BİR SOSYAL AĞDAN ALINAN VERİLERİN ANLAMSAL KUTUPLANDIRILMASI

Yüksek Lisans Tezi

Dilber ÇETİNTAŞ

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Taner TUNCER

HAZİRAN-2018

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİR SOSYAL AĞDAN ALINAN VERİLERİN ANLAMSAL KUTUPLANDIRILMASI

Yüksek Lisans Tezi

Dilber ÇETİNTAŞ
(**131129106**)

Danışman: Doç. Dr. Taner TUNCER

Haziran-2018

T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİR SOSYAL AĞDAN ALINAN VERİLERİN ANLAMSAL KUTUPLANDIRILMASI

YÜKSEK LİSANS TEZİ

Müh. Dilber ÇETİNTAŞ

131129106

Tezin Enstitüye Verildiği Tarih : 06/06/2018
Tezin Savunulduğu Tarih : 29/06/2018

Tez Danışmanı : Doç. Dr. Taner TUNCER (F.Ü)
Diğer Jüri Üyeleri : Doç.Dr. Erdinç AVAROĞLU (Mersin Ü.)
Dr. Öğr. Üyesi Ebubekir ERDEM (F.Ü)

Haziran-2018

ÖNSÖZ

Bu tez çalışmasında bir sosyal ağdan alınan verilerin anlamsal kutuplandırılması amaçlanmıştır. Twitter üzerinden alınan verilerin doğal dil işleme yöntemleri kullanarak olumlu, olumsuz, nötr olarak sınıflandırılması hedeflenip bu kutuplanmayı etkileyen sosyal ağ ilişkileri incelenmiştir. Sonuçlar umut verici düzeyde pozitif sonuçlanmıştır.

Çalışmamın her aşamasında yol göstericiliği ile yanımda olan değerli danışmanım Sayın Doç. Dr. Taner TUNCER'e sonsuz teşekkürlerimi sunuyorum.

Çalışmamın ilerlemesinde çok yardımı olan ve bugünlere gelmemede emek harcayan aileme çok teşekkür ederim.

Dilber ÇETİNTAŞ

ELAZIĞ - 2018

İÇİNDEKİLER

ÖZET

Bu tez çalışmasında twitter kullanıcılarının atmış olduğu tweetlerden yararlanarak internet üzerinde dolaşan kirli ve düzensiz bilginin daha elveririşli hale getirilmesi amaçlanıştır. Günümüzde sosyal ağ kullanımının artması ve kullanıcıların yoğun talep göstermesi paylaşılan içeriklerin kontrolünü gündeme getirmiştir. Kötü niyetli kişilerin algı yönetimi, provakasyon, sahtekarlık amaçlı metinlerin önüne geçilmesi mevcut verileri sınıflandırma fikrini ortaya çıkarmıştır. Sınıflandırmadan kasıt verinin toplumun etik değerleri, dil ve lehçe göz önüne alınarak anlamsal olarak olumlu, olumsuz, nötr olarak belirtilmesidir.

Bu yüzden analiz amaçlı NodeXL ile Twitter üzerinden alınan datalar gereksiz kelime ve noktalama işaretlerinden arındırılmış, yanlış yazımlar, lehçe farklılıkları elimine edilmiştir. Yapılandırma aşamasından sonra tweetler daha doğru sonuçlar alınabilmesi adına herbir tweet beş farklı gönüllü tarafından olumlu, olumsuz, nötr olarak etiketlenip sayı değeri yüksek olan etiket değeri sonuç olarak kabul edilmiştir. Hazırlanan data seti knime programı yardımıyla kullanılan kelime sıklığına göre özellik vektörleri oluşturulup 70'a 30 oranında öğrenme ve tahmin grupları DVM(Destek Vektör Makineleri), Naive Bayes, Karar Ağaçları algoritmalarıyla sınıflandırılmıştır. Son aşamada ise ağ ilişkilerinin anlam üzerindeki etkisi araştırılmıştır.

Anahtar Kelimeler: NodeXL, Twitter, Metin Dönüşümü, Veri madenciliği

SUMMARY

Semanticaly Polarizing Data From A Social Network

In this thesis, it is aimed to make dirty and irregular information circulating on the internet more convenient by taking advantage of the tweets that twitter users have taken. Increasing use of social networking today and the intensive demand of users have brought about the control of shared content. Prevention of malicious people 's perception management, provocation, fraudulent texts reveals the idea of classifying existing data. The notion of classification is that the ethical values of the community are to be semantically positive, negative, neutral, taking into account language and dialect.

Therefore, NodeXL for analysis and data received via Twitter are eliminated from unnecessary words and punctuation, misspellings, polish differences have been eliminated. After the configuration phase, the tweet was tagged as positive, negative, neutral by five different volunteers each tweet in order to get more accurate results, and the tag value with high number was accepted as the result. The prepared data set is created by feature vectors according to the word frequency used with the help of the knime program, and the learning and estimation groups of 70 to 30 are classified by SVM (Support Vevtor Machine), Decision Tree and Naive Bayes algorithms.

Key Words: NodeXL, Twitter, Text Conversion, Data Mining

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

SEMBOLLER ve KISALTMALAR

API **:** Application Programming İnterface

DVM **:** Destek Vektör Makinesi

NodeXL **:** Network Overview Discovery and Exploration for Excel

KA **:** Karar Ağaçları

1. GİRİŞ

İnternetin paylaşım ortamı olarak kullanılmasıyla oluşan sosyal ağlar iletişimi daha etkin ve hızlı hale getirmesiyle sosyal medya gücünü ortaya çıkarmıştır. Sadece iletişimle sınırlı kalmayan bu sistem reklam alanında da kullanılarak süper güç haline gelmiştir. Sosyal ağlar kişisel kullanımlarının yanında firmaların bütçesine katkı sağlayacak şekilde kampanyaların, tanıtımların aktif olarak yayımlandığı yer olması itibariyle de bu yükselişi ciddi anlamda tetiklemiştir. Bu anlamda ülkemizde hem kullanıcılar hem de firmalar tarafından büyük önem arz etmektedir.

İnternet kullanıcılarının ortak bir ilişki, hızlı bilgi paylaşımı, anlık konum iletileri, yeni arkadaşlıklar için kullandıkları sanal oluşumlardır. Kullanıcılarına sanal bir profil oluşturarak etkileşimi başlatan bu ağlar web sitelerinin gelişimiyle beraber sınırlar ötesinde sosyalleşme kavramını gerçekleştirmeye başlamıştır. Bu nedenle sosyal ağlar topluluklar arasında giderek yayılmakta kullanımı neredeyse zorunlu hale gelmektedir.

Sosyal ağ yazılımlarının sağlamış olduğu sosyal içerik her yaştan ve her gruptan kullanıcının katılmasını desteklemektedir. Ses, resim, video, görüş içeriklerini kapsayan ve her geçen gün yenisini ekleyen bu platforma dahil olma yıllar içerisinde katlanarak artmaktadır.

Kendi amaçları doğrultusunda bu online dünyaya katılanlar zamanlarının büyük bir kısmını burada geçirmeye başlamıştır. Gündem olaylarını, çevresinin özel yaşantı paylaşımlarını bu ortamdan takip cazip gelmiştir.

Ayrıca bireyler sosyal ağlar sayesinde ortak çalışmalar yapmakta, çalışmalarını ve araştırmalarını sosyal ağlar aracılığıyla paylaşabilmektedirler. Bireylerin aktif performansları nedeniyle artan hareketlilik ve etkileşimler ve oluşan online dünya içine girerek daha çok renkleri, inanışları ve ülkeleri katmıştır. Günümüzde, birçok sosyal ağ sitesi ortaya çıkmış ve insanların iletişimini, etkileşimini, işbirliğini, çalışmasını ve hatta öğrenme sürecini bile yeniden şekillendirmiştir [1].

Bunun en önemli ve güçlü örneğini Amerika Başkanlık seçimlerinde Barrack Obama'nın sosyal ağları ve sosyal medyayı bir arada kullanarak yaptığı kampanya gözler önüne seriyor. Bunun yanında dünyada başta twitter olmak üzere Facebook ve Friendfeed gibi

sosyal ağları kullanarak şirketlerine maddi kazanç elde etmeyi başaran onlarca örnek de bu alanda nasıl bir güç ile karşı karşıya olduğumuzun en önemli kanıtları arasındadır [2].

1.1 Tezin Amacı

Sosyal ağlar hem ticari hem siyasi anlamda toplum bilincinin anlaşılabilmesi açısından hedef ortamlar olmaktadır. Kullanıcıların tepkilerinin ölçülerek alınan feedbackler doğrultusunda ileriye dönük adımlar atmak sosyal ağları gelecek planlarını yaparken başvurulacak önemli bir basamak noktasına getirmiştir.

Bu avantajlarının yanısıra çok hızlı bir yayın ortamı olması olumsuz istenmeyen algıların da kısa sürede çok fazla kullanıcıya ulaşmasına ve istenmeyen durumların oluşumuna sebebiyet vermektedir.

Bu tez çalışmasında bir sosyal ağdan alınan verilerin anlamsal olarak kutuplandırılması amaçlanmaktadır. Türkçenin anlam zenginliğinden kaynaklı aynı cümleden farklı anlamların çıkarımının yapılmasının önüne geçilerek anlam kargaşasının önüne geçmek, doğal dil ile yapılan tüm işlemleri bu sisteme uyarlayarak olumsuzlukları önceden farkedip hızlı bir şekilde yayılmasının önüne geçebilmek hedeflenmektedir.

1.2 Teze Bakış

Tez çalışmasının ikinci bölümünde daha önce yapılan benzer çalışmalar sunulmaktadır, üçüncü bölümde kullanılan kavram ve tanımları yer almaktadır , dördüncü bölümde yapılan çalışmalar ele alınırken , beşinci bölümde çalışmamız sonucunda elde edilen değerlerden ve etkilerinden bahsedilmiştir beşinci bölümde; tez çalışması süresince elde ettiğimiz sonuçlar aktarılmıştır.

2. LİTERATÜR TARAMASI

Kullanıcılara sunduğu kullanım kolaylığı sayesinde tüm kesimlere hitap eden ve kişilerin fikir paylaşımını gerçekleştirdiği Twitter çalışmamız için uygun sosyal ağ olmuştur. Bu nedenle araştırmamızı daha önce Twitter üzerinde yapılan çalışmalar üzeride yoğunlaştırdık. Bu çalışmalarda büyük veriler üzerinde yapılan dil işlemleri ve metin madenciliği yol gösterici olmuştur.

"Twitter Verileri ile Türk Televizyonları İzlenme Oranı Sıralamaları Tahmini "[3] çalışmasında program ismi, oyuncu ismi, program isminden oluşturulmuş etiketler yardımıyla 01-11-2014/ 28-06-2015 tarih aralığındaki tweetler toplanmıştır. İkinci bir modül ratinglerin tutulduğu web sayfasına giderek izlenme oranı sonuçları alınmıştır. Böylelikle sistemde toplanan veriler tweetler, twitter kullanıcıları, kanallar, programlar, izlenme oranı ve sıralamaları olmuştur. Toplanan bu datalar doğal dil işleme aşamasından geçirilerek 3 farklı kişi tarafından olumlu, olumsuz, nötr olarak gruplandırılmıştır. n-gram eğitim seti ve Naive Bayes, Destek Vektör Makinesi, Rastgele Orman sınıflandırıcılarının kullanıldığı projede %78,59 oranla en yüksek başarı destek vektör makinesinde en düşük başarı %76,26 ile Naive Bayes'de elde edilmiştir. İlk 5, ilk 10, ilk 15 şeklinde kümeleme yoluna gidildiğinde sayı arttıkça doğruluk oranının arttığı gözlemlenmiştir.

"Twitter Verileriyle Duygu Analizi ve Türkçe Duygu Kütüphanesi"[4] tezinde oluşturulan duygu kütüphanesine 1580 kayıt eklenerek olumlu kelimelere (+) puan, negatif kelimelere(-) puan verilerek cümlenin duygusu belirlenmeye çalışılmıştır. Hava ve haftanın günlerinin duygular üzerindeki etkisini ölçmek için alınan belirli etiketteki tweetler emoji ve kelime içeren düşünce kütüphanesi yardımıyla gruplanmaya çalışılmıştır. Cuma ve Cumartesi günleri olumlu, pazartesi olumsuz tweetler paylaşıldığı gözlemlenmiştir. Hava durumuna göre yapılan analizde Sonbahar ayında en fazla verinin toplandığı, kış aylarında olumsuz tweet yaz aylarında olumlu tweet yayınladıkları ölçülmüştür.

"Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi "[5] adlı çalışmada duygu analizi alanında yapılan çalışmalardan farklı olarak uzaklık ve benzerlik fonksiyonlarının kullanarak performans değerlendirilmesi yoluna gidilmiştir.

Twitter'da Türkçe Veriler Üzerinde Hakaret Suçu Analizi"[6] başlıklı analiz çalışmasında hakaret içerikli tweetleri bulabilmek adına hakaret içeren ve hakaret içermeyen iki kelime grubu oluşturularak bu kelimelere göre tarama işlemi gerçekleştirilmiştir. TF-IDF ağırlıklandırma yönteminin Destek Vektör Makine algoritmasının kullanıldığı durumda %95,4 başarı elde edilmiştir.

"Twitter Üzerindeki Türkçe Mesajlarda Veri Madenciliği ile Duygu Analizi"[7] konulu tezde rastgele alınan 62347 tweet doğal dil işlemlerinin ardından 100 gönüllü tarafından anlamsal olarak ifade ettiği duyguya göre 'Mutluluk', 'Kızgınlık', 'Üzüntü' ve 'Şaşkınlık' etiketlenmiştir. En az iki gönüllü tarafından aynı etiketlenen sonuç etiketi olarak kabul edilmiştir. Karar ağacı ile sınıflandırma yapıldığında %85,372'lik başarı oranı Fuzzy Rule Learner kullanıldığında %83,608'lik başarı oranı elde edilmiştir.

Meriç Meral tarafından yapılan "Twitter Verilerini Anlamsal Sınıflandırma"[8] çalışmasında 9 farklı alandan alınan tweetler kelime tabanlı ve n-gram yöntemiyle işlenip 8321 tanesi gönüllü kişilerce etiketlendikten sonra Naive Bayes, Rastgele Orman, Destek Vektör Makineleri kullanılarak olumlu, olumsuz, nötr olarak sınıflandırılmıştır.En yüksek başarı sözlük tabanlı yöntem rastgele orman sınıflandırıcısından elde edilmiştir.

Malak Abdullah ve Mirsad Hadzikadic tarafından yapılan çalışmada kolay ve etkin kullanımı sayesinde kişilerin politik görüşlerini yansıttığı twitter ' dan Donald Trump 'ın yer aldığı başkanlık seçimlerindeki sosyal medyanın etkisi hesaplanmaya çalışılmıştır. Çalışmada adayların seçmenler üzerindeki etkisini ölçmek için 85000 tweet toplanmıştır. Tarafsız olabilmek ve sadece secimle ilgili tweetleri alabilmek için #voteTrump, #trump2016, #Donald_Trump #antiTrump hashtagleri kullanılmıştır. Sadece İngilizce tweetler alınırken retweetlerden kaçınılmıştır. Verilerin temizlenmesi aşamasında boşluklar, rakamlar, noktalama işaretleri silinip, tüm harfler lowercase olarak ayarlanmıştır. Üçüncü aşamada alınan tweetlerdeki duygular destekleyen ya da desteklemeyen duygu olarak öfkeli, şaşırmış, keyifli… sınıflandırmaları kullanılmıştır. Sonuç olarak %58 desteklemeyen %42 destekleyen tweet; duygu analizinde %66 keyifli bulguları elde edilmiştir. Çalışmanın doğruluğu %68 olarak ölçülmüştür [27].

Shreya Ahuja ve Gaurav Dubey twitter verilerini kullanarak kullanıcıların duygu ve fikirlerini bulmayı amaçlayan bir çalışma yapmışlardır. İlk olarak Phyton Araçlarından Tweepy kullanılarak Twitter API aracılığıyla belirlenen zaman ve istenilen dildeki tweetler toplanmıştır. Alınan her bir tweet token'lara ayrılmıştır. Verilere pozitif ya da negatif

puanlar atayabilmek için 2477 kelime içeren Finn Arup Nielsen tarafından oluşturulmuş AFINN Dictionary kullanılmıştır. Kişisellik değerini ölçmek için ise sadece Phyton üzerinde çalışan TextBlob kütüphanesi tercih edilmiştir. Bu kütüphane 0.0 değerini döndürürse objektif, 1.0 değerini döndürürse subjektif olduğunu belirlemiştir. 0.4 oranı ve üzerindeki kişisellikteki değerlerin pozitif kutupsallığa 0.2 altındaki değerlerin negatif kutupsallığa ait olduğu fark edilmiştir. Nötr kutuptaki değerlerin pozitiflerle hemen hemen aynı oranlara sahip olduğu gözlemlenmiştir [28].

Jinny Yan ve Suzanne J. Mathews adlı kişilerin yaptığı çalışmada Çince atılmış tweetler alınarak yazar tespitinin yapılması amaçlanmıştır. Karakter tabanlı olan Çince gibi bir dilde 50000'den fazla karakter bulunması işlemi zorlaştırmıştır. Çince dili için 140 karakterin artırılmaması bir başka negatif etken olmuştur. Ele alınan on yazarın her birinin yayınladığı 100 tweet alınarak n-gram yöntemi sonucu elde edilen veriler k-means ve Expectation Maximization kümeleme yöntemleri WEKA aracılığıyla kullanılarak hangi yazara ait olduğu saptanmaya çalışılmıştır. Üç yazar için başarı oranı %44,53, beş yazar için %29,24, on yazar için %20,52 doğruluk oranlarına ulaşılmıştır [29].

3. ARKAPLANDAKİ KAVRAMLAR

3.1.Veri Madenciliği

Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır [9]. Günümüzde teknolojinin hızla ilerlemesi ve her alanda yer bulmasıyla tüm bilgiler sayısal ortamlara aktarılmaya başlanmıştır. Mağazalardaki satışlar, müşteri bilgileri, ürün bilgileri, kamera görüntüleri veri tabanlarında saklanmaya başlanmıştır. Her geçen gün artan verilerden anlamlı sonuçlar çıkarmak kişi ve kurumlara stratejik karar verme noktasında avantaj sağlayacaktır. Bu çıkarımlarda önemli olan nokta tahmin edilebilir bir bilgi olmamasıdır. Ayrıca verilen kararın daha doğru olabilmesi için alınan verilerin doğru depolanması, temizleme ve sınıflandırma işlemlerinin iyi yapılmış olmasını gerektirmektedir.

3.2 Twitter

Jack Dorsey tarafından 2006 yılında iletişim amaçlı sms sağlayıcı olarak başlatılan yaygın olarak kullanılan sosyal paylaşım ağıdır. 2011 yılında Türkçe dil desteği vermeye başlamıştır. Paylaşılan tweetlerin tüm twitter kullanıcıları tarafından görülebilmesi fakat kullanıcı isterse kısıtlama yapabilmesi sağlanmıştır. Ayrıca kullanıcı takip etmek istediği kurum /kuruluş/ kişi tweetlerine üye olması sağlanarak güncel paylaşımlarından haberdar olabilme imkanı sunulmuştur. Twitter 2017 Eylül ayına kadar 140 karakterle sınırlı olarak insanların duygu ve düşüncelerini paylaştıkları sosyal mesajlaşma ağı olarak bilinmekteydi 2017'nin 26 Eylül'de yapılan değişiklikle karakter sayısı Japonca, Korece, Çince hariç geriye kalan tüm diller için iki kat arttırarak 280 karakter olarak güncelledi [4].

2018 analiz sonuçlarına göre 670 milyon Twitter kullanıcısı bulunmaktadır ve bunların %60 oranının aktif olduğu görüntülenmektedir [10]. Gün geçtikçe bu oranın daha da yükselmesi twitter'ın kendi dilini oluşturmasına sebebiyet verdiği gözlemlenmektedir. Bu terimlerden bazıları:

RT(Retweet): Herhangi bir kişi ya da kuruluşun yapmış olduğu paylaşımı beğenip kendi sayfasında beğenme durumudur. Kısaca alıntı yapılan kaynağı belirtme durumudur.

Hashtag : # kullanılarak kelimeleri etiketleme işlemidir. Atılan tweetler arasında yapılan filtreleme olarak düşünülür. Aynı etiket başlığı altında gruplar oluşturmayı amaçlar.

TT (Trend Topic): Konunun popülerliğini belirtir. Genellikle bir hashtag'in defalarca paylaşılması sonucu oluşur.

Mention: Yapılan paylaşımlarda kullanıcı adlarının kullanılması durumudur.

3.3. Twitter API

Twitter API twitter içerisinde yer alan özellikleri başka uygulamalarda kullanmaya yarayan ara yüzdür. Twitter API'si twitter verisine programatik olarak okuma ve yazma yapılmasını sağlar. Kısacası, Twitter'ı REST API'lerine ve Streaming API'lerine ulaşabilmek için kullanılan Phyton paketidir [7]. Şekil 3.1 API yardımıyla çekilen bir tweet örneğini göstermektedir.

```php
<?php
    header('Content-type: text/html; charset=utf8');
    require "tmhOAuth.php";



    //Consumer ve Accsess degerleri
    $consumer_key=' iTFEZdt3JLqpv1Kh2FiFkvfHi';
    $consumer_secret='zpLEXYmUO8HpPt2tb7tGVbN33RFsyq6N6ywqfX7RsRfg2iPrY2';
    $access_token=' 920896272533524480-Ck0vdAhVtUHy87CBbPVjhUYeew3MbX3';
    $access_token_secret='k424XVZNPllmkWO1PBFpbGysb37FTFwNTa3Hn61y7dg0h';
    //sınıfın başlatılması
    $twitter= new tmhOAuth( $consumer_key,$consumer_secret,$access_token,$access_token_secret);
    //kullanıcı adı
    $username='DilberCetintas';
    //tivit sayısı
    $count=1;

    $tweets=$twitter->get('GET https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name='.$username.'&count='.$count);
    print_r($tweets);

    foreach($tweets as $tweet){

        $id=$tweet->id_str;
        $text=$tweet->text;
        $created_at= date("Y-m-d M:i:s", strtotime($tweet->created_at));

        echo'<a href="https://twitter.com/'.$username.'/statuses/'.$id.'">
        '.nl2br($text).'</br>
        '.$created_at.'

        </a>';
    }
```

Şekil 3.1. API Yardımıyla Çekilen Tweet Örneği.

Twitter API twitterdan veri çekme yada ekleme işlemlerinde kullanılır. İşlem yapılabilmesi için birkaç izin alınması gerekmektedir. Bu izinler hizmet sağlayıcının hizmetinden yararlanabilmemiz için kim olduğumuzu anlamaya yönelik oluşturmuş olduğu anahtarlardır. Twitter API bu anahtarları kullanarak yetki verme ya da vermeme kararı almaktadır. Eğer anahtar geçerli ise Oauth kullanarak kimlik doğrulamayı gerçekleştirmektedir. Şekil 3.1'de uygulama ve erişim anahtarları kullanılarak Twitter

hesabına giriş yapmaya gerek kalmadan DilberCetintas adlı kullanıcı hesabına sahip kişinin atmış olduğu tweet, id, tweet tarihi gibi bilgiler elde edilmiştir.

3.4. Duygu Analizi

Duygu analizi temel olarak bir metin işleme (text processing) işlemi olup verilen metnin duygusal olarak ifade etmek istediği sınıfı belirlemeyi amaçlar. Duygu analizinin ilk çalışmaları duygusal kutupsallık (sentimental polarity) olarak geçmekte olup verilen metni olumlu (pozitif), olumsuz (negatif), ve nötr olarak sınıflandırmayı amaçlamaktadır [11]. Duygu analizi çalışmalarında sadece metinler kullanılmamaktadır. Kişilerin duygu durumlarını ifade eden emoji kullanımlarının da duygu sınıflandırmasında önemli bir etmen olduğunu gösteren birçok çalışma mevcuttur . Örneğin Oxford Sözlük tarafından düzenli olarak her yıl seçilen yılın kelimesi, 2015 yılında bir kelime değil bir emoji(Face with Tears of Joy/ Mutluluk Göz Yaşları) olmuştur [25]. Sadece Twitter'da, 2014–2016 yılları arasında 110 milyardan fazla emoji tweet edilmiştir [26].

Metin madenciliği Yöntemleri ile Duygu Analizi [25] adlı çalışmada emojiler 15 farklı gruba ayrılıp en az birini içeren 8.000 ile 10.000 arasında tweet alınmıştır. Alınan bu tweetler sınıflandırıldığında Naive Bayes ile %52,09 başarı elde edilmiştir. Bu durum sadece emoji kullanımının duygu belirtimindeki etkisini göstermektedir.

Bing Liu'ya göre gerçek dünyada büyük organizasyonlar ve tüm iş dalları hizmetler ya da ürünler hakkında tüketicilerin görüşlerini bilmek isterler. Aynı zamanda karşılıklı olarak da tüketici ürünlerin mevcut kullanıcıları tarafından yapılan görüşleri bilmek ister [12]. Sadece üretici ve tüketici değil finanstan tıp alanına kadar birçok şahıs duygu analizi yöntemleriyle yapılacak yorumlara ihtiyaç duyar. Bu durum sosyal medyada aktarılanları daha da önemli kılmaktadır. İyi kararlar alabilmek adına sıklıkla sosyal medyaya başvurulması bu ortamlardaki duygu analizini zorunlu hale getirmektedir. Fakat analiz işlemlerini gerçekleştirirken dil kaynaklı bazı problemlerle karşılaşılmaktadır. "Böyle bir yemek nasıl beğenilir" denildiğinde amaç olumsuzluk belirtmek olmamasına rağmen soru olarak algılanması problemlerden biridir. Başka bir örnek vermek gerekirse "Sınavı kazanmayı çok istedi fakat çalışması yetersiz oldu" şeklinde olumlu bir cümlenin ardına gelen olumsuz cümle tüm anlamı olumsuz hale getirebilmektedir. Ya da "okula gelmemiş olman kalacağın anlamına gelmez" şeklindeki ters ifade olumsuz olarak görünse de olumlu anlam ifade etmektedir. Bu olumsuzlukları gidermek amaçlı yapılan her bir adım Duygu (Sentimental) analizin gelişim süreçlerini oluşturmaktadır.

3.4.1 Metin Önişleme

Metni kelimelere ayırma, kelimelerin anlamsal değerlerini bulma (isim, sıfat, fiil, zarf, zamir vb.) kelimeleri köklerine ayırma ve gereksiz kelimeleri ayıklama, yazım kurallarına uygunluğunu tespit etmek ve var olan hataları düzeltmek gibi metin belgelerin yapıtaşı olan kelimelerle ilgili işlemleri içeren süreçtir. Metin madenciliğinin en büyük sorunu, işleyeceği veri kümesinin yapısal olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışan metin madenciliği alanında ön işleme aşaması veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir [12]. Twitterdan alınan verilerde yazım hatalarının, karakter sayısının sınırlı olması sebebiyle kelimelerde yapılan yanlış kısaltmaların, url paylaşımlarının yer alması veri analizini zorlaştırmaktadır. Bu problemlerin üstesinden gelebilmek için mention (@ile başlayan kelimeler), hashtag ve url silinmelidir. Kullanılan diğer önişleme kavramları:

Snowball: Birçok dilde kök bulma algoritmalarına sahip karakter işleme dilidir. Türkçe sondan eklemeli bir dil olduğu için ve her eklenen yapım eki kelimeye yeni bir anlam kazandırdığı için kök bulma da zorluklar yaşanabilmektedir.

Stopword İşlemi: Tekrar eden, istenmeyen, belirli bir karakter sayısından az olan, anlam ifade etmeyen kelimelere verilen addır. Bu tür kelimelerin diskalifiye edilmesi sırasında kullanılır. Bu durum bilgiye erişimi kolaylaştırdığı gibi sonucu değiştirme olasılığı olan kelimelerin elimine edilmesine katkı sağlar.

Punctuation Erasure: Semantik analiz çalışmalarında noktalama işaretlerinin anlam ifade etmemesi sebebiyle metinde yer alan tüm yazım işaretlerini silmek amacıyla kullanılır.

Number Filter: Analizi daha verimli hale getirebilmek için kullanılan sayısal ifadelerin kaldırılmasıdır.

N Chars Filter: Belirli karakter sayısından az alan ve istatistiği etkilemeyecek kelimeleri azaltır. Türkçe dili için örneklemek gerekirse “ve” bağlacı sık kullanılır fakat duygu analizi konusunda anlam ağırlığı yoktur.

Bag of Words: Bu aşamada gruplanan tüm dokümanlardaki tüm kelimelerin kullanım sıklıkları hesaplanır ve bir havuzda toplanır. Daha sonra ise bu kelimelerin değerleri (Word Weighting) hesaplanır [12]. Kelime ağırlığı, tüm metin içindeki geçme sıklığı olarak belirlenir.

3.4.2 TF (Term Frequency) Özellik Çıkarımı

Dokümanlar bilgisayarda bir anlam ifade etmez. Metinlerin işlenebilmesi, yapısallaştırılması ve basitleştirilmesi aşamasına özellik çıkarımı denir. Örneğin bir miktar resimden içinde çimen bulunanların tespit edilmesi isteniyor olsun. Bilindiği üzere çimenler yeşildir ve resimlerden yeşil tonun ağırlıkta olanlarının çimen içermesi ihtimali yüksektir. Öyleyse sisteme giren bir resmin tamamının işlenmesi yerine resmin histogramının (renk kodlarının dağılımının) işlenmesi buna bir örnek olabilir. Resim, basitçe histogramından çok daha karmaşık ve büyük bir veridir. Bu veri histogramı (tekrar dağılımı) çıkarılmak suretiyle küçültülmüş ve istenen amaca yönelik olarak işlenmiştir [13].

Metinleri ağırlıklandırma da kelime-torba yöntemleri yer alır. Yani torba içindeki kelimelerden sayısı fazla olanın ağırlığı daha fazladır. Bu yönteme göre her bir kelime ağırlandırılarak temsili vektör oluşturulur. Bu vektörleri oluştururken TF, TF-IDF kullanılır. TF(Term Frequency) kelimenin grup içi ağırlığını, IDF gruplar arası ağırlığını hesaplamada kullanılır.

Terim frekansına göre sık geçen kelime konuyla daha ilişkilidir ve ağırlık oranı daha fazladır. T teriminin d dökümanındaki frekansı denklem 3.1 deki gibi hesaplanır.

$$\mathrm{TF}(t,d)=0.5 * (0.5 * f(t,d))/(MaxFreq(d)) \quad \textbf{(3.1)}$$

IDF (Inverse Document Frequency) fonksiyonunda az sayıda bulunan kelimelerin daha ayırt edici olacağı fikrinden yola çıkmıştır. Genellikle dokümanları ayırt etme işlemlerinde tercih edilir. IDF terim hesaplaması denklem 3.2' deki gibi hesaplanır. Burada n toplam doküman sayısını k ise terimin geçtiği doküman sayısını belirtir.

$$IDF(t) = 1 + \log(n/k) \quad \textbf{(3.2)}$$

Bu modele göre terimin önemi, belge içerisinde o terimin geçme sayısıyla doğru orantılıyken belge havuzu içerisinde o terimin geçme sıklığıyla ters orantılıdır.

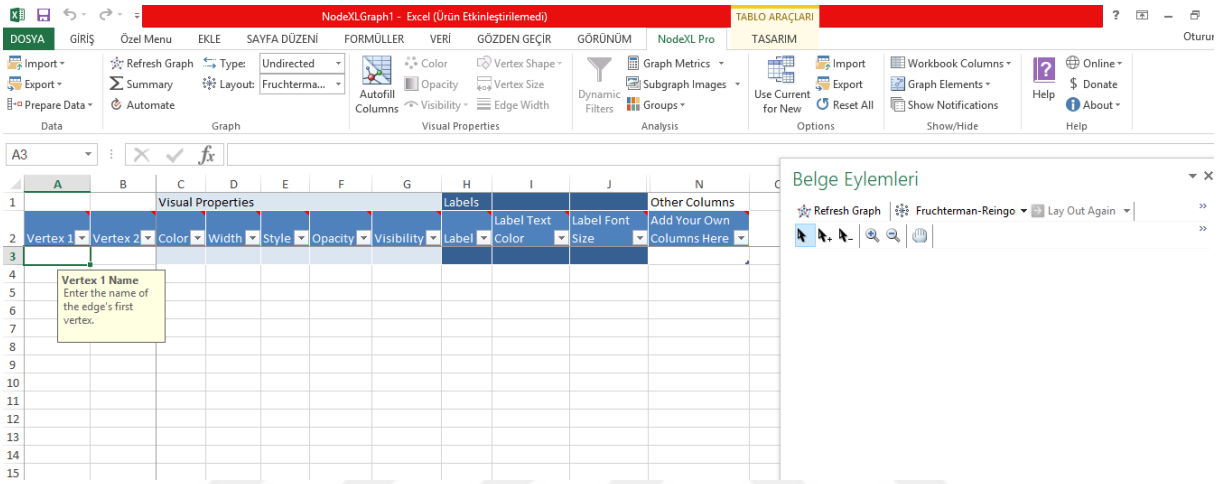
3.5 NodeXL ve Sosyal Ağ Analizi

Sosyal ağ analizi sosyal yapıyı, düğümler ve bu düğüm çiftlerini birbirine bağlayan ilişkiler kümesi şeklindeki bir sosyal ağ olarak algılamaktadır. Düğümler, kişilerden oluşabildiği gibi gruplar, kurumlar ve hatta uluslar da olabilir. İlişkilerden kastedilen ise bu düğümlerin birbirlerine bağımlılığı ve birbirleriyle olan işbirliğidir. Ağ analizi sosyal sistemde yer alan kişiler arasındaki ilişki biçimlerini ortaya koyar ve bu ilişkilerin sosyal yapı içindeki yerleşimlerini ve zaman içindeki değişimlerini inceler. Böylelikle toplulukların yapısı incelenmekte, ağ topluluklar arasında kolayca gözlemlenemeyen ilişkiler görselleştirilerek var olan bağlantılar modellenmeye çalışılmaktadır. Kısacası bireyler arası ilişkilerin sayısallaştırılıp bilimsel hale getirilmesidir. Böylece öngörülemeyen gizli kalmış bilgilerin keşfedilme imkanı doğar. Bu amaçla birçok üniversitenin katkısıyla Microsoft Excel eklentisi olan NodeXL (Network Overview and Discovery Exploration for Excel) geliştirilmiştir. Şekil 3.2 NodeXL ile sosyal ağlarda yapılabilecek analizleri göstermektedir.

Ağ Grafik	Fruchterman-Reingold, Harel-Koren Fast Multiscale, Circle, Spiral, Horizontal Sine Wave, Vertical Sine Wave, Grid, Polar, Polar Absolute, Sugiyama, Random algoritmaları kullanılarak ağın grafiği çizilebilir. Gruplanmalar ve gruplar arasındaki bağlantılara göre ağın grafiği çizilebilir.
Düğüm	Düğümler kare, üçgen, daire, karo, küre sembolleriyle gösterilebildiği gibi etiket isimleri veya resimleri ile de gösterilebilir. Düğümlerin şekli, rengi, boyutu, opaklığı, görünürlüğü gibi özellikleri kenarların belirli niceliklerine göre ayarlanabilir.
Bağlantı	Grafik yönlü veya yönsüz bağlantılarla çizilebilir. Bağlantıların rengi, şekli (kesintisiz, tire, tire nokta, tire nokta nokta), uzunluğu, opaklığı, görünürlüğü ve etiketi bağlantının niceliklerine göre ayarlanabilir.
Filtreleme	Düğümleri istenilen niceliklere göre filtreleyerek karmaşık ağlar sadeleştirilebilinir.
Sayısal Analizler	**Açıklaması**
Betimsel İstatistikler	Düğüm, tek yönlü bağlantı, karşılıklı bağlantı, toplam bağlantı ve döngü sayıları hesaplanabilir.
Ağ İstatistikleri	Yoğunluk, max. diameter, ortalama diameter, modülerlik, minimum-maksimum-ortalama-medyan derece, minimum-maksimum-ortalama-medyan girdi derecesi, minimum-maksimum-ortalama-medyan çıktı derecesi, minimum-maksimum-ortalama-medyan arasındalık, minimum-maksimum-ortalama-medyan yakınlık, minimum-maksimum-ortalama-medyan özvektör, minimum-maksimum-ortalama-medyan pagerank, minimum-maksimum-ortalama-medyan kümelenme katsayısı, minimum-maksimum-ortalama-medyan değerleri hesaplanabilir.
Gruplanma	Düğümlerin niceliklerine göre ağa katkıları, bağlantılı bileşenler, kümelenme ve motiflere göre gruplandırma yapılarak her gruptaki aykırı önemli düğümler belirlenebilir.

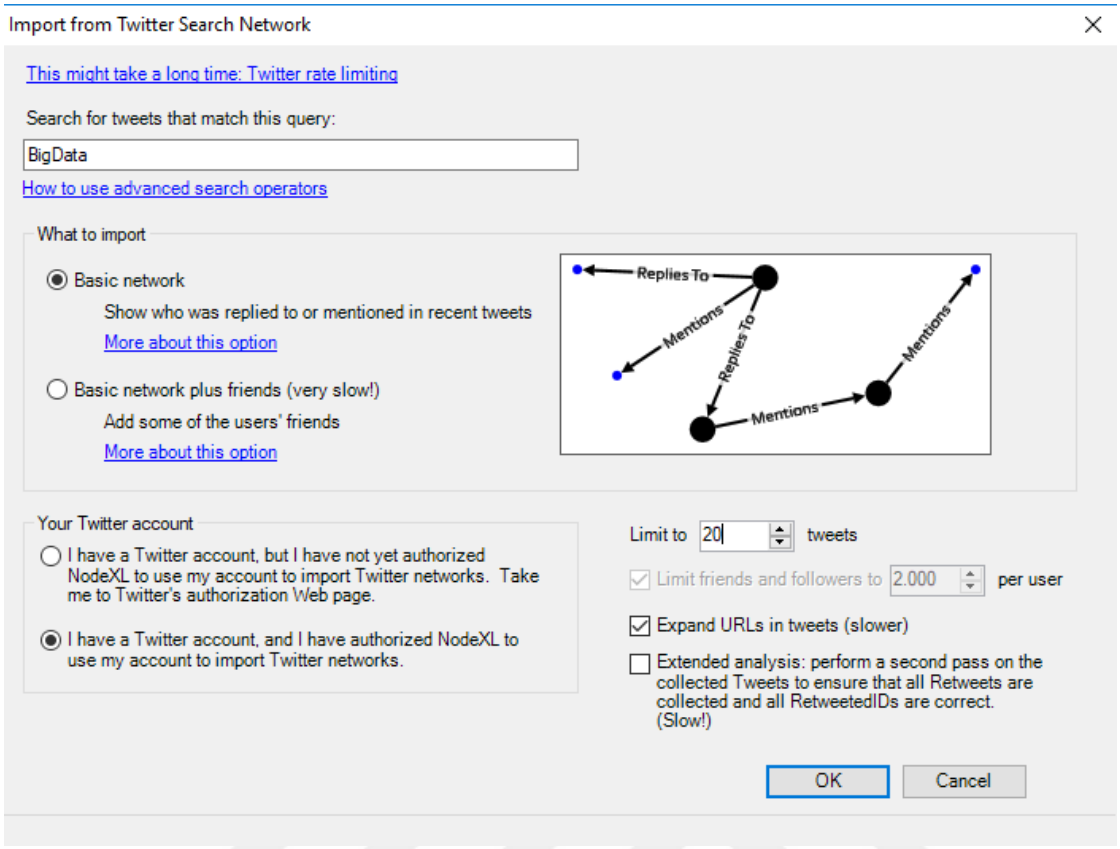
Şekil 3.2 NodeXL ile Yapılabilecek Ağ Analizleri[14]

Altı çalışma sayfası NodeXL'in temel şablonunu oluşturur. Bunlar sırasıyla "Edges (bağlantılar tablosu)", "Vertices (düğümler tablosu)", "Groups (gruplar tablosu)", "Group Vertices (düğüm grupları tablosu)" ve "Overall Metrics (genel ölçüler tablosu)" 'dir. NodeXL bu temelde verileri oluşturduktan veya aldıktan sonra ağı çizip hesaplamalar yaparak verileri işler ve sonuç verir [15]. Şekil.3.3 NodeXL açılış ekranı ve ilgili ara yüzleri göstermektedir.



Şekil 3.3. NodeXL Açılış Ekranı.

NodeXL ile ilişkisel veritabanlarından, csv formatındaki dosyalardan, facebook, twitter, youtube gibi sosyal ağlardan veri almak oldukça kolaydır. Ne tür verilere ya da hangi kullanıcılara göre arama yapılacağı konusunda kullanıcı esnek bırakılmıştır. Şekil. 3.4 BigData Etiketiyle ağdan 20 veri alınmasını göstermektedir.



Şekil 3.4. BigData Etiketiyle 20 Veri Alma.

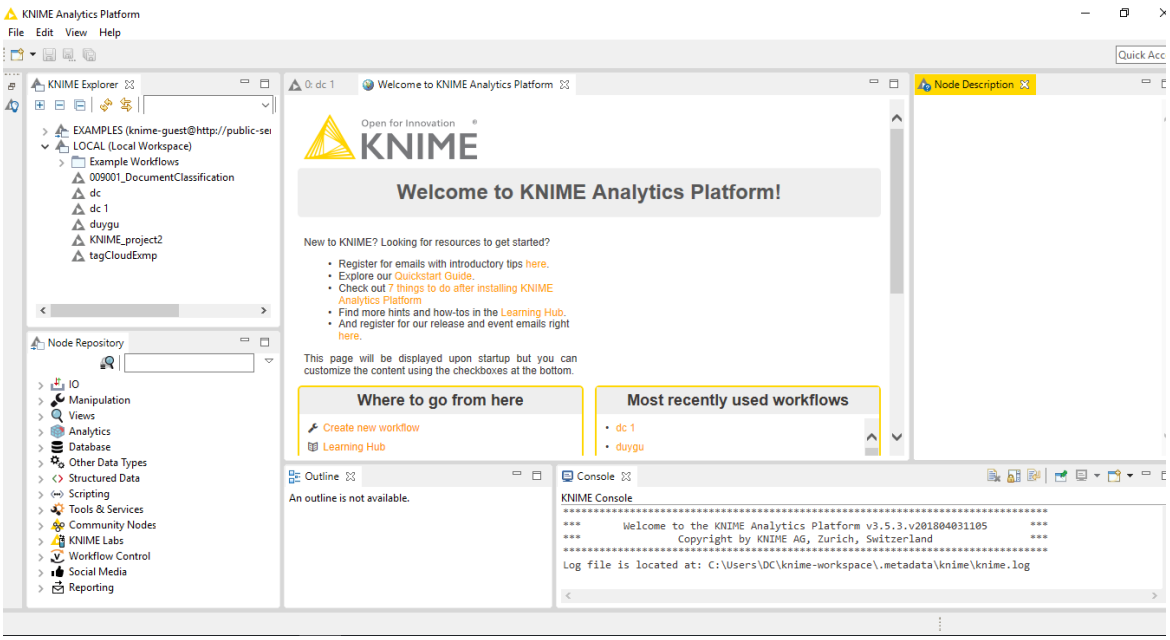
3.6 Knime

KNIME (Konstanz Information Miner**)** 2004 yılında Konstanz'daki bir yazılım ekibi tarafından geliştirilmiştir. Knime işlevselliği artırmak adına eklenti yüklenebilir ve ppt, xls, doc, pdf formatında rapor hazırlanabilir. KNIME, modül olarak ifade edebileceğimiz işlem hattı yapısıyla makine öğrenimi ve veri madenciliği ihtiyaçlarına yönelik bir çok bileşene sahiptir ve bu bileşenler uygulama içerisinde düğüm (node) olarak ifade edilir. Düğümler vasıtasıyla kod yazmadan detaylı işlemler gerçekleştirmek mümkündür. İlişkilendirilen düğümler akış sırasına göre çalıştırılır ve dönüşleri konsol üzerinden takip edilebilir. Ayrıca her düğüme ait çıktı ayrı ayrı görüntülenebilmektedir[17]. Şekil 3.5. Knime açılış ara yüzünü göstermektedir.

- KNIME'ın çekirdek mimarisi, yalnızca mevcut sabit disk alanı ile sınırlı olan büyük verilerin işlenmesine izin verir (diğer açık kaynak veri analiz araçlarının çoğu ana bellekte çalışır ve bu nedenle mevcut RAM ile sınırlıdır). Örneğin. KNIME, 300 milyon müşteri adresinin, 20 milyon hücre görüntüsünün ve 10 milyon moleküler yapıların analiz edilmesini sağlar [16].

- KNIME eklentileri Metin madenciliği, Resim madenciliği ve zaman serileri analizi yöntemlerinin entegrasyonunu sağlar [16].
- KNIME, çeşitli diğer açık kaynak projelerini destekler, örneğin; Weka makina öğrenme algoritmaları, the statistics package R project, LIBSVM, JFreeChart, ImageJ, the Chemistry Development Kit [16].

KNIME bir Java projesi olmasına karşın JavaScript, Python, Perl dillerinde fonksiyon çevirici (wrapper) yazmayı destekler [16].



Şekil 3.5. Knime Açılış Ekranı.

3.7 Sınıflandırma Algoritmaları

Makine Öğrenmesi eldeki verilerden yararlanarak model oluşturan bilgi çıkarımı için kullanılan algoritmalar bütünüdür. Makine öğrenmesi bilgisayarlara, programlama yapılmadan, öğrenme yeteneği sağlayan bir yapay zekâ tekniğidir [18]. Duygu analizi çalışmalarında sınıflandırma konusunda makine öğrenmesi yardımcı elemanlardır. Kendi kendini eğitebilmesi ve gelen yeni verilere göre değişiklik yapabilmesi geliştirilen programlara ciddi avantaj sağlar.

Makine Öğrenmesi denince, akla ilk olarak sınıflandırma ve sınıflandırıcılar gelir. Genel olarak bir sınıflandırma problemi, Makine Öğrenmesi denetimli veya denetimsiz öğrenme algoritmalarıdır. Denetimli sınıflandırma yapılırken öncelikle hangi sınıfa ait oldukları belli, önceden etiketli yeterince büyük bir eğitim kümesinin olması gerekir.

Denetimli sınıflandırma algoritması (Naive Bayes, KDM, Karar Agaçları (KA) vb.) bu eğitim kümesindeki örüntüleri öğrenerek bir model üretir. Artık bu model istenilen etiketsiz örnekleri, eğitim kümesinden öğrendiği örüntülere göre, sınıflandırabilir [18].
Bu tez çalışmasında sınıflandırma algoritmalarından Naive Bayes, Destek Vektör Makinası (DVM) kullanılmıştır. DVM'nin birçok çalışmada yüksek başarı verdiği tespit edilirken, Naive Bayes algoritmasının doğru veri kümeleri için tercih edilmesi durumunda iyi sonuçların saptandığı gözlemlenmiştir.

3.7.1 Naive Bayes Sınıflandırıcı

Naive Bayes sınıflandırıcı istatistiksel tekniklerin kullanıldığı bir yöntemdir. Bayes teoremine dayanarak yapılan işlemlerde olasılıklar hesabı oldukça önem arz etmektedir. Bu teoreme göre ele alınan iki olayın bağdaşmaz olaylar olması gerekmektedir. Eğitim kümesi ile eğitilen naive bayes sınıflandırıcısı, kullanılan özniteliklerin her birinin sınıflarla olan ilişkisini olasılık oranı olarak hesaplar ve o değerleri içeren modeli çıktı olarak verir. Daha sonra naive bayes sınıflandırıcısı, test örneklerini, bu modeldeki öznitelik-sınıf olasılıklarını kullanarak, özniteliklerin bağımsızlığı varsayımıyla sınıflandırır [18].
Bu sınıflandırmayı gerçekleştirebilmek için her bir hipotez için bayes olasılığını hesaplamak gerekse de yüksek sayıda verilerde yeterli öznitelik kullanılarak başarılı olunabilir. Şekil.3.6 hava durumuna göre futbol oynama durumu Naive Bayes algoritmasıyla hesaplanmasını göstermektedir.

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)} \qquad \textbf{(3.3)}$$

3.3 denkleminde A özelliğine sahip nesnenin C sınıfına ait olup olmama durumunu hesaplayan denklem verilmiştir. Herbir özelliğin birincil olasılıkları hesaplandıktan sonra yeni gelen değer 3.4 formülü uygulanarak sonsal değeri bulunur.

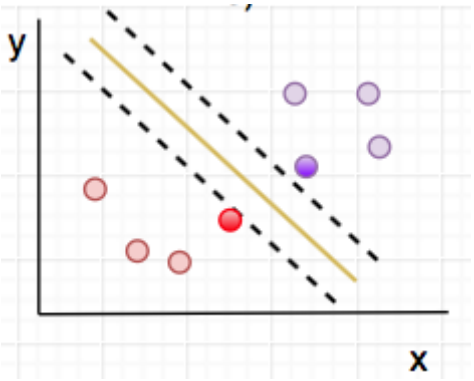
$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C)P(C)}{P(A_1 A_2 \ldots A_n)} \qquad \textbf{(3.4)}$$

Şekil 3.6. Naive Bayes Örneği[22]

3.7.2 Destek Vektör Makinesi

Destek Vektör Makineleri sınıflandırmayı bir doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yerine getirir. Destek Vektör Makinesi yöntemi, veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır [19]. Destek Vektör Makinasında amaç, sınıfları birbirinden ayıracak optimal ayırma hiper düzleminin elde edilmesidir. Başka bir ifadeyle, