W_{TWIE_{t,d}} = TfIdf(t,d).N_{ct}.\frac{\log((D_{tc'})'+1.5)}{\log(D_{tc'}+1.5)}$
TFPDF	$W_{TFPDF_{t,d}} = \sum_{c=1}^{c=D}\frac{F_{jc}}{\sqrt{\sum_{k=1}^{k=K} F_{kc}{}^2}}\exp(\frac{n_{jc}}{N_c})$	TICW	$W_{TFIDFCF_{t,d}} = \log(tf_{ij}+1).\log\left(\frac{N+1}{n_j}\right).\frac{n_{cij}}{N_{ci}}$
DTFIDF	$W_{DTFIDF_{t,d}} = C_{t,d}.log_2(\frac{N_t}{P_t})$	TFICF	$W_{TFICF_{t,d}} = lof(1+f_{ij}).log\left(\frac{N+1}{n_j+1}\right)$
PTW	$W_{PTW_{t,d}} = tf.log\left(1+\frac{A}{B}\frac{A}{C}\right)$	CDW	$W_{CDW_{t,d}} = \frac{MCPD}{(PositiveDF+NegativeDF)}$
NTW	$W_{NTW_{t,d}} = \frac{1}{\sum_{i=1}^{L} T_{ji}}\left(R_j - \frac{1}{L-1}\sum_{\substack{i=1,\\ i\neq S_j}}^{L} T_{ij}\right)$	ITW	$W_{ITW_{t,d}} = tf_{ik}.log\left(\frac{N}{n_k}+0.01\right).\frac{ttn_k}{tc_k.\lvert C_i\rvert}$
TFRF	$W_{TFRF_{t,d}} = log\left(2+\frac{a}{c}\right)$	QFICF	$W_{QFICF_{t,d}} = \log(tp+1).log\left(\frac{\lvert C\rvert}{cf}+1\right)$
RelDF	$W_{RelDF_{t,,d}} = P(t\lvert rel) - P(t) = \frac{r}{R} - \frac{n}{N}$	CITW	$W_{CITW_{t,d}} = tf_{ij}.\left(1+log\frac{D}{d(t_i)}\right).\left(1+log\frac{C}{CS_\delta(t_i)}\right)$

IQIW bu çalışmada, otomatik soru kategorizasyonu için önerilen QFICF yöntemi geliştirilerek elde edilen ve geleneksel yöntemlerden daha başarılı olduğu gösterilen IQFQFICF (Inverse Question Frequency-Question Frequency-Inverse Category Frequency) yönteminin kısaltmasıdır [20]. *TFPDF* (Term Frequency-Proportional Document Frequency) konu çıkarma çalışmalarında ilgili konudaki belirleyici terimlerin ağırlığını artırmak için önerilmiş bir yöntemdir [40]. *TWIE* (Term Weighting Method for Identifying Emotions) yöntemi mesaj veya dokümanlarda duygu tespiti için önerilmiş ve TFIDF yönteminden daha başarılı olmuştur [41]. *TICW* (TFIDFCF, Term Frequency-Inverse Document Frequency-Class Frequency) yöntemi klasik TFIDF yöntemine sınıf frekansı parametresi eklenerek elde edilmiş ve metin sınıflandırmada geleneksel yöntemlerden daha başarılı olmuştur [42]. *CW* (ConfWeigh) istatistiksel güven aralıklarına dayalı metin sınıflandırma için önerilmiş bir yöntemdir [43]. *DTFIDF* (Delta-TFIDF) duygu içeren terimlerin ağırlığını artıran ve duygu analizi çalışmaları için öneren bir yöntemdir [44].

CITW (Class Indexing based Term Weighting) ilgili sınıftaki sık gözlenen ve gözlenmeyen terimlerin ayırt edilmesini sağlayan bir yöntemdir [45]. *DBM25* (Delta BM25) yöntemi ise geleneksel yöntemlerden Okapi BM25 ve yeni yaklaşımlardan DTFIDF yönteminin kombinasyonu ile elde edilmiş ve kısa metinler üzerinde TF ve varyasyonlarından daha başarılı olmuştur [3]. Tablo 1'de verilen yöntemler ilgili çalışmalarda kullanılan matematiksel notasyonları ile verilmiştir. Ancak, yöntemler incelendiğinde ağırlık hesabında kullanılan bazı parametrelerin ortak olduğu görülmektedir. Verilen bu matematiksel notasyonlarda W ağırlıklandırma yöntemini, t bir terimi ve d bir dokümanı (bu çalışmada Tweet) göstermek üzere bu ortak parametrelerden;

- ✓ n ve N toplam doküman sayısını,
- ✓ $n_{i,j,k}$ ve $d(t_i)$ terimin doküman frekansını,
- ✓ $tf_{t,d}$, $f_{i,j}$, tf terimin terim frekansını,
- ✓ C kategori setini,
- ✓ r, a, A ve tp terimin ilgili kategorideki doküman frekansını,
- ✓ c ve fn terimin diğer kategorilerdeki doküman frekansını,
- ✓ B ilgili kategoride terimi içermeyen doküman sayısını,
- ✓ cf terimin sınıf frekansını temsil eder.

Diğer parametrelerle ilgili detaylı bilgiler ilgili çalışmalarda bulunabilir.

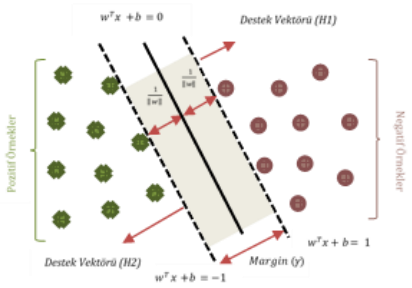
3.6 Sınıflandırma

Veri kümesi sayısallaştırıldıktan sonra bir öğrenme algoritmasına verilmiştir. Test aşamasında ise bu öğrenme algoritması ile eğitilmiş model kullanılarak gelen her bir yeni mesaj sınıflandırılmıştır. Çalışmamızda sınıflandırıcı olarak duygu analizi çalışmalarında yaygın olarak kullanılan Destek Vektör Makinesi (*DVM*) algoritması kullanılmıştır.

3.6.1 Destek vektör makinesi

DVM sınıflandırma problemlerinin çözümü için önerilmiş bir makine öğrenmesi yöntemidir [46]. Temelde iki sınıflı problemleri çözmekle ilgilenir ve sınıflandırma problemlerinde

doğrusal ve doğrusal olmayan olmak üzere iki farklı türü mevcuttur [47]. DVM yönteminde temel amaç, iki sınıfı birbirinden ayırt edebilecek ve her iki sınıfa maksimum uzaklıkta bir aşırı düzlem veya diğer adıyla doğrusal sınıflandırma fonksiyonunu elde etmektir. Pozitif ve negatif kategorideki örnekler sırasıyla +1 ve -1 ile temsil edilmiş ise sınıflandırıcı fonksiyonu Denklem (3) ile formüle edilir. Amaç sınıflandırıcı fonksiyonunu x örneği pozitif ise $sign(f(x)) = +1$, negatif ise $sign(f(x)) = -1$ olacak şekilde elde etmektir. İki sınıflı (örneğin negatif, pozitif vb. gibi) bir doğrusal sınıflandırma problemi üzerinde DVM algoritmasının çalışma prensibi Şekil 2'de verilmiştir.



Şekil 2: DVM ile iki sınıflı doğrusal sınıflandırma.

Şekil 2'den anlaşılacağı üzere iki sınıfı birbirinden ayıran farklı ve çok sayıda vektör mevcuttur. DVM bu vektörlerden destek vektörleri arasında maksimum mesafeye sahip aşırı düzlemi bulmayı amaçlamaktadır.

$$f(x) = w^T x - b \tag{3}$$

Pozitif ve negatif kategorilerden birer örneğin ayırıcı düzleme olan uzaklığı sırasıyla $wx^+ + b = +1$ ve $wx^- + b = -1$ olmak üzere maksimum mesafe aşağıdaki gibi olur.

$$\gamma = \frac{w(x^+ - x^-)}{\|w\|} = \frac{2}{\|w\|} \tag{4}$$

Denklem (4)'te γ değerinin maksimum olması, $y \in \{+1, -1\}$ kategorileri göstermek üzere;

$$eğer\ y_i = +1\ ise\ w^T x - b \geq +1 \tag{5}$$
$$eğer\ y_i = -1\ ise\ w^T x - b \leq -1 \tag{6}$$

Denklem (5) ve Denklem (6)'da verilen kısıtları sağlayacak şekilde $\frac{1}{2}\|w\|^2$ değerinin minimum olarak elde edilmesiyle gerçekleşmektedir. Burada $\|w\|$, ağırlık vektörü olarak adlandırılan w normal düzleminin normudur. Bu işlem bir optimizasyon problemi olarak ele alınır ve ℓ örnek sayısı olmak üzere w aşağıdaki eşitlik ile elde edilir.

$$y_i(w^T x_i - b) \geq 1, \forall_i = 1, \ldots, \ell \tag{7}$$

Denklem (7) ile iki sınıfı birbirinden en iyi şekilde ayırt edebilecek aşırı düzlem öğrenilmiş olur. Bu aşamadan sonra gelen her yeni örnek Denklem (8) ile sınıflandırılır. Sonuç sıfırdan büyük ise kategori pozitif, küçük ise negatif olarak belirlenir.

$$f(x) = sign(w^T x_{yeni} + b) \tag{8}$$

Doğrusal olmayan DVM ise verinin doğrusal olarak sınıflandırılamadığı durumlarda kullanılır. Bu durumda veri bir çekirdek fonksiyonu aracılığıyla daha yüksek boyutlu bir uzaya taşınır ve sınıflandırma bu şekilde yapılır. Doğrusal olmayan DVM'de bu amaçla Doğrusal, Sigmoid, Polinomial ve Radyal Tabanlı Fonksiyon gibi çekirdekler kullanılır [48]. DVM ikili sınıflandırma problemlerini çözmek için geliştirilmiş olsa da çok kategorili sınıflandırma problemlerinde de başarıyla uygulanabilmektedir. Çok kategorili sınıflandırma işleminin uygulanabilmesi için bire-karşı-bir (one-against-one) ve bire-karşı-hepsi (one-against-all) gibi yöntemler önerilmiştir [49]. Çalışmamızda DVM, çok kategorili problemlerde bire-karşı-bir yöntemini kullanan LibSVM kütüphanesi ile uygulanmıştır [50].

