International Journal of Social Sciences

Uluslararası Sosyal Bilimler Dergisi

Sosyal Medyadaki Film Yorumlarının Fikir Madenliği ile Otomatik Sınıflaması

Migena Ceyhan[1]

Zeynep Orhan[2]

Dimitrios Karras[3]

Özet

Günümüzde dijital yaşam insan ilişkilerini önemli ölçüde değiştirmiştir. Pek çok konuda fikirler, duygular ve düşünceler sanal ortamda paylaşılmaktadır. Bunlar farklı alanlarda kullanılabildiği takdirde önemli ve faydalı uygulamalar geliştirilebilir. Her geçen gün artan uygulamalar da bu düşünceyi desteklemektedir.

Bu çalışma, sosyal medyada kullanıcıların yaptığı film yorumlarından olumlu ve olumsuz olanların karakteristik özelliklerini duygu ve düşünce analiziyle öğrenmekte, daha sonra bunları kullanarak yeni yorumları otomatik olarak sınıflamayı, zamandan ve insan gücünden tasarruf etmeyi, binlerce yorumu saniyeler içinde inceleyerek sonuçları bilgisayar kullanıcısına hızlı ve kolay anlaşılabilir özet şeklinde ulaştırmayı amaçlamaktadır.

Anahtar Kelimeler: Sosyal medya, film yorumu, fikir madenliği, otomatik snıflama

Automatic Classification of Film Comments on Social Media with Intellectual Mining

Summary

Today, digital life has significantly changed human relations. Ideas, emotions and thoughts are shared in a virtual environment on many topics. If these can be used in different areas, important and useful applications can be developed. Increasing applications support this idea.

The aim of this study is to teach the characteristics of the positive and negative ones from the movie comments made by users on social media with emotion and thought analysis, then automatically classify new comments using them, to save time and manpower, to examine the results in seconds, and quickly and easily to the computer user, and to deliver in an understandable summary.

Keywords: Social media, movie interpretation, opinion mining, automatic classification

---

[1]Department of Mathematics and Informatics, University of Shkodra "Luigj Gurakuqi", Shkoder, Albania. E-mail: migena.ceyhan@unishk.edu.al

[2] Computer Science Department, Union College, Schenectady, New York, USA

[3] Computer Engineering Department, Epoka University Tirana, Albania

Giriş

Duygu Analizi ve Düşünce Madenciliği

Duygu analizi (DA), bir dökümanın olumlu, olumsuz ya da nötr şeklinde yorumlanmasıdır. DA için fikir çıkarımı (Opinion Extraction), fikir madenciliği (Opinion Mining), duygu madenciliği (Sentiment Mining), öznellik analizi (Subjectivity Analysis) gibi isimler de kullanılmaktadır 0.

Duygu Analizi Gerekliliği

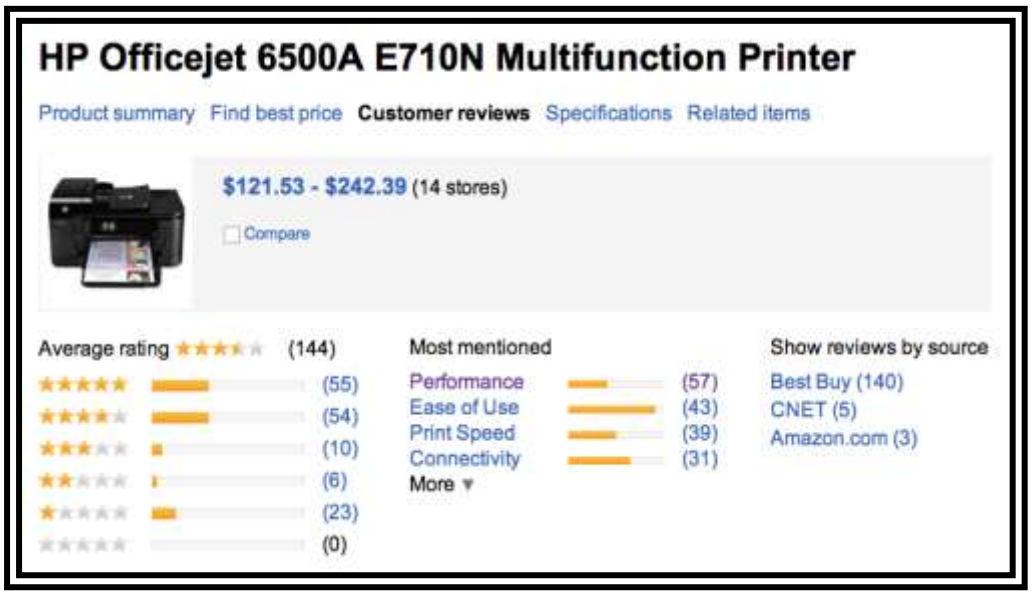
Internetin son yıllardaki gelişimi ile fikirlerin her yerde bulunabilmesi mümkün hale gelmiştir. Bloglar, Facebook, Twitter, haber portalları, e-ticaret siteleri gibi sosyal ağlar bu durumu kolaylaştırmıştır. Bunca bilgi kaynağı her ne kadar faydalı da olsa bu kaynakların çokluğu ve miktarı göz önüne alınınca kullanıcılar için hazmı zor bir bilgi yığını olmaktadır. Örneğin bir şey almak istediğimizde konu ile ilgili pek çok bilgiye internetten ulaşmak mümkündür. Şekil 1, Şekil 2 ve Şekil 3Şekil 3'de görüldüğü gibi tek bir ürün için yüzlerce yorum arasında kaybolmak kullanıcıların kabusudur. Bunun yerine daha anlaşılır bir özet sunmak faydalı olacaktır 0.

Dijital dünyadaki gelişmelerle birlikte hayatımızın her alanında teknoloji kullanılmaya başlamış ve sanal dünyada çok önemli ve değerli veri ortaya çıkmıştır. Bu veri dinamik bir yapıya sahip olduğu için sürekli artmakta ve yeni şekiller almaktadır. Günümüzde insanlar bir filmi, müziği ya da teknolojik cihaz gibi ürünü denedikten sonra internette sosyal medya siteleri, blog veya forumlar aracılığıyla ürün hakkındaki görüşlerini, beğendikleri ve beğenmedikleri kısımları eleştirirler. Özellikle büyük şirketler bu yorumları inceleyerek ortaya koydukları ürünün negatif ve pozitif yönlerini tespit eder, kendilerini geliştirirler. Ayrıca tüketiciler ürünü kullanmadan önce bu yorumlara bakarak genel bir fikir edinirler. Ancak inceleme süreci çok uzun sürmektedir ve çok zahmetlidir. 0

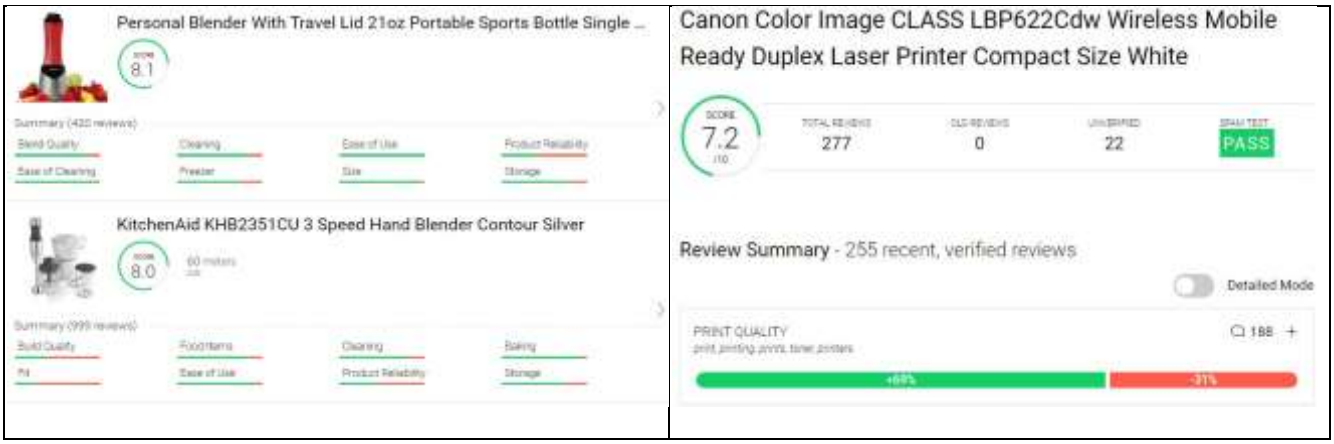
Kaynakların analiz edilip kullanılabilir uygulamalara dönüşmesi bu bilgi okyanusundan en iyi şekilde faydalanılmasını sağlayacaktır. Şekil 4'te de görüldüğü gibi dijital dünyayı kullanan kişi sayısı

milyarlarla ifade edilen seviyelere gelmiştir. Bu alanda yapılacak çalışmaların etkisi ve ekonomik büyüklüğünü anlamak açısından istatistikler önemli bir göstergedir. İstatistikler bize bu konuda daha iyi fikirler vermektedir.0 İnternetteki veri miktarının büyük bir hızla artması ve internet kullanımının ticarete etkisi üretici ve tüketici ilişkilerini de değiştirmektedir. Artık hem üreticiler hem de tüketiciler sosyal medyadan faydalanarak her konuda fikir ve tecrübe paylaşımında bulunmaktadır. Bir konuda, bir ürün veya marka hakkında ilk başvuru kaynağı bu veriler olmaktadır. Sadece Twitter ele alındığında bile günde milyonlarca Tweet atılmakta ve pek çok yeni hesap açılmaktadır. Ancak bu inanılmaz sayılar kullanıcıların takip kapasitesini de aşmaktadır ve yardımcı araçlara ihtiyaç her geçen gün artmaktadır. Bu noktada bilgiyi otomatik yorumlayıp sınıflayacak duygu (sentiment) analizi uygulamaları ortaya çıkmaktadır.0

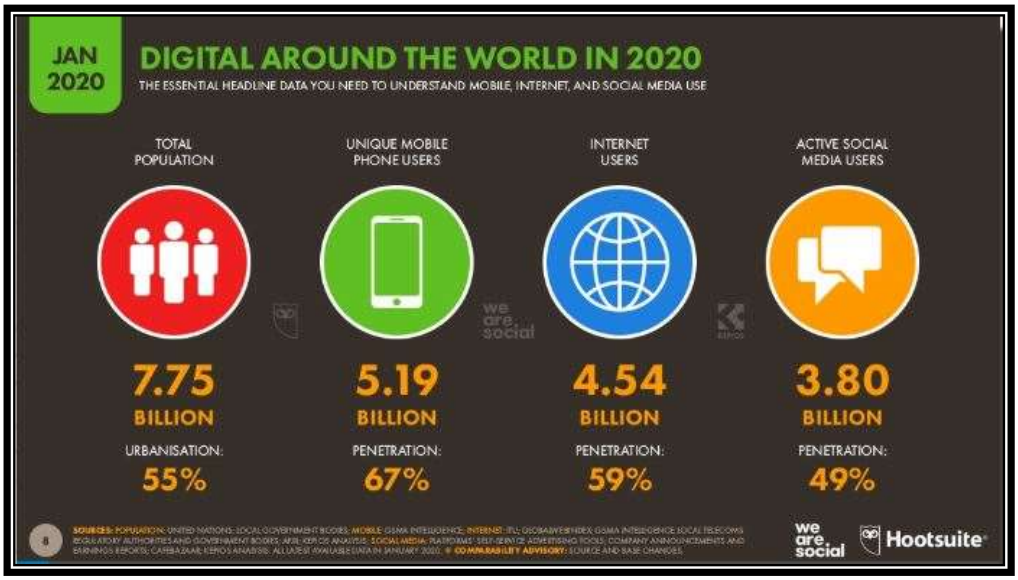
Şekil 1: HepsiBurada ürün yorumları



Şekil 2: Bing ürün yorumları0



Şekil 3 : the**review**index Amazon ürün İnglizce yorumları



Şekil 4 : Dünyada internet, sosyal medya ve mobil kullanım istatistikleri

Duygu Analizi Nerelerde ve Nasıl Kullanılır

Duygu analizi pek çok alanda kullanılmaktadır. Bunlara bazı örnekler0:

Sinema dünyası: Yapılan yorumlar olumlu mu olumsuz mu? Hasılat ne kadar olur? Filmin devamı çekilebilir mi? gibi soruların cevapları aranmakta ve buna göre stratejiler belirlenmektedir.

Ticaret hayatı ve üretim: İnsanlar ürünler hakkında neler düşünüyor? Ürünün beğenilen ve beğenilmeyen yönleri nelerdir? gibi araştırmalara kaynaklık ederek üretim stratejilerine yön verilmektedir.

Kamuoyu yoklaması: Tüketici güveni ne seviyededir? Umutsuzluk artıyor mu? gibi kamunun bir konu hakkındaki görüşleri analiz edilerek uygun politikalar üretilmektedir.

Politika: Politik hayat, özellikle seçim sonuçları tahmini için kullanılmaktadır.

Tahmin: Herhangi bir konuda duygu analizi ile tahmin yapmak için uygun görülmektedir.

Girişimciler ve işadamlarının neden arama işlemleri ve sosyal ağları pazarlama planlarına dahil etmeleri gerektiğine ve bunlar olmadan piyasada kalmalarının neden mümkün olmadığına istatistikler güzel cevap vermektedir. Aşağıda bu konuda bazı çarpıcı veriler bulunmaktadır:

Sosyal ağları ve sosyal medyayı düzenli kullanan kişiler internet kullanıcılarının %84'ünü oluştrumakta ve bu oran hızla büyümektedir.

Google (açık arayla) en popüler arama motorudur. 2020 itibarıyla Google, toplam masaüstü arama trafiğinin %79'undan fazlasını oluşturmaktadır.

Google her gün 5,5 milyardan fazla arama gerçekleştiriyor.

Tüketicilerin %97'si bir şirketi internet üzerinden araştırıyor.

Kullanıcılar aramalardaki sonuçlarda %34 oranında ilk çıkan sonuca yönlenmektedir.

Arama sonuçlarında en üst 4 sonuca yönlenme oranı %83'tür. **Hata! Başvuru kaynağı bulunamadı.**

Olumlu eleştiriler 18-34 yaş aralığındaki gençlere duyulan güvenin %91 oranında artmasına neden oluyor.

Google araması yapan kullanıcıların %72'si, aradıkları yerden 8 km uzaklıktaki bir mağazayı ziyaret etti.

Müşterilerin %88'i bir mobil cihazdan arama yaptıktan sonra 24 saat içinde bir mağazayı aradığını veya ziyaret ettiği biliniyor. **Hata! Başvuru kaynağı bulunamadı.**

DA'nın getirileri için ise şunlar söylenebilir0:

Markalara internette 360 derece bakış açısı sunarak, etkin şekilde itibar yönetimi yapma olanağı sağlamak.

Manuel olarak yapılması gittikçe imkansızlaşan bir fonksiyonu otomatik olarak gerçekleştirmek.

Anlık, günlük, haftalık, aylık raporlarla, şirketin ürün geliştirme, fiyatlandırma, müşteri ilişkileri politikalarını doğru şekilde belirlemeleri için gerekli aracı sunmak.

Markanın ürün ve hizmetleriyle ilgili fikir ve görüşleri dinleyerek, krizlere acil ve anında müdahale etmek, kriz nedeniyle oluşacak maliyetlerin artmasını önlemek.

Yalnızca marka ve ürün adıyla değil, sektöre ilişkin anahtar sözcüklerle müşteri beklentilerini görme olanağı sağlamak.

Şirketlerin yalnızca kendi markalarını değil, rakip şirketlere ilişkin içerikleri de takip ederek, rekabette avantajlı duruma geçmelerini sağlamak.

Merkezi olmayan web platformlarını tarayıp, tek arayüzden, hedeflenen tüm içerik kaynaklarını takip etmek.

Müşteri ilişkileri ve itibar yönetimi maliyetlerini düşürerek kurumlara ekonomik fayda sağlamak

DA, kişilerin olaylar, hizmetler, ürünler, kurumlar, başka kişiler v.s. hakkındaki duygu ve düşüncelerini tespite yarar. Olumlu geri bildirimler firmayı teşvik ederken, olumsuz geri bildirimler caydırıcı görevler üstlenirler. Bazen firmalar kendileri ile diğer firmaları karşılaştırmak istediklerinde de DA'dan yardım alırlar. Bunun dışında pazarlama yöneticileri, politikacılar, online ürün yöneticileri, reklamcılar, girişimciler ve tüketici yorumlarına ihtiyaç duyan herkes için önemli bir alandır 0.

Duygu Analizi İle İlgili Temel Bilgiler

DA pek çok farklı alandan da destek almaktadır. Doğal dil işleme, metin analizi, veri madenciliği, yapay zeka bunlardan sayılabilir. Konu hakkında özellikle son yıllarda çalışma sayısı çok artmıştır. Ancak Türkçe yapılan çalışmalar çok fazla değildir.

DA farklı seviyelerde ele alınabilir 0:

Basit seviye: Eldeki veriyi bütün olarak sadece pozitif veya negatif şeklinde sınıflandırma yapar.

Karmaşık seviye: Veriyi sadece pozitif negative sınıflandırma değil derecelendirme yapar, örneğin 1-5 arası bir pozitif/negatif derecesi atar.

İleri seviye: Hedefi, kaynağı ve karmaşık tutumları inceler ve analiz eder. Örneğin cep telefonu inceleniyorsa cep telefonunun hangi özellikleri göz önüne alınacak (ekran, fiyat, performans, işlemci vs) ve bunların ayrı ayrı olumlu ve olumsuz yorumları nasıl çıkarılacak gibi konularla ilgilenir.

DA kullandığı veriyi ele alırken de farklılık gözetebilir0:

Belge Seviyesinde (Document-Level): Bir belge üzerinde çalışılır. En basit seviye karşılaştırmadır.Bir belge bir düşünceyi karşılar.

Cümle Seviyesinde (Sentence-Level): Bir belge bir çok duyguyu barındırabilir. Cümleler ayrılmalıdır. Genel yaklaşım bir önceki cümleyi baz almaktır. Son zamanlarda, cümlenin özelliğine göre yaklaşımlar vardır. (Şart cümleleri, soru cümleleri, esprili cümleler vb.)

Özellik Temelli (Aspect-based): Diğerlerine göre yaklaşım farklıdır. Konuşulan nesnenin özellikleri üzerine çalışır.

Duygu Analizinde Kullanılan Yöntemler

DA için makine öğrenme yöntemlerinden faydalanılmaktadır. Makine öğrenmesi kısaca insana ait olan bir özelliğin bilgisayara yazılım sayesinde kazandırılması çalışmasıdır ve yapay zekânın önemli bir alt konusudur0. Makine öğrenmesi üç ana başlıkta sınıflandırılmaktadır 0 :

Denetimli Öğrenme (*Supervized Learning*): En basit tanımla denetimli öğrenme, sınıflandırma (*classification*) işlemi ile yapılır. Denetimli öğrenme, türü bilinen yani etiketli örneklerden, türü bilinmeyen bir örneğin sınıfını yani etiketini bulmak için kullanılan bir yöntemdir. Sınıflandırma işlemi için eğitim ölçüt kümesine ihtiyaç vardır. Sınıflandırma işleminde verilen bir örnek için eğitim ölçüt kümesi temel alınarak ait olacağı sınıfı bulma yöntemine dayalıdır. Denetimli öğrenme için etiketli veri bulmak zor ve pahalı iştir, çünkü verilerin etiketlenebilmesi için insan desteğine ihtiyaç vardır.

Denetimsiz Öğrenme (*Unsupervized Learning*): Denetimsiz öğrenme kümeleme (*clustring*) işlemi ile yapılır. Denetimsiz öğrenmede örneklerin etiketleri belli değildir. Kümeleme işlemi birbirine benzeyen örneklerin aynı küme içerisine alınması mantığına dayanır. Denetimsiz öğrenme için etiketsiz veri bulmak kolay ve ucuzdur .

Yarı Denetimli Öğrenme (*Semi Supervised Learning, SSL*): Denetimli ve denetimsiz öğrenme sistemlerinin kabaca birleşimi de diyebileceğimiz yarı denetimli öğrenme işlemi aslında bir sınıflandırma işlemidir. Elimizde bulunan az miktardaki etiketli veri ile aynı ortamdaki benzer özelliklere sahip etiketsiz verileri etiketleme işlemidir.

Yöntem

DA için programda aşağıda verilen yöntem izlenmiştir:

Algorİtma:

Algoritma basamakları şunlardır:

Verileri topla.

Verilerde olumlu ve olumsuz yorumlarda kullanılan önemli ve belirleyici kelime ve kelime gruplarını çıkar.

Verilerin düzeltilmesini ve denetlenmesini sağla.

Verileri metin haline getir.

Verileri morfolojik analizden geçir.

Morfolojik analizden birden fazla çıkan sonuçları teke indir.

Metinleri işleyerek bunlardan kullanılması planlanan özellikleri çıkar.

Değerlendirme yöntemlerini belirle

Naïve Bayes yöntemini yazıp belirlenen özelliklerle dene.

Weka programında Naïve Bayes yöntemi kullanılması için yukarıdaki özelliklerin csv dosyalarını hazırla ve Naïve Bayes'i test et.

Yeni bir yöntemle sadece olumlu veya olumsuz yorumlarda önemli olan kelime veya kelime gruplarından çıkarılan sözlük yardımıyla bir sınıflandırma yap.

Programdaki ve Weka'daki sonuçları karşılaştır.

Gerçekleme:

Algoritmamızda bahsedilen aşamaların detayları aşağıda belirtilmiştir.

Verilerin toplanması

Veriler elle toplandı. Veriler elde edilirken www.sinemalar.com ve www.beyazperde.com sitelerinden yararlanıldı. Bu siteler Türkiye'deki popüler siteler olduğu ve yeterli sayıda yorum içerdiği için tercih edildi. Olumlu veriler için IMDb(Internet Movie Database) (www.imdb.com) puanı yüksek filmlerden yararlanılırken, olumsuz veriler düşük puanlı filmlerden elde edildi. Eğitim verisi olarak 919 tane olumlu, 615 tane olumsuz yorum toplandı. Test verisi için 25 olumlu 25 olumsuz olmak üzere toplam 50 yorum toplandı.

Tablo 1'de veri bilgileri gösterilmiştir. Tablo 2'de bu verilerin ilk toplandığı hali excel satırı olarak verilmiştir. Filmin adı ile birlikte yapılan yorum ve bu yorumda bazı yazım yanlışlarının düzeltilmiş hali ve yorumda geçen ve olumlu veya olumsuz olmaya etki eden kelimeler ile yorumun sınıfı bilgisi tutulmuştur.

Tablo 1:Eğitim ve test verisi bilgileri

Yorum Tipi	Toplam	Eğitim	Test
Olumlu	919	894	25
Olumsuz	615	590	25

Tablo 2:Olumlu ve olumsuz eğitim verisinden örnekler.

Filmin Adı	Yorum	Yorum düzeltilmiş	Yorumda önemli yerler	Olumlu(1) / Olumsuz(2)
Inception(Başlangıç)	Geleceğin kült filmlerinden biridir. Muhakkak izlemek lazım, zihinsel arşivde bulundurmak lazım. Oyunculuklar, senaryo harikaydı. İlla ki eksik bir şey söylemek gerekirse son sahnelerin fazla uzatıldığını söyleyebilirim, zaten 148 dakika bir film bence (istisnai durumlar dışında) çoktur. Başarılı bir film. Mutlaka izleyin derim...	Geleceğin kült filmlerinden biridir. Muhakkak izlemek lazım, zihinsel arşivde bulundurmak lazım. Oyunculuklar, senaryo harikaydı. İlla ki eksik bir şey söylemek gerekirse son sahnelerin fazla uzatıldığını söyleyebilirim, zaten 148 dakika bir film bence (istisnai durumlar dışında) çoktur. Başarılı bir film. Mutlaka izleyin derim...	kült, muhakkak izlemek lazım, harika, başarılı, mutlaka	1
Lucy	Cidden çok değişik bir film. Ve bence biraz da saçma olmuş.	Cidden çok değişik bir film. Ve bence biraz da saçma olmuş.	saçma, değişik	2

Ön İşlemler

Verilerde olumlu ve olumsuz yorumlarda kullanılan önemli ve belirleyici kelime ve kelime grupları iki işaretleyici tarafından çıkarıldı. Daha sonra bunların çapraz kontrolü yapıldı. Bazı yazım yanlışları da düzeltildi, ancak veri sayısı çok fazla olduğu ve sosyal medya metinleri fazlaca yazım yanlışı içerdiği için tamamen giderilemedi. Veri içinde geçen bazı karakterler ve gereksiz bilgilerin bir kısmı da atıldı.

Morfolojik Analiz

Veriler metin haline getirildi. Kelime kökleri kullanılacağı için morfolojik analiz yapmak gerekti. Bunun için özel bir yazılım kullanıldı. Veriler morfolojik analiz (referans ver indirme linki yaz) için uygun formata C programıyla geçirildi.

Tablo 3: Morfolojik analiz girdi formatı, çıktısı ve belirsizliklerin giderilmiş hali

Morfolojik analiz girdisi	**Morfolojik analiz çıktısı**	**Belirsizliklerin giderilmiş hali**
<DOC> <TITLE> <S> xwqLucyxwq1xwq2 </S> </TITLE> <S> Cidden çok değişik	<DOC> <DOC> <TITLE> <TITLE> <S> <S> xwqLucyxwq1xwq2 *UNKNOWN* </S> </S> </TITLE> </TITLE> <S> <S> Cidden cidden +Adverb çok çok +Det çok çok +Adverb çok çok +Adj çok çok +Postp+PCAbl	<DOC> <DOC> <TITLE> <TITLE> <S> <S> xwqLucyxwq1xwq2 *UNKNOWN* </S> </S> </TITLE> </TITLE> <S> <S> Cidden cidden+Adverb çok çok+Adverb değişik değişik+Adj bir bir+Det

bir	değişik değişik +Adj	film film+Noun+A3sg+Pnon+Nom
film	bir bir +Det	. .+Punc
.	bir bir +Adverb	</S> </S>
</S>	bir bir +Adj	<S> <S>
	bir bir +Num+Card	Ve ve+Conj
<S>	film film +Noun+A3sg+Pnon+Nom	bence ben+Pron+Pers+A1sg+Pnon+Equ
	. . +Punc	biraz biraz+Adj
Ve	</S> </S>	da da+Conj
bence	<S> <S>	saçma saçma+Noun+A3sg+Pnon+Nom
biraz	Ve ve +Conj	olmuş ol+Verb+Pos+Narr+A3sg
da	bence ben +Noun+A3sg+Pnon+Equ	</S> </S>
saçma	bence ben +Pron+Pers+A1sg+Pnon+Equ	</DOC> </DOC>
olmuş	biraz biraz +Adverb	
</S>	biraz biraz +Adj	
	da da +Conj	
</DOC>	saçma saçma +Noun+A3sg+Pnon+Nom	
	saçma saçma +Adj	
	saçma saç +Verb+Neg+Imp+A2sg	
	saçma saç +Verb+Pos^DB+Noun+Inf2+A3sg+Pnon+Nom	
	olmuş ol +Verb+Pos+Narr+A3sg	
	olmuş ol +Verb+Pos+Narr^DB+Adj+Zero	
	</S> </S>	
	</DOC> </DOC>	

Morfolojik analiz için Deniz Yüret'in bloğundan ulaşılabilecek, Kemal Oflazer'in Türkçe için hazırlanmış sonlu durum makinesini içeren MORPHOLOGYCAL TAGGER uygulaması kullanıldı. Uygulamaya http://deniz.yuret.com/turkish/tr-tagger.tgz adresinden erişilebilir. TAGGER'i çalıştırmadan önce makinede Xerox Finite State yazılımı bulunmalıdır. Bu yazılımı http://www.stanford.edu/~laurik/.book2software/ adresinden indirmek mümkündür. Yazılımı kurduktan sonra metnin belirli bir formatta hazırlanması gerekir. Bu format için hazırlanmış örnek input TAGGER içinde sample-input.txt olarak mevcuttur. Örnek bir morfolojik analiz girdisi ve çıktısı formatı olumsuz verinin ilk satırı için Tablo 3'te il iki sütunda verilmiştir.

Belirsizliklerin Giderilmesi

Morfolojik analizden birden fazla sonuç da çıkabildiği için bunları elemek gerekti. Bunun için de bir programdan faydalanıldı. Türkçe bir kelimeden eklenen eklerle bir çok farklı kelime türeyebildiği için TAGGER olası ihtimalleri oluşturur. Deniz Yüret'in MORPHOLOGYCAL DISAMBIGUATORı bulunan ihtimallerden belirsizliği gidererek sonucu bulmak için kullanılan bir uygulamadır. http://deniz.yuret.com/turkish/tr-disamb.tgz adresinden indirilebilir.TAGGER'ın yukarıdaki çıktısı DISAMBIGUATOR'a girdi olarak verilir. Bunun çalışması için perl programının kurulması gereklidir. Çıktı olarak üretilen satır Tablo 3 son sütunda verilmiştir.

Özelliklerin Çıkarılması

Metinler işlenerek bunlardan kullanılması planlanan özellikler çıkarıldı. Kullanılan özellikler aşağıda anlatılmıştır:

Tüm kelimelerin tekl kelime kökü ve pozitif/negatif oluşu:

Öncelikle kelimeler teker teker ve birbirinden bağımsız olarak kökleri ve olumluluk durumu ile kullanıldı. Burada olumluluk demekle kelimenin asıl anlamında kullanılması durumunda pos eki bir olumsuz ek alması durumunda da neg eki eklenmesidir. Örneğin *severim* kelime kökü sev ve asıl anlamında olduğu için buradan *sevPos* özelliği çıkacaktır. Ama kelime *sevmem* olsaydı özellik olarak *sevNeg* çıkacaktı.

Tüm kelimelerintekli kelime kökü, pozitif/negatif oluşu ve kipi:

Burada da kelimeler teker teker ve birbirinden bağımsız olarak kökleri, olumluluk durumu ile kip eklerinden bazıları kullanıldı. Kip olarak emir, gereklilik, istek kiplerine bakıldı. Örneğin *sev* kelime kökü sev ve asıl anlamında olduğu ve emir kipi olduğu için buradan *sevPosİmp* özelliği çıkacaktır. Ama kelime *sevmemeliyim* olsaydı özellik olarak *sevNegNeces* çıkacaktı.

Tüm kelimelerin İkili kelime kökü ve pozitif/negatif oluşu:

Birinci özelliğe benzemektedir. Tek farkı teker teker yerine ikili kelime gruplarına bakmasıdır. Örneğin *bu filmi sevdim* cümlesinden *buPosFilmPos* ve *filmPosSevPos* özellikleri çıkacaktır.

Tüm kelimelerin İkili kelime kökü, pozitif/negatif oluşu ve kipi:

İkinci özelliğe benzemektedir. Tek farkı teker teker yerine ikili kelime gruplarına bakmasıdır. Örneğin *bu filmi seyredin* cümlesinden *buPosNokipFilmPosNokip* ve *filmPosNokipSeyretPosİmp* özellikleri çıkacaktır. Ama kelime *sevmem* olsaydı özellik olarak *sevNeg* çıkacaktı.

Önemli kelimelerin tekli kelime kökü ve pozitif/negatif oluşu:

Birinci özelliğin sadece önemli kelimeler için kullanılması ile elde edilmiştir.

Önemli kelimelerintekli kelime kökü, pozitif/negatif oluşu ve kipi:

İkinci özelliğin sadece önemli kelimeler için kullanılması ile elde edilmiştir

Önemli kelimelerin ikili kelime kökü ve pozitif/negatif oluşu:

Üçüncü özelliğin sadece önemli kelimeler için kullanılması ile elde edilmiştir.

Önemli kelimelerin ikili kelime kökü, pozitif/negatif oluşu ve kipi:

Dördüncü özelliğin sadece önemli kelimeler için kullanılması ile elde edilmiştir.

Değerlendirme yöntemlerini belirlenmesi

Program başarısını ölçmek için sıkça kullanılan değerlendirme yöntemlerinden doğruluk (Acc-accuracy), netlik (P-Precision), ve kapsama (R-Recall) ile P ve R için harmonik ortalama olan F ölçümü (FM-FMeasure) kullanıldı. Ayrıca tahminlerin sınıflara göre dağılımını gösteren hata matrisi (CM-Confusion matrix) bulundu.

True Positive (TP -Doğru Pozitif): Olumlu tahmin ettiginiz gerçekte olumlu yorumlar.

True Negative (TN - Doğru Negatif): Olumsuz tahmin ettiginiz gerçekte olumsuz yorumlar.

False Positive (FP -Yanlış Olumlu): Olumlu tahmin ettiginiz gerçekte olumsuz yorumlar.

False Negative (FN - Yanlış Olumsuz): Olumsuz tahmin ettiginiz gerçekte olumlu yorumlar.

Bunları örnekle anlatacak olursak 50 tane olumlu 50 tane de olumsuz verimiz olsun. Sınıflama yapan bir sistemle biz bunlara bir değer atayalım. 50 olumlunun 40 tanesi olumlu, 10 tanesi olumsuz ve 50 olumsuzun 30 tanesi olumsuz 20 tanesi de olumlu olarak bulunmuş olsun. Bu durumda aşağıdaki gibi bir CM oluşacaktır:

Tablo 4:Hata Matrisi (Confusion Matrix)

Asıl	**Atanan**	
	Olumlu	**Olumsuz**
Olumlu	40 (TP)	10 (FP)
Olumsuz	20 (FN)	30 (TN)

Toplam eleman sayısı N=100'dür. A değeri hesaplanırken doğru sınıflanan tüm örneklerin tüm veriye oranına bakılır. P, R, ve FM değerleri her sınıf için ayrı hesaplanır. Pc c sınıfına aitken c sınıfı olarak bulunan verilerin tüm c sınıfı olarak bulunan verilere oranıdır, yani Pc c için bulunan değerlerin doğruluk oranıdır. Rc c sınıfına aitken c sınıfı olarak bulunan verilerin tüm c sınıfına ait verilere oranıdır, yani Rc c sınıfına ait verilerin ne kadarının kapsanabildiğini veya c sınıfı olarak bulunabildiğini gösterir. FM ise P ve R harmonic ortalamasıdır ve sistemlerin P yüksekken R değeri düşük veya R yüksekken P değeri düşük sonuçlar vermesinin önüne geçmek için kullanılır ve FM=2PR/(P+R) olarak hesaplanır.

Tablo 5: Olumlu ve Olumsuz P,R hesaplamaları

	P	**R**
Olumlu	P_Ol = $\frac{40}{40+20}$	R_Ol = $\frac{40}{40+10}$
Olumsuz	P_Olz = $\frac{30}{30+10}$	R_Olz = $\frac{30}{30+10}$

FM_Ol=$\frac{2 \text{ x } P_{Ol} \text{ x } R\_Ol}{P\_Ol+R\_Ol} = \frac{2x0.67x0.8}{0.67+0.8} = 0.73$ Denklem 1

FM_Olz=$\frac{2 \text{ x } P_{Olz} \text{ x } R\_Olz}{P\_Olz+R\_Olz} = \frac{2x0.75x0.6}{0.75+0.6} = 0.67$ Denklem 2

Acc = $\frac{TP+TN}{TP+FP+TN+FN} = \frac{40+30}{40+10+30+20} = \frac{70}{100} = 0.7$ Denklem 3

Bu makalenin devamında P, R, FM ve Acc değerleri yüzde olarak gösterilecektir.

Naïve Bayes Yöntemi

Makine öğrenme yöntemlerinden denetimli öğrenme metodlarından Naïve Bayes metodu uygulandı. Bu metodun eğitim setini hazırlamak pahalı ve zor olmasına rağmen bu metodu seçme nedenimiz Naïve Bayes ile test etmenin kolaylığı ve yüksek başarısıdır. Ayrıca sinema yorumları toplamak biraz zor olsa da sınıflandırma işlemi çok zor olmamıştır.

Naïve Bayes (NB) sınıflandırıcısı, Bayes teoremine dayanan basit ve güçlü bağımsızlık varsayımı kullanan bir olasılık sınıflandırma yöntemidir. E-posta sınıflandırmasından, belge sınıflamasına, duygu düşünce analizine kadar pek çok alanda başarıyla kullanılan temel bir metin sınıflandırma tekniğidir. Sadelik ve fazlaca basit varsayımlarına rağmen NB karmaşık pek çok gerçek hayat probleminde başarıyla uygulanmaktadır. Bu yöntemden daha başarılı teknikler kullanılsa da NB işlemci ve bellek kullanımı açısından daha verimlidir. Ayrıca daha az sayıda eğitim verisine ve diğer alternatif yöntemlerden daha kısa eğitim zamanına ihtiyaç duyar 0.

Naïve Bayes Sınıflandırıcısı Ne Zaman Kullanılır?

NB sınırlı işlemci ve bellek kaynağına sahip olunan durumlarda uygundur. Ayrıca eğitim zamanının kısalığının önemli olduğu uygulamalarda eğitim çok hızlı olduğundan en uygun seçimlerden biridir. Pek çok uygulamada alternatifleri test etmek ve karşılaştırma yapmak için bir yöntem olarak da kullanılmaktadır.

Naïve Bayes Teoremi

NB sınıflandırıcısı sınıflandırmada kullanılan özelliklerin birbirinden bağımsız olduğunu varsayar. Her ne kadar bu varsayım pek çok durumda yanlış olsa da bu sınıflandırıcının makul gibi görünmeyen etkinliğinin arkasında yatan bazı teorik sebepler olduğu gösterilmiştir 0. Olasılık hesapları düşük kaliteli olsa bile sınıflandırma kararları oldukça iyidir 0. Seçilen sınıf hakkındaki olasılık hesabı çok gerçekçi olmasa da, asıl amaç gerçek olasılıkları hesaplamak değil göreceli olarak sınıfları birbiriyle karşılaştırıp karar vermek olduğu için genelde karar doğru olmakta ve model iyi çalışmaktadır0.

Metin sınıflandırma uygulamasında genelde kelimeler, kelime grupları veya bunların bazı nitelikleri özellik olarak kullanılmaktadır. Maximum a posteriori (MAP) karar kuralı ile aşağıdaki gibi bir sınıflandırıcı elde edilmektedir:

$$c_{map} = \arg\max_{c \in C}\left(P(c \mid d)\right) = \arg\max_{c \in C}\left(P(c)\prod_{1 \le k \le n_d} P(t_k \mid c)\right)$$ Denklem 4

Burada t_k belgenin terim/kelimeleri, C sınıflandırmada kullanılan olası sınıfların kümesi, $P(c \mid d)$ d belgesi verildiğinde c sınıfı olma koşullu olasılığı, $P(c)$ c sınıfının ön olasılığı (prior) ve $P(t_k \mid c)$ ise c sınıfı verildiğinde t_k terimi/kelimesi olma koşullu olasılığıdır. Bu formüle göre bir belgenin sınıfını bulmak için o sınıf verildiğinde her bir terim/kelimenin olma olasılıklarının çarpımını (likelihood-olabilirlik) hesaplamak ve bunu o sınıfın ön olasılığı (prior) ile çarpmak gerekmektedir. Bütün sınıflar için bu işlem yapılarak çıkan sonuçlar içinden en büyük olanı seçilmektedir.

Bilgisayarların sayıları belli bir netliğe kadar ifade edebilme kapasitesi ve bazı hesap hatalarını göz önüne alınca bu çarpımdaki sonuçlar bazı problemlere neden olabilir. Sonuçlarda bazı değerler çok küçük olup bellekte tam gösterilemediği ve sığmadığı için sıfıra yuvarlanabilir, analizin sonuçlarını geçersiz kılabilir. Bunu önlemek için olasılıkların çarpımını maksimum yapmak yerine bunların logaritmasının toplamını maksimum yapmak yöntemi tercih edilir ve formül aşağıdaki şekliyle kullanılır:

$$c_{map} = \arg\max_{c \in C} \left(\log P(c) + \sum_{1 \le k \le n_d} \log P(t_k \mid c) \right)$$

Denklem 5

Yeni formülde maksimum olasılıklı sınıf yerine maksimum log skorlu sınıf seçilir. Logaritma fonksiyonu monoton artan bir fonksiyon olduğu için MAP karar kuralı aynı kalır. Burada dikkat edilmesi gereken bir konu da belirli bir sınıfta bazı terimlerin/kelimelerinhiç geçmemesi durumudur. Bu durumda koşullu olasılık değeri 0 olur. Eğer ilk kural kullanılırsa çarpım 0 olur, ikinci kural kullanılırsa logaritma 0 tanımsız olur. Bunu önlemek için bir ekleme veya Laplace smoothing yöntemi ile sayma sonuçlarına bir eklenir. Formül aşağıdaki şekilde değiştirilir:

$$P\ (t \mid c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'}) + B'}$$

Denklem 6

Burada B' terim/kelime dağarcığı V içindeki eleman sayısıdır 0.

Naïve Bayes Çeşitleri Ve Kullanım Alanları

NB yönteminin bazı çeşitleri bulunmaktadır. Multinomial NB(MNB), Binarized Multinomial NB (BMNB) ve Bernoulli NB (BNB) bunlar arasında sayılabilir. Bunlar farklı sonuçlar verebilir çünkü farklı modeller kullanmaktadır. Genelde MNB sınıflama probleminde birden fazla geçişlerin önemli olduğu durumlarda kullanılır. Örnek olarak konu sınıflama (Topic Classification) verilebilir. BMNB ise frekansların sınıflandırmada çok önemli bir rol oynamadığı durumlar için uygundur. Duygu analizi BNMB kullanımı için iyi bir uygulama alanıdır, çünkü bu metinlerde "kötü" gibi bir olumsuz kelimenin kaç defa geçtiği değil geçip geçmediği önem kazanmaktadır. Son olarak BNB yöntemi ise

bazı terim/kelimeleringeçmeme durumunun önemli olduğu istenmeyen e-posta(spam e-mail) veya istenmeyen içerik tespiti gibi uygulamalar için faydalı olmaktadır0.

Multinomial Naïve Bayes Modeli

Bu yöntemde t (terim/kelime) için koşullu olasılık değeri c sınıfına ait belgelerdeki t'nin göreceli frekans değerine göre aşağıdaki gibi hesaplanmaktadır0:

$$P\ (t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$ Denklem 7

t için c sınıfındaki belgelerin eğitim verisinde birden çok geçişi de göz önüne alınmaktadır. Bu yöntemin eğitim ve test algoritması aşağıda verilmiştir:

Tablo 6:MNB Eğitim ve Test Algoritmaları

MNBEğitim(C,D)

V←DağarcığıÇıkar(D)

N←BelgeleriSay(D)

for each cεC

do Nc←BelgeSay(D,c)

prior[c]←Nc/N

$text_c$←SınıfBelgeleriniBirleştir(D,c)

for each tεV

do Tct←TerimSay($text_c$,t)

for each tεV

do condprob[t][c]←(Tct+1)/$\sum_{t'}$(Tct'+1)

return V, prior, condprob

BNBTest(C,V,prior,condprob,d)

W←BelgedenTCıkar(V,d)

for each cεC

do skore[c]←log prior[c]

for each tεW

do skor[c]+=log condprob[t][c]

return $\text{argmax}_{c\varepsilon C}$ skor[c]

Binarized (Boolean) Multinomial Naïve Bayes Modeli

Bu yöntem MNB çok benzemekle birlikte t terim/kelimesinin bütün geçiş sayısı/frekansı yerine sadece geçip geçmediğine bakmaktadır. Buradaki varsayım t için geçiş sayısı yerine geçip geçmediğinin modelin başarısına etki ettiğidir0. MNB'deki eğitim ve test algoritması aynı kalmakta sadece frekans yerine t varsa 1 yoksa 0 kullanılmaktadır.

Bernoulli Naïve Bayes Modeli

BNB, dağarcıktaki her t için belgede varsa 1 yoksa 0 değeri üretir. T için geçiş sayılarına bakmadığı ve geçmeyen t değerlerini de göz önüne alması yönüyle MNB'den epeyce farklıdır. MNB geçmeyen t değerlerine bakmamaktadır 0. BNB uzun belgeleri sınıflarken geçiş sayılarına bakmadığı ve ilgisiz terimlerin geçip geçmediğine de duyarlı olduğu için çok hata yapabilmektedir. BNB için eğitim ve test algoritması aşağıda verilmiştir:

Tablo 7:BNB Eğitim ve Test Algoritmaları

BNBEğitim(C,D)

V←DağarcığıÇıkar(D)

N←BelgeleriSay(D)

for each cεC

do Nc←BelgeSay(D,c)

prior[c]←Nc/N

for each tεV

do Nct←BelgeSayTerim(D,c,t)

condprob[t][c]←(Nct+1)/(Nc+2)

return V, prior, condprob

BNBTest(C,V,prior,condprob,d)

Vd←BelgedenTCıkar(V,d)

for each cεC

do skore[c]←log prior[c]

for each tεV

do if tεVd

then skor[c]+=log condprob[t][c]

else skor[c]+=log(1-condprob[t][c])

return $\text{argmax}_{c\varepsilon C}$ skor[c]

Weka

Weka, makine öğrenimi amacıyla Waikato Üniversitesinde geliştirilmiş yazılımın ismidir. Günümüzde yaygın kullanımı olan çoğu makine öğrenimi algoritmalarını ve metotlarını içermektedir 00. Java dilinde geliştirilmiş olması ve kütüphanelerinin jar dosyaları halinde geliyor olması sayesinde, JAVA dilinde yazılan projelere kolayce entegre edilebilmesi kullanımını daha da yaygınlaştırmıştır .Weka, tamamen modüler bir tasarıma sahip olup, içerdiği özelliklerle veri kümeleri üzerinde görselleştirme, veri analizi, iş zekası uygulamaları, veri madenciliği gibi işlemler yapabilmektedir. Weka yazılımı, kendisine özgü olarak bir .arff uzantısı desteği ile gelmektedir. Ancak Weka yazılımının içerisinde CSV dosyalarını da ARFF formatına çevirmeye yarayan araçlar mevcuttur. 0

WEKA içerisinde yüklü olan sınıflandırma algoritmalarından herhangi birisini kullanarak mevcut veri kümesi üzerinde sınıflandırma yapılabilir. Ayrıca test ve sağlama (validation) için ayrı kümeler kullanmak da mümkündür.

WEKA Naïve Bayes Sonuçları

Weka ile yapılan deneylerin NB sonuçları **Hata! Başvuru kaynağı bulunamadı.**'te verilmiştir.

Tablo 8 Weka NB sonuçları

Ölçülen Değerler	**Asıllar**	**Atananlar**							
	Referanslar	**Özellik0**	**Özellik1**	**Özellik2**	**Özellik3**	**Özellik4**	**Özellik5**	**Özellik6**	**Özellik7**
TP	25	24	24	24	24	24	23	25	25
FN	0	2	2	2	2	3	3	6	6
FP	0	1	1	1	1	1	2	0	0
TN	25	23	23	23	23	22	22	19	19
Acc	100	94	94	94	94	92	90	88	88

Naïve Bayes Program Sonuçları

Programımızın NB sonuçları MBNB yöntemi ile Tablo 9'de verilmiştir. Bu yöntemde sadece özellik geçtiyse 1 geçmediyse 0 olarak alınmıştır.

Tablo 9:Programın NB sonuçları(Binary için)

Ölçülen Değerler	**Asıllar**	**Atananlar**							
	Referans	**Özellik 0**	**Özellik 1**	**Özellik 2**	**Özellik 3**	**Özellik 4**	**Özellik 5**	**Özellik 6**	**Özellik 7**
TP	25	24	24	23	23	23	23	24	24
FN	0	2	2	2	2	2	2	7	6
FP	0	1	1	2	2	2	2	1	1
TN	25	23	23	23	23	23	23	18	19
Acc	100	94	94	92	94	92	92	84	86

Programımızın NB sonuçları MNB yöntemi ile Tablo 10'te verilmiştir. Bu yöntemde özelliklerin geçiş sayıları göz önüne alınmıştır.

Tablo 10:Programın NB sonuçları (Binary için)

Ölçülen Değerler	Asıllar	Atananlar							
	Referans	Özellik 0	Özellik 1	Özellik 2	Özellik 3	Özellik 4	Özellik 5	Özellik 6	Özellik 7
TP	25	24	24	23	23	23	23	24	24
FN	0	2	2	2	2	3	3	6	8
FP	0	1	1	2	2	2	2	1	1
TN	25	23	23	23	23	22	22	19	19
Acc	100	94	94	92	92	90	90	86	86

Sözlük Yöntemi

Kendi geliştirdiğimiz yöntemle sadece olumlu veya olumsuz yorumlarda önemli olan kelime veya kelime gruplarından çıkardığımız sözlük yardımıyla bir sınıflandırma yaptık. Buna göre test yorumları için basit bir skor hesaplaması yaptık. Test yorumunun içindeki kelimeler pozitif kelimeler sözlüğünde geçiyorsa pozitif skora bu kelimenin olumlular sözlüğünde geçme oranı kadar ekledik. Aynı hesapmayı negative için de yapıp skorların en yükseğini sonuç olarak atadık. Bunun sonuçları da Naïve Bayes ile karşılaştırıldı. Programımızın yeni yöntemle sonuçları Tablo 11'da verilmiştir.

Tablo 11:Programın yeni yöntem ile sonuçları

Ölçülen Değerler	Asıllar	Atananlar							
	Referans	Özellik 0	Özellik 1	Özellik 2	Özellik 3	Özellik 4	Özellik 5	Özellik 6	Özellik 7
TP	25	25	25	24	24	23	23	22	22
FN	0	0	0	1	1	2	2	3	3
FP	0	18	18	8	8	6	5	5	5
TN	25	7	7	17	17	19	20	20	20
Acc	100	64	64	82	82	84	86	84	84

Tüm Sonuçların karşılaştırılması

Tablo 12 yöntemlerin özelliklere göre doğruluk oranlarını, Tablo 13 yeni yöntemin özelliklere göre doğruluk oranlarını, Tablo 14 ve Tablo 15 olumlu ve olumsuz sınıfların MBNB ve MNB yöntemleriyle özelliklere göre P, R, FM değerlerini vermektedir. Şekil 5-Şekil 12 arasındaki grafikler bu değerleri grafiklerle görselleştirmektedir.

Doğruluk oranlarına göre NB yöntemleri yaklaşık aynı sonuçları vermektedir. En iyi sonuçlar NB yöntemleri için tekil tüm kelime kökleri ile elde edilmiş ve %94 doğruluk sağlanmıştır. İkili kelimeler kullanıldığında başarı oranı biraz düşmüştür. Sadece önemli kelimeler kullanıldığında ise düşüş daha fazla olmuş ve özellikle ikili kelimeler kullanıldığında doğruluk oranı %84-86 seviyelerine inmiştir. Tekil kelimelerinikililere göre daha başarılı olması üretilen özellik sayısının daha az olması, ikili kelimeleringörülme olsılıklarının daha düşük olması nedeniyle öğrenme aşamasında elde edilememiş olması ve gereksiz bazı özellikler çıkmasına neden olarak özellik sayısının büyüklüğünü artırması gibi nedenlere bağlanabilir. Ayrıca NB yöntemlerinden MNB ve MBNB yakın sonuçlar vermiştir. Bundaki en önemli etken, kullanılan verideki yorumların kısa olması ve frekans veya var/yok bilgisi kullanılmasının yaklaşık aynı sonucu vermesidir.

Yeni önerilen yöntemde ise tam NB yöntemlerine göre daha farklı bir sonuç ortaya çıkmıştır. Tekil kelime köklerinin tamamı kullanıldığında olumlu ve olumsuz yorumlarda geçen pek çok ortaklıktan dolayı doğruluk oranı %64 çıkmıştır. Ancak bu yöntemin amacı zaten tüm kelimeleri kullanmak değildir. Bu nedenle beklenildiği şekilde önceden oluşturulan olumlu ve olumsuz önemli kelimeler sözlüğü kullanılarak skor hesaplaması yapılınca doğruluk oranları %64 seviyesinden %84-86 seviyelerine çıkarak önemli bir iyileşme göstermiştir. Ayrıca tüm kelimelerde ikili gruplara bakıldığında bu ikililerin olumlu ve olumsuzda ortak çıkma olasılığı düştüğü için yine başarı açısından teklilere göre daha iyi sonuç vermiştir.

P, R ve FM değerlerine baktığımızda, NB için doğruluk oranlarındaki durumla karşılaşmaktayız. Tüm tekil kelimeler daha dengeli P ve R değerleri vermiştir. Önemli kelimeler kullanıldığında ve bunların tekli veya ikili olmasına göre P ve R değerleri sınıflar arasında daha dengesiz dağılmıştır.

Yeni yöntemde de P, R değerlerinin sınıflar arası dalgalanmaları özellikle tüm kelimeler kullanıldığında fazladır, önemli kelimeler kullanıldığında ise daha sağlıklı sonuçlar vermektedir. Skorlar arasındaki fark çok yakın çıkabildiği ve bu fark için bir eşik değeri belirlenmediği için bu sonuçlar doğal kabul edilebilir. Bu durumu düzeltmek için sonraki çalışmalarda, öncelikle olumlu ve olumsuz yorumlarda geçen ortak özelliklerin belli yöntemlerle elenmesi yapılıp, aradaki farkın da deneysel olarak en uygun seviyesi seçilerek karar verilebilir.

Tablo 12: Metotların özelliklere göre doğruluk oranları

Doğruluk								
Metot	**Özellik0**	**Özellik1**	**Özellik2**	**Özellik3**	**Özellik4**	**Özellik5**	**Özellik6**	**Özellik7**
NBBin	94	94	92	92	92	92	84	86
NBFR	94	94	92	92	90	90	86	86
WekaNB	94	94	94	94	92	90	88	88
Yyöntem	64	64	82	82	84	86	84	84

Tablo 13:Yeni yöntemin özelliklere göre doğruluk oranları

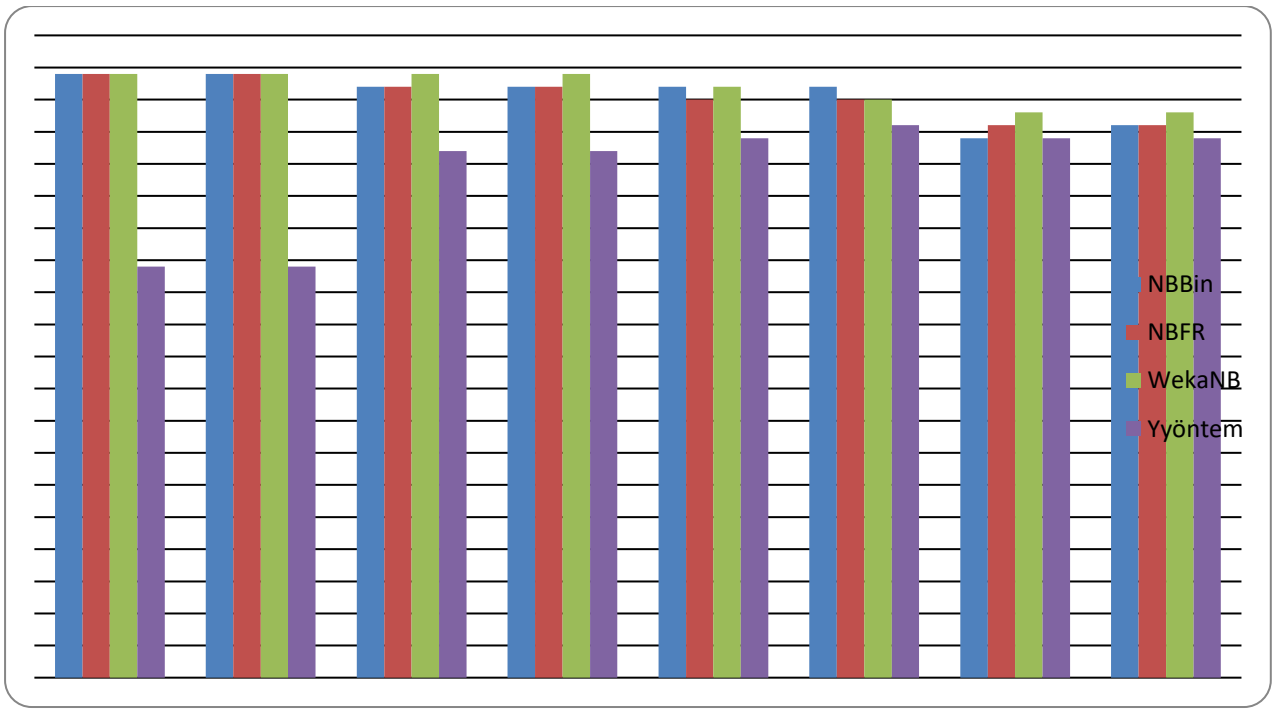
Yeni yöntem								
Ölçüm	**Özellik0**	**Özellik1**	**Özellik2**	**Özellik3**	**Özellik4**	**Özellik5**	**Özellik6**	**Özellik7**
P_Ol	58	58	75	75	79	82	81	81
P_Olz	100	100	94	94	90	90	87	87
R_Ol	100	100	96	96	92	92	88	88
R_Olz	28	28	68	68	76	80	80	80
FM_Ol	74	74	84	84	85	87	85	85
FM_Olz	44	44	79	79	83	85	83	83

Tablo 14:Olumlu ve olumsuz sınıfların MBNB ve özelliklere göre P, R, FM değerleri

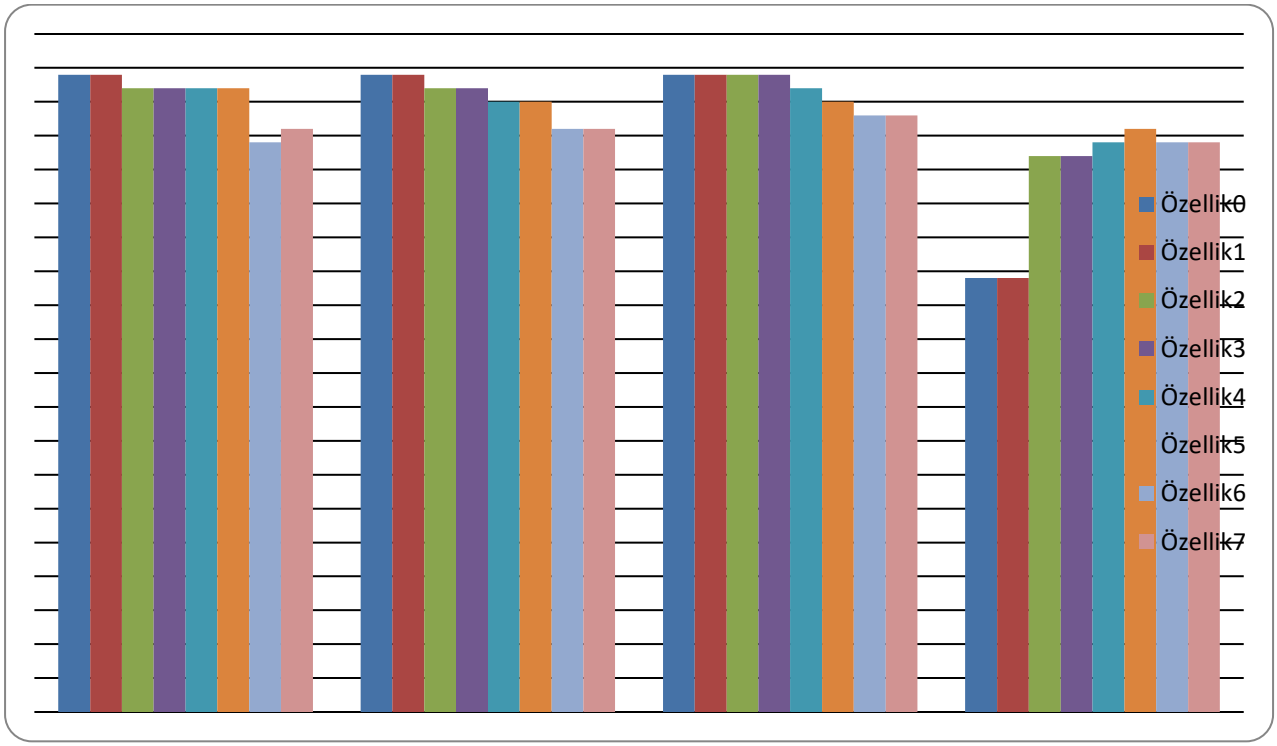
MBNB								
Ölçüm	**Özellik0**	**Özellik1**	**Özellik2**	**Özellik3**	**Özellik4**	**Özellik5**	**Özellik6**	**Özellik7**
P_Ol	92	92	92	92	92	92	77	80
P_Olz	96	96	92	92	92	92	95	95
R_Ol	96	96	92	92	92	92	96	96
R_Olz	92	92	92	92	92	92	72	76
FM_Ol	94	94	92	92	92	92	86	87
FM_Olz	94	94	92	92	92	92	82	84

Tablo 15:Olumlu ve olumsuz sınıfların MNB ve özelliklere göre P, R, FM değerleri

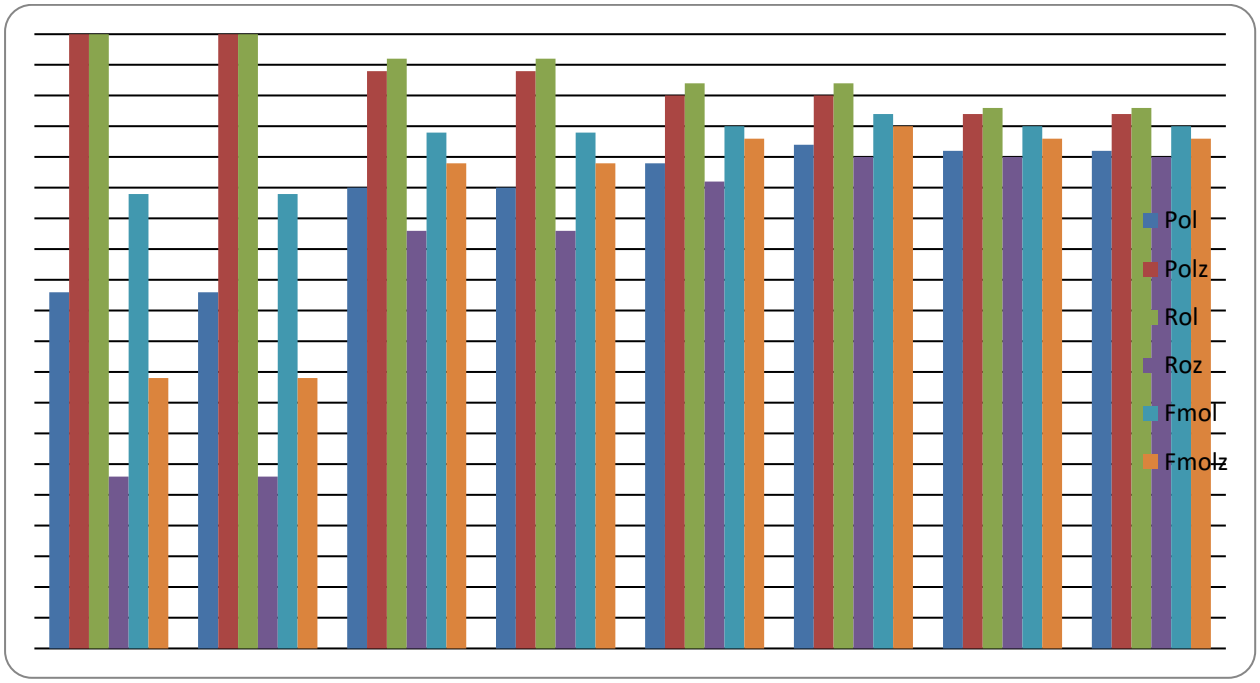
MNB								
Ölçüm	**Özellik0**	**Özellik1**	**Özellik2**	**Özellik3**	**Özellik4**	**Özellik5**	**Özellik6**	**Özellik7**
P_Ol	92	92	92	92	88	88	80	80
P_Olz	96	96	92	92	92	92	95	95
R_Ol	96	96	92	92	92	92	96	96
R_Olz	92	92	92	92	88	88	76	76
FM_Ol	94	94	92	92	90	90	87	87
FM_Olz	94	94	92	92	90	90	84	84



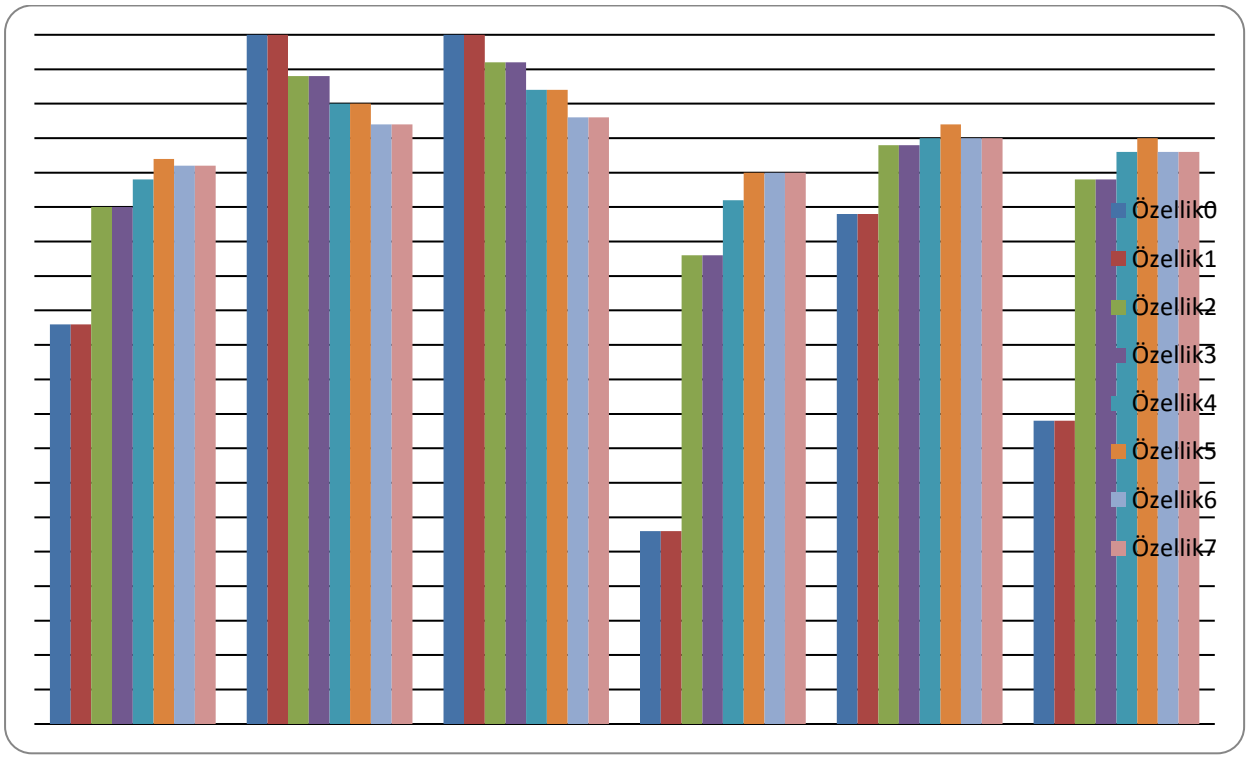
Şekil 5: Özelliklere göre yöntemlerin doğruluk oranları grafiği



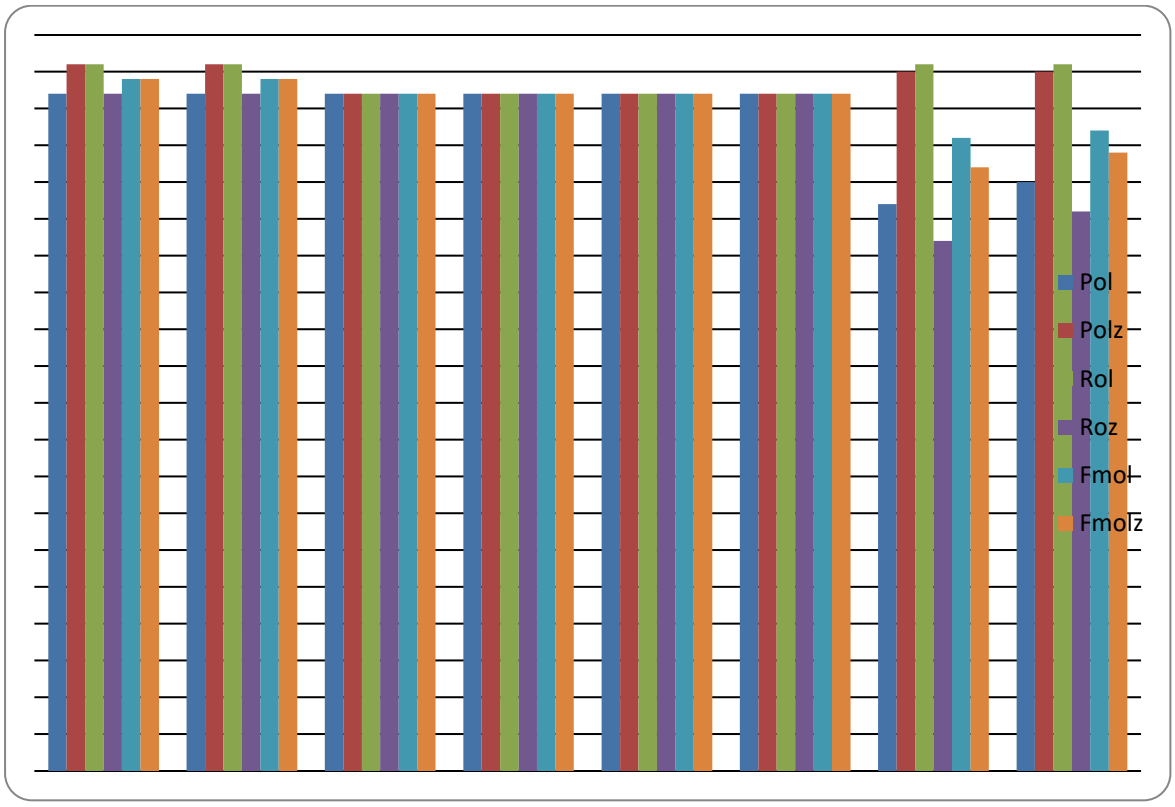
Şekil 6: Yöntemlere göre özelliklerin doğruluk oranları grafiği



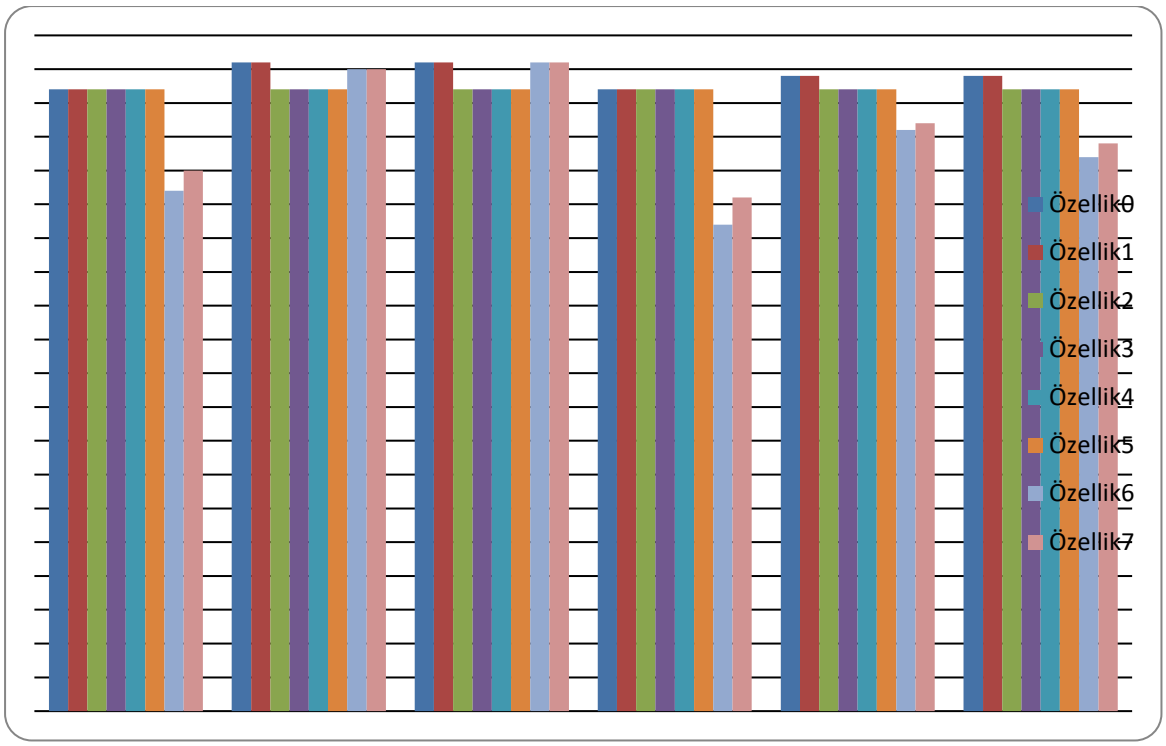
Şekil 7: Yeni Yöntemin Olumlu ve olumsuz sınıflar için her bir özellikteki P, R, FM değerleri



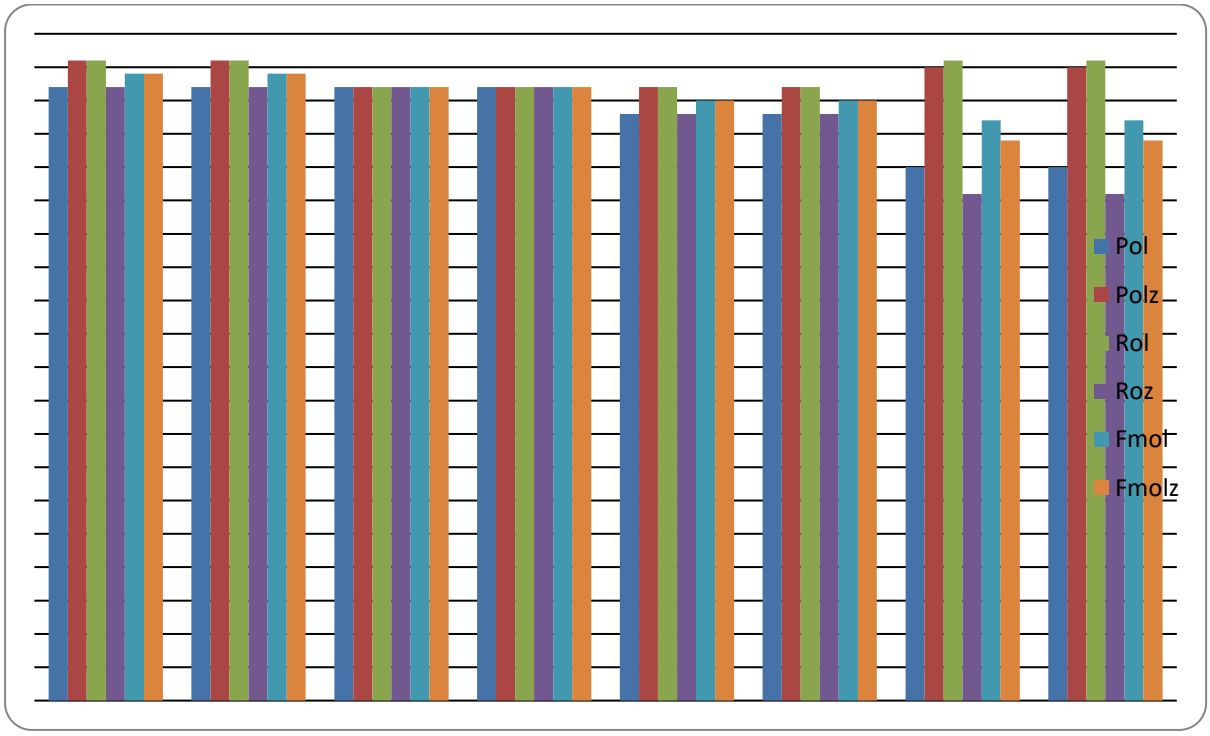
Şekil 8: Yeni Yöntemin Olumlu ve Olumsuz sınıflar için özelliklere göre P, R, FM karşılaştırması



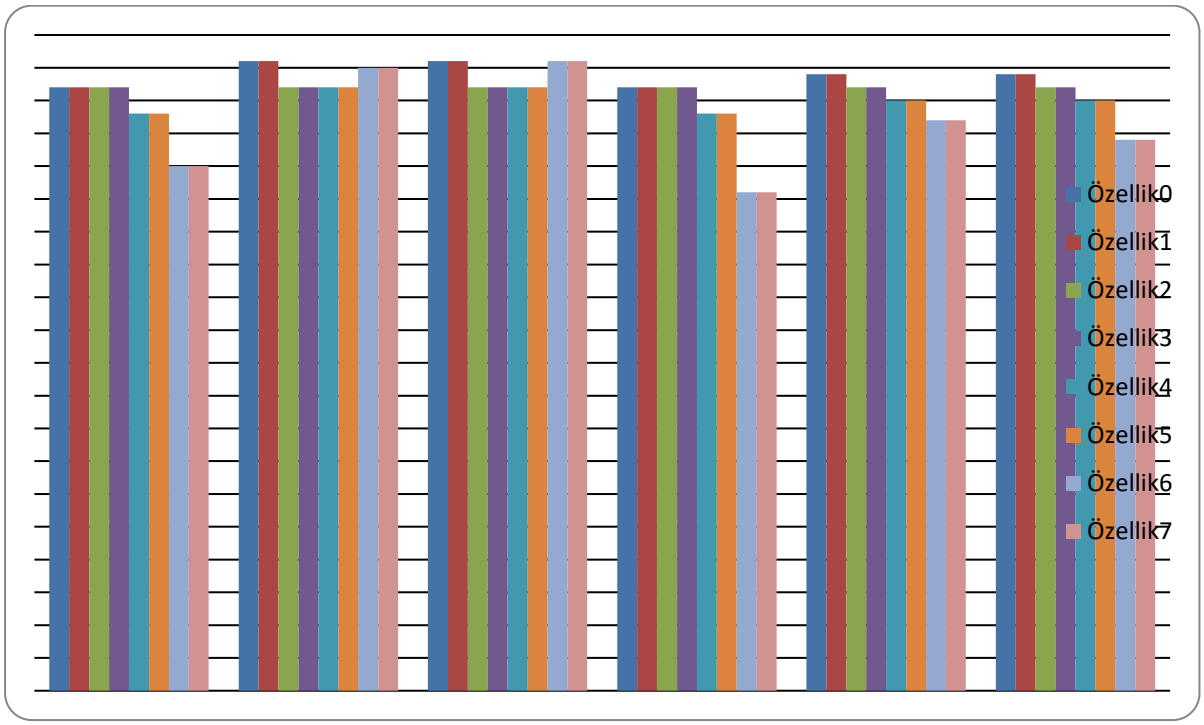
Şekil 9: MBNB yönteminin Olumlu ve olumsuz sınıflar için her bir özellikteki P, R, FM değerleri



Şekil 10: MBNB yönteminin Olumlu ve Olumsuz sınıflar için özelliklere göre P, R, FM karşılaştırması



Şekil 11: MBNB yönteminin olumlu ve olumsuz sınıflar için her bir özellikteki P, R, FM değerleri



Şekil 12: MNB yönteminin Olumlu ve Olumsuz sınıflar için özelliklere göre P, R, FM karşılaştırması

Sonuç Yorum Ve İrdeleme

Duygu analizi ve düşünce madenciliği teknolojinin ilerlemesi ve elektronik ortamda insanların duygu ve düşüncelerini dile getirdikleri verilerin artması ile birlikte önem kazanan bir alan olmuştur. Son yıllarda bu alana yönelik çalışmaların üssel olarak artması ve bu konuda yapılacak çalışma ve uygulamalara astronomik bütçeler ayrılması konuyu oldukça cazip hale getirmektedir. Günümüz imkanları kullanılarak bu alanda insanların faydalanacağı pek çok uygulama geliştirmek mümkündür.

Başka ülkelerde bu alanda çok fazla çalışılırken henüz Türkçe üzerinde bu konuda yeni yeni kıpırdanmalar olmaktadır. Biz de bu gelişmekte olan ve heyecan verici alanda bir çalışma yapmak istedik. Sinema yorumlarını kullanarak önemli sonuçlar elde ettik. Ancak bu konuda geliştirilebilecek çok fazla şey olduğunu da gördük. Her ne kadar NB ve basit tekniklerle uygulama yapmış olsak da bu basit ve performansı yüksek yöntemlerle çok iyi sonuçlar elde ettik. Bununla birlikte, yeni yöntem olarak bahsettiğimiz sözlük çıkartma yönteminde önemli terimlerin elle seçilen işlemi daha meşakatli olduğu halde, geleneksel makine öğrenimi yöntemlerine kıyasla daha başarılı sonuç vermemiştir. Bununla birlikte daha fazla yorumdan elde edilen sözlükler daha iyi sonuç verecektir. Sözlük yönteminden daha iyi sonuç alabilmek için önemli terimler daha dikkatli seçilebilir, sayıları arttırılabilir ve özellikle ikili terimler daha özenle seçilebilir. Yeni yorumlardan gelen yeni önemli

sözler sözlüğe eklenebilir. Aynı şekilde kipleri içeren yöntem orijinal bir buluş olmakla beraber, kipli terimlerin youmlarda çok az geçmesinden dolayı beklenen yüksek başarı gösterilmemiştir, fakat bu tür yöntemlerde tüm terimleri kullanmak yerine belirgin terimleri içeren sadece küçük kelime kümelerinden özellikler seçilerek başarı oranını arttırabilir. Sözlük hazırlama yöntemi, kiplerle birlikte ve özellikle makine öğrenimi yöntemleriyle birlikte kullanılınca hibrid yöntem daha iyi sonuç verecektir. Bu tür çıkarımların bizi ve bu alanda ileride çalışma yapmak isteyenleri aydınlatacaklarına inanmaktayız.

Kaynaklar

Jurafsky, D. What is sentiment analysis, https://web.stanford.edu/class/cs124/lec/sentiment.pptx.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, ChengXiang Zhai Comprehensive Review of Opinion Summarization, http://kavita-ganesan.com/sites/default/files/survey_opinionSummarization.pdf

Hidayet Takçı, Duygu Analizi (Sentiment Analysis), http://verimadencisi.blogspot.com.tr/2013/08/duygu-analizi-sentiment-analysis.html,2013.

Allie Gray Freeland, SEO and Social Secrets Every Entrepreneur Must Know and Why, https://www.huffpost.com/entry/6-seo-and-social-secrets-_b_5505296, 2017

Türkçe otomatik sentiment analizi (otomatik duygu analizi uygulaması, http://www.sentimentanalizi.com/, 2014

Vasilis Vryniotis, Machine Learning Tutorial: The Naïve Bayes Text Classifier, http://blog.datumbox.com/machine-learning-tutorial-the-Naïve-bayes-text-classifier, 2013.

Zhang, Harry, The optimality of Naïve Bayes, AA 1.2 (2004): 3.

Christopher D. Manning, Prabhakar Raghavan ve Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, http://nlp.stanford.edu/IR-book/html/htmledition/properties-of-Naïve-bayes-1.html, 2008

Roman Daneghyan, 54 SEO Statistics That Will Impact Your Business in 2020, https://www.safaridigital.com.au/blog/seo-statistics-2019, 2019

Allie Gray Freeland, SEO and Social Secrets Every Entrepreneur Must Know and Why, https://www.huffpost.com/entry/6-seo-and-social-secrets-_b_5505296, 2017

Ender Ahmet Yurt, Duygu Analizi, http://www.slideshare.net/webender/duygu-analizi, 2014

Fatih Amasyalı, Yeni Makine Öğrenmesi Metotları ve İlaç Tasarımına Uygulanması, Doktora Tezi, Fen Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi, İstanbul, 2008.

Ethem Alpaydın, Yapay Öğrenme kitabı, Boğaziçi Üniversitesi Yayınevi, İstanbul, 2010

Weka, http://tr.wikipedia.org/wiki/Weka

Ian H. Witten; Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, Weka: Practical Machine Learning Tools and Techniques with Java Implementations, Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems.

Şadi Evren ŞEKER (2013) (Türkçe). İş Zekası ve Veri Madenciliği (Weka ile) ISBN 9786051276717. Cinius.