4 Araştırma bulguları

TTM ve AS140 veri kümeleri yöntemlerde açıklanan önişleme adımlarından geçirilmiştir. Her iki veri kümesi ile ilgili istatistiksel bilgiler ve önişlemden sonra elde edilen öznitelik bilgileri Tablo 2'de verilmiştir. Önişlemden sonra mesajlar duygu sınıflandırması işlemine tabi tutulmuş ve böylece mesajlar etiketlenmiştir. Bu aşamadan sonra, his simgeleri ve özel terimler sınıflandırma sonuçlarını etkileyebileceği düşünülerek veri kümelerinden çıkarılmıştır. His simgeleri ve özel terimlerin çıkarılmasıyla terim frekans filtresi uygulanmamasına rağmen bazı mesajlardan terim elde edilememiş ve bu mesajlar da veri kümelerinden çıkarılmıştır. Böylece TTM ve AS140 veri kümelerinden sırasıyla 1240 ve 92 adet mesaj elenmiştir. Tablo 2'de verilen istatistikler incelendiğinde ise Türkçe mesajların daha fazla özel ve normalize edilen terim içerdiği tespit edilmiştir. Kelime torbası ve N-gram modelde öznitelikler gövdeleme ve durak kelime çıkarımı işlemleri uygulandıktan sonra elde edilmiştir. N-gram modelde genellikle karakter seviyesi arttıkça başarı arttığından *trigram* öznitelikleri kullanılmıştır.

AS140 verisinde, daha önce ilgili çalışma kapsamında his simgeleri çıkarılmış olmasına rağmen çok az sayıda his simgesi gözlenmiştir. BoW ve N-gram modellerinde elde edilen öznitelik sayıları incelendiğinde AS140 verisinden yaklaşık olarak iki kat daha fazla öznitelik elde edilmiştir. Bu durum AS140 verisinde ki mesajların TTM verisindeki mesajlardan içerik bakımından daha zengin olduğunu göstermektedir. Her iki veri kümesi de başlangıçta aynı sayıda mesaj içermesine rağmen elenen mesaj sayısı ve elde edilen öznitelik sayılarında farklılıklar oluşmuştur. Bu durumun temel nedeni, dil farkının yanı sıra her iki veri kümesinin farklı zamanlarda ve sorgu anahtarı olarak sadece his simgeleri kullanılarak rastgele toplanmış olmasındandır. Bunun yanı sıra, her iki veri kümesi duygu sınıflandırmasına tabi tutulmuş ve GDA yöntemi ile yeniden etiketlenmiştir. Böylece, örnek kategorilerinin his simgelerinden bağımsız olması sağlanmış ve sınıflandırma başarısının artırılması amaçlanmıştır.

Tablo 2: TTM ve AS140 veri kümeleri ile ilgili istatistikler.

Veri Kümesi	Tweet	Hashtag {#}	Kullanıcı Adı {@}	URL {http://}	His Simgesi { :), -), :(}	Normalize Edilen Terim	Toplam Terim		Toplam Benzersiz Terim		Ortalama Terim	
							BoW	N-gram	BoW	N-gram	BoW	N-gram
TTM	20000	1512	12436	3463	21944	2530	74588	339157	4579	5058	3.7	16.9
AS140	20000	349	9792	977	13	1763	154543	782576	15128	8361	7.7	39.1

Duygu sınıflandırmasından sonra her iki veri kümesi için elde edilen kategori örnek dağılımı Şekil 4'te verilmiştir. Şekil 4'ten anlaşılacağı üzere duygu sınıflandırmasından sonra veri kümelerindeki kategori-örnek dağılımı değişmiştir. Her iki veri kümesinde de negatif kategorili mesaj sayısının daha fazla olduğu tespit edilmiştir. Böylece önceden his simgeleri kullanılarak etiketlenen (mesajlar toplanırken içerdiği his simgesine göre pozitif veya negatif etiketli olduğu varsayıldığında) veri kümelerinde çoğunlukla negatif mesajların yanlış bir şekilde pozitif olarak etiketlendiği görülmüştür. Bu durumun bir sonucu olarak ise her iki veri kümesi için de kategoriler arası dengesizlik meydana gelmiştir. Duygu sınıflandırması aşamasında etiketlemenin yanı sıra pozitif ve negatif topiklerde en sık gözlenen kelimeler de tespit edilmiştir. TTM verisi için pozitif ve negatif konularda en sık gözlenen gövdelenmiş kelimeler ise Şekil 3'te verilen kelime bulutu ile gösterilmiştir. Bu kelimeler incelendiğinde her iki veri kümesi için de GDA ile etiketleme işleminin deneysel sonuçlarla örtüşecek şekilde oldukça başarılı olduğu tespit edilmiştir.

4.1 Deneysel sonuçlar

Bu bölümde, geleneksel ağırlıklandırma yöntemleri ve Tablo 1'de verilen yeni yaklaşımların performansları ve sınıflandırma sonucu üzerindeki etkileri incelenmiştir. *LibSVM* paketiyle uygulanan DVM yönteminde tüm parametreler için ön tanımlı değerler ve doğrusal çekirdek kullanılmıştır. GDA yöntemi ise *MALLET* makine öğrenmesi kütüphanesi ile uygulanmıştır. Deneysel sonuçlar elde edilirken *hold-out* olarak isimlendirilen geçerleme yöntemi kullanılmış ve veri kümesinin %30'u test, %70'i eğitim seti olacak şekilde bölünmüştür [51]. BTWS yöntemi vektör olarak temsil edilen bir dokümanda gözlenen ve gözlenmeyen terimlere ağırlık atayabilmektedir. Ancak dokümanda gözlenmeyen terimlere de ağırlık atandığında sınıflandırma sonucunda çok önemli bir değişiklik olmamıştır. Bu nedenle, BTWS yönteminde sadece dokümanda gözlenen terimlere ağırlık atanmıştır. STWS yönteminde ITD ve ITS parametreleri için en başarılı sonuçların elde edildiği kombinasyon kullanılmıştır. RelDF yönteminde R değeri ilgili çalışmada belirtildiği gibi 1 olarak alınmıştır. BM25 ve DBM25 yöntemlerinde k_1 ve b değerleri sırasıyla 1.2 ve 0.95 olarak alınmıştır [52]. Bu yöntemlerde doküman uzunluğu ise kelime torbası veya N-gram özniteliklerinin uzunlukları toplamı olarak kabul edilmiştir. Deneysel sonuçlar elde edilirken, özellikle bazı yeni yaklaşımlarda bulunan logaritma çarpanından dolayı ağırlık değerlerinin NaN veya Infinity değerlerini alabildiği tespit edilmiştir. Bu durum, Twitter mesajlarından oluşan veri kümelerinin çok seyrek yapıda olması ve 140 karakter sınırından dolayı terim frekanslarının düşük olmasından kaynaklanmaktadır. Çalışmamızda, bu durumdaki terimler için ilgili ağırlıklandırma yöntemindeki ağırlık değeri sıfır olarak kabul edilmiştir. Ağırlıklandırma yöntemleri her iki veri kümesi üzerinde uygulandıktan sonra CITW, TICW ve TM2 yöntemlerinde terim ağırlıklarının çok büyük bir oranda bahsedilen nedenlerden dolayı sıfır olarak atandığı gözlenmiştir. Veri kümesinin seyrek olması ve kategori sayısının iki olması bu üç yöntemde ağırlık değerinin sıfıra indirgenmesinde oldukça etkili olmuştur. Ayrıca duygu sınıflandırmasından sonra veri kümelerinde dengesiz kategori dağılımının elde edilmesi (Şekil 4) en büyük etken olmuştur. Bu nedenle, bu yöntemlerin özellikle dengesiz veri kümeleri üzerinde uygulanan iki kategorili duygu analizi için uygun olmadığı tespit edilmiştir. Sınıflandırıcının performans değerlendirmesinde ise Denklem (9) ile elde edilen *doğruluk (Acc)* metriği kullanılmıştır [53].

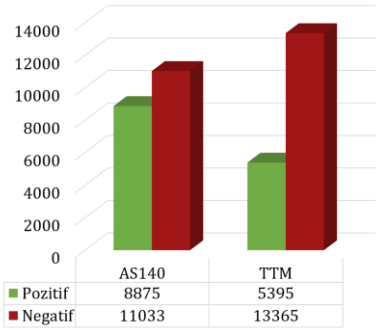
$$Acc = \frac{(TP + TN)}{N} \tag{9}$$

Yukarıdaki eşitlikte TP, TN ve N sırasıyla kategori ataması doğru yapılan, kategorisi doğru reddedilen ve test kümesindeki toplam örnek sayılarını temsil eder. TTM ve AS140 veri kümeleri üzerinde geleneksel yöntemler ve Tablo 1'de verilen ağırlıklandırma yöntemleri ile elde edilen sınıflandırma başarıları ise sırasıyla Şekil 5 ve Şekil 6'da verilmiştir. Deneysel sonuçlar incelendiğinde, genel olarak ağırlıklandırma yöntemlerinin her iki veri kümesi üzerinde benzer davranışlar sergilediği görülmüştür. Bu bağlamda yeni yaklaşımlar, genellikle geleneksel yöntemlerden daha başarılı olmuştur. Öznitelik modeli açısından değerlendirildiğinde ise kelime torbası model, N-gram modelden genelde daha başarılı olmuştur. Ancak her iki veri kümesi için de PTW ve RelDF yöntemleri kullanıldığında N-gram model daha başarılı olmuştur. Sınıflandırma başarısı açısından her iki veri kümesinde en başarılı sonuçlar *PTW* yöntemiyle elde edilmiştir. En başarılı üç ağırlıklandırma yöntemi ise sırasıyla PTW, ITW ve RelDF yöntemleri olmuştur. Yöntem bazında değerlendirildiğinde kategori sayısının iki olması ve veri kümesinin seyrek olması BTWS, CDW, CW, STFS ve TFPDF gibi yeni yaklaşımlarda sınıflandırma başarısının geleneksel yöntemler ile elde edilen başarıyla aynı düzeyde kalmasına neden olmuştur. Tüm deneylerde her iki veri kümesi için de en düşük başarı *TFPDF* ağırlıklandırma yöntemi ile elde edilmiştir. Bu durum TFPDF yöntemi sınıf frekansı ile doğru orantılı iken veri kümesinin iki kategorili olmasından ileri gelmiştir. Yani sınıf sayısı az olduğundan, TFPDF yönteminde terimlerin ağırlık değerleri arasında ayırt edici seviyede yüksek bir fark oluşmamıştır.



Şekil 3: TTM verisi için pozitif ve negatif konularda en sık gözlenen kelimeler.

Elde edilen genel başarı oranı açısından ise ağırlıklandırma yönteminin yanı sıra konu bilgisine dayalı duygu sınıflandırmasının da sonuçlar üzerinde olumlu etkisinin olduğu görülmüştür. Önceki çalışmalarda *klasik yöntemlerle* elde edilen başarı oranının yaklaşık olarak %87 olduğu [30] değerlendirildiğinde çalışmamızda duygu analizi başarısı %10 oranında daha yüksek (PTW yönteminin her iki veri seti üzerinde göstermiş olduğu sırasıyla %97.6 ve %97.5 başarı oranları göz önünde bulundurulduğunda) olmuştur.

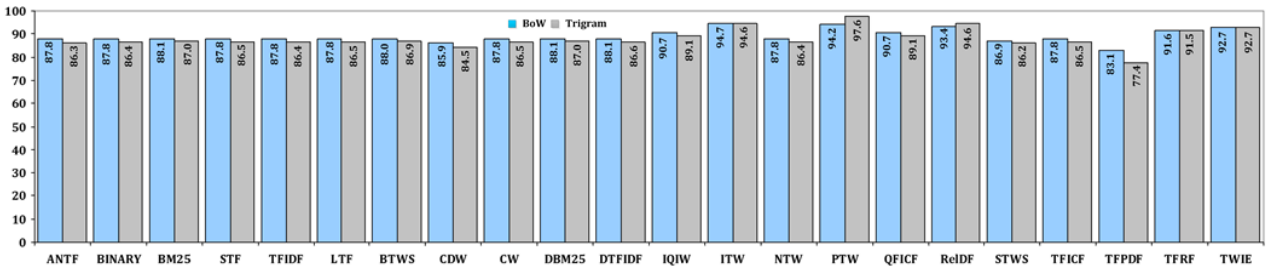


Şekil 4: GDA tabanlı yarı otomatik duygu sınıflandırmasından sonra AS140 ve TMM veri kümeleri için kategori-örnek dağılımı.

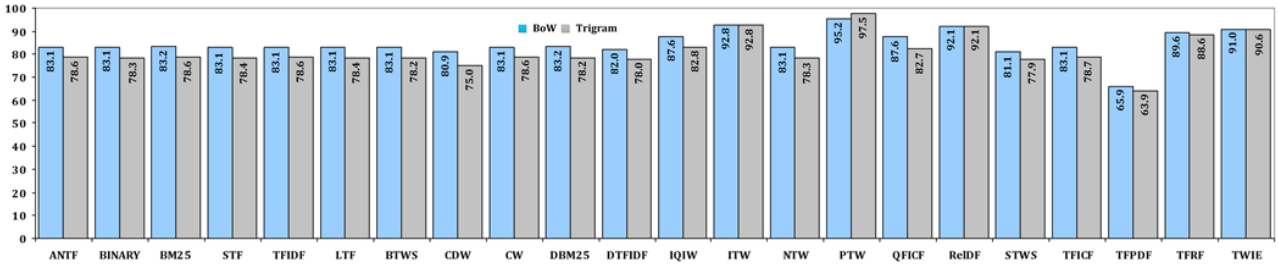
Ağırlıklandırma yöntemlerinin sınıflandırma sonucu üzerindeki etkisinin yanı sıra performans bakımından da analizi gerçekleştirilmiştir. Böylece en hızlı yöntemin tespit edilmesi amaçlanmıştır. Bu bağlamda, her bir yöntem için elde edilen çalışma zamanı *milisaniye* (x1000) cinsinden Şekil 7'de verilmiştir. Tüm yöntemler *Java* programlama dili kullanılarak uygulanmıştır. Java programlama dilinde sınıfların yüklenmesi vb. gibi nedenlerden dolayı programın her koşmasında çalışma zamanında küçük farklılıklar görülebilmektedir. Bu nedenle çalışmamızda her bir ağırlıklandırma yöntemi üç kez koşturulmuş ve ortalama koşma süreleri dikkate alınmıştır. Koşma süreleri incelendiğinde, geleneksel yöntemlerin yeni yaklaşımlardan daha hızlı olduğu ve öznitelik uzay boyutu arttıkça aradaki hız farkının da arttığı görülmektedir. Yeni yaklaşımlar uygulanırken yöntemin ağırlık hesaplama aşamasında ihtiyaç duyacağı tüm istatistiksel veriler önceden hesaplanmış verilen sürelere bu hesaplama zamanı eklenmemiştir. Bu bakımdan değerlendirildiğinde, geleneksel yöntemlerin yeni yaklaşımlara göre her koşulda daha hızlı olduğu açıktır.

4.2 Sonuç ve gelecek çalışmalar

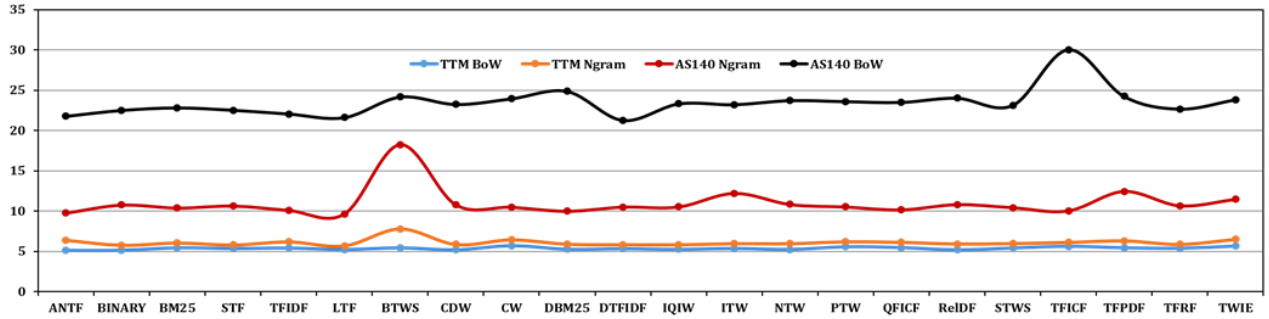
Bu çalışmada bilgi getirimi, doküman benzerliği, konu çıkarma ve metin sınıflandırma çalışmalarında yaygın olarak kullanılan geleneksel yöntemler ve bu alanlar için yakın zamanda önerilmiş yeni yöntemler incelenmiştir.



Şekil 5: Geleneksel yöntemler ve yeni yaklaşımlar kullanılarak TTM verisi için elde edilen sınıflandırma başarıları (%).



Şekil 6: Geleneksel yöntemler ve yeni yaklaşımlar kullanılarak AS140 verisi için elde edilen sınıflandırma başarıları (%).



Şekil 7: Ağırlıklandırma yöntemlerinin farklı öznitelik modellerinde TTM ve AS140 verileri üzerinde koşma süreleri (x1000ms).

Bunun yanı sıra, özellikle duygu analizi için önerilmiş yöntemler de yeni yaklaşımlara dahil edilmiştir. Böylece, toplam 25 adet ağırlıklandırma yöntemi (CITW, TICW ve TM2 dahil) Twitter duygu analizi problemine uygulanarak sonuçlar üzerindeki etkisi incelenmiştir. Veri kümesi olarak aynı sayıda Türkçe ve İngilizce Twitter mesajı içeren iki farklı veri kümesi kullanılmıştır. Her iki veri kümesi de GDA yöntemi ile oluşturulan konu model ile duygu sınıflandırmasından geçirilmiştir. Bu aşamadan sonra his simgeleri ve özel terimler mesaj içeriklerinden çıkarılmıştır. Öznitelikler ise kelime torbası ve N-gram olmak üzere iki farklı model ile elde edilmiştir.

Deneysel sonuçlar incelendiğinde, geleneksel yöntemlerin daha hızlı olduğu ancak hem öznitelikler arası hem de öznitelik-sınıf arası ilişkiyi dikkate alan yeni yaklaşımların genellikle daha başarılı olduğu görülmüştür. Bu yöntemlerden CITW, TICW ve TM2'nin dengesiz veriler üzerinde uygulanan iki kategorili duygu analizi için uygun olmadığı tespit edilmiştir. Bu durumun ise sınıf sayısının iki, veri kümesinin seyrek yapıda ve duygu sınıflandırmasından sonra veri kümelerinin dengesiz olmasından kaynaklandığı düşünülmektedir. Duygu analizi probleminde en başarılı ağırlıklandırma yönteminin ise PTW olduğu tespit edilmiştir. PTW yöntemi öznitelik seçme tabanlı ve doğrudan sınıf frekansına bağlı olan bir yöntemdir. Ağırlık hesabında terimin ilgili kategori için ne kadar iyi bir gösterge olduğunu dikkate alır. Bu yönüyle, PTW aslında bir öznitelik seçme yöntemi olarak da kullanılabilir. Bu nedenlerden dolayı PTW diğer yöntemlere göre daha başarılı olmuştur. Kategori sayısının iki olması ise sınıf bilgisini kullanan yöntemlerde, düşük performans gösteren yaklaşımların aksine PTW, ITW ve RelDF gibi yöntemlerde ilgili terime daha yüksek ağırlık atanmasını sağlayan önemli bir etken olmuştur.

Gelecek çalışmalarda benzerlik tabanlı yeni bir ağırlıklandırma yönteminin önerilmesi, bu yöntemin metin sınıflandırma ve duygu analizi alanlarında uygulanarak analiz edilmesi düşünülmektedir.

5 Kaynaklar

[1] Patra A, Singh D. "A survey report on text classification with different term weighing methods and comparison between classification algorithms". *International Journal of Computer Applications*, 75(7), 2013.

[2] Prabowo R, Thelwall M. "Sentiment analysis: A combined approach". *Journal of Informetrics*, 3(2), 143-157, 2009.

[3] Paltoglou G, Thelwall M. "A study of information retrieval weighting schemes for sentiment analysis". *48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, USA, 11-16 July 2010.

[4] Çetin M, Amasyalı MF. "Supervised and traditional term weighting methods for sentiment analysis". *In Signal Processing and Communications Applications Conference (SIU)*, Girne, KKTC, 24-26 April 2013.

[5] Aizawa A. "An information-theoretic perspective of tf–idf measures". *Information Processing & Management*, 39(1), 45-65, 2003.

[6] Salton G, Buckley C. "Term-weighting approaches in automatic text retrieval". *Information processing & management*, 24(5), 513-523, 1988.

[7] Robertson S, Zaragoza H, Taylor M. "Simple BM25 extension to multiple weighted fields". *13th ACM International Conference on Information and Knowledge Management*, New York, USA, 08-13 November 2004.

[8] Lan M, Tan CL, Low HB. "Proposing a new term weighting scheme for text categorization". *Association for the Advancement of Artificial Intelligence*, Boston, USA, 16-20 June 2006.

[9] Reed JW, Jiao Y, Potok TE, Klump BA, Elmore MT, Hurson A R. "TF-ICF: A new term weighting scheme for clustering dynamic data streams". *In ICMLA'06. 5th International Conference on Machine Learning and Applications*, Florida, USA, 14-16 December 2006.

[10] Polettini N. "The vector space model in information retrieval-term weighting problem". *Entropy*, 1-9, 2004.

[11] Chen LS, Chang CW. "A new term weighting method by introducing class information for sentiment classification of textual data". *International Multi Conference of Engineers and Computer Scientists*, Hong Kong, China, 16-18 March 2011.

[12] Deng ZH, Luo KH, Yu HL. "A study of supervised term weighting scheme for sentiment analysis". *Expert Systems with Applications*, 41(7), 3506-3513, 2014.

[13] Gasanova T, Sergienko R, Akhmedova S, Semenkin E, Minker W. "Opinion mining and topic categorization with novel term Weighting". *5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Baltimore, Maryland, USA, 27 June 2014.

[14] Jung Y, Park H, Du D. "A Balanced term-weighting scheme for improved document comparison and classification". *Preprint*, 2001.

[15] Kansheng SHI, Jie HE, Liu HT, Zhang NT, Song WT. "Efficient text classification method based on improved term reduction and term weighting". *The Journal of China Universities of Posts and Telecommunications*, 18(1), 131-135, 2011.

[16] Liu Y, Loh H. T, Sun A. "Imbalanced text classification: A term weighting approach". *Expert Systems With Applications*, 36(1), 690-701, 2009.

[17] Deng ZH, Tang SW, Yang DQ, Li MZLY, Xie KQ. "A comparative study on feature weight in text categorization". *In Advanced Web Technologies and Applications*, Hangzhou, China, 14-17 April 2004.

[18] Mladenić D, Brank J, Grobelnik M, Milic-Frayling N. "Feature selection using linear classifier weights: interaction with classification models". *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, United Kingdom, 25-29 July 2004.

[19] Debole, F, Sebastiani F. *Supervised Term Weighting for Automated Text Categorization*. Editor(s): Spiros S. Text Mining and its Applications, 81-97, Germany, Berlin Heidelberg, Springer, 2004.

[20] Quan X, Wenyin L, Qiu B. "Term weighting schemes for question categorization". *Pattern Analysis and Machine Intelligence*, 33(5), 1009-1021, 2011.

[21] Go A, Bhayani R, Huang L. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, California, USA, Project Report, CS224N, 2009.

[22] Srividhya V, Anitha R. "Evaluating preprocessing techniques in text categorization". *International Journal of Computer Science and Application*, 47(11), 2010.

[23] Brücher H, Knolmayer G, Mittermayer MA. "Document classification methods for organizing explicit knowledge". University of Bern, Switzerland, Technical Report, 140, 2002.

[24] Coban O, Ozyer B, Ozyer G. T. "A comparison of similarity metrics for sentiment analysis on Turkish twitter feeds". *International Conference on SocialCom*, Chengdu, China, 19-21 December, 2015.
[25] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. "Sentiment analysis of twitter data". *In Proceedings of the Workshop on Languages in Social Media,* Portland, Oregon, USA, 23 June 2011.
[26] Kouloumpis E, Wilson T, Moore JD. "Twitter sentiment analysis: The good the bad and the omg!". *International Conference on Web and Social Media*, Barcelona, Catalonia, Spain, 17-21 July 2011.
[27] Kaya M, Fidan G, Toroslu I. H. "Sentiment analysis of turkish political news". *International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, China, 4-7 December 2012.
[28] Walsh, B. "Markov chain monte carlo and gibbs sampling". University of Sao Paulo, Brazil, Lecture Notes for EBB 581, 2004.
[29] Blei DM, Ng AY, Jordan MI. "Latent dirichlet allocation". *The Journal of machine Learning research*, 3, 993-1022, 2003.
[30] Çoban Ö, Özyer G. T. "Sentiment classification for Turkish twitter feeds using LDA". *24th IEEE Signal Processing and Communications Applications Conference (SIU)*, Zonguldak, Turkey, 16-19 May 2016.
[31] Salton G, Wong A, Yang CS. "A vector space model for automatic indexing". *Communications of the ACM*, 18(11), 613-620, 1975.
[32] Lewis DD. "An evaluation of phrasal and clustered representations on a text categorization task". *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 21-24 June 1992.
[33] Akın AA, Akın MD. "Zemberek, an open source NLP framework for Turkic Languages". *Structure*, 10, 1-5, 2007.
[34] Porter MF. "An algorithm for suffix stripping". *Program*, 14(3), 130-137, 1980.
[35] Kanaris I, Kanaris K, Houvardas I, Stamatatos E. "Words versus character n-grams for anti-spam filtering". *International Journal on Artificial Intelligence Tools*, 16(06), 1047-1067, 2007.
[36] Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. "Text classification using string kernels". *The Journal of Machine Learning Research*, 2, 419-444, 2002.
[37] Manning C. D, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Online Edition, Cambridge, United Kingdom, Cambridge University Press, 2008.
[38] Xu H, Li C. "A Novel term weighting scheme for automated text categorization". *7th International Conference on Intelligent Systems Design Applications,* Rio de Janeiro, Brazil, 22-24 October 2007.
[39] Nanas N, Uren V, De Roeck A. "A comparative evaluation of term weighting methods for information filtering". *15th International Workshop on Database and Expert Systems Applications,* Zaragoza, Spain, 3-3 September 2004.
[40] Bun KK, Ishizuka M. "Topic extraction from news archive using TF*PDF algorithm". *In Proceedings of the Third International Conference on Web Information Systems Engineering*, *Singapore*, 14 December, 2002.
[41] De Silva J, Haddela P. S. December. "A term weighting method for identifying emotions from text content". *2013 International Industrial and Information Systems (ICIIS) Conference*, Peradeniya, Sri Lanka, 17-20 December 2013.
[42] Liu M, Yang J. "An improvement of TFIDF weighting in text categorization". *International Proceedings of Computer Science and Information Technology*, IACSIT Press, Singapore, 2012.
[43] Soucy P, Mineau G. W. "Beyond TFIDF weighting for text categorization in the vector space model". *International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, July 30-August 5, 2005.
[44] Ren F, Sohrab MG. "Class-indexing-based term weighting for automatic text classification". *Information Sciences*, 236, 109-125, 2013.
[45] Srividhya V, Anitha R. "Evaluating preprocessing techniques in text categorization". *International Journal of Computer Science and Application*, 2010, 49-51, 2010.
[46] Cortes C, Vapnik V. "Support-vector networks". *Machine learning*, 20(3), 273-297, 1995.
[47] Burges C. J. "A tutorial on support vector machines for pattern recognition". *Data mining and knowledge discovery*, 2(2), 121-167, 1998.
[48] Gunn S. R. "Support Vector Machines for Classification and Regression". Department of Science and Mathematics Engineering, University of Southampton, Southampton, UK, ISIS Technical Report, 14, 1998.
[49] Fradkin D, Muchnik I. "Support vector machines for classification". *Discrete Methods in Epidemiology*, 70, 13-20, 2006.
[50] Chang CC, Lin CJ. "LIBSVM: A library for support vector machines". *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 2011.
[51] Kohavi R. "A study of cross-validation and bootstrap for accuracy estimation and model selection". *International Joint Conference on Artificial Intelligence*, Quebec, Canada, 20-25 August 1995.
[52] Jones KS, Walker S, Robertson SE. "A probabilistic model of information retrieval: development and comparative experiments". *Information Processing & Management*, 36(6), 809-840, 2000.
[53] Sheela LJ. "A Review of Sentiment Analysis in Twitter Data Using Hadoop". *International Journal of Database Theory and Application*, 9(1), 77-86, 2016.

Ek A

Çalışmada kullanılan materyal, açık kaynak kodlu yazılım ve kütüphaneler.

Materyal/Kütüphane	URL	Ref.
MALLET	http://mallet.cs.umass.edu/topics.php	-
Twitter API	https://dev.twitter.com/rest/public/search	-
Lucene	http://lucene.apache.org/	-
Sentiment140	http://help.sentiment140.com/	[21]
Zemberek	https://code.google.com/p/zemberek/	[33]
LibSVM	https://www.csie.ntu.edu.tw/~cjlin/libsvm/	[50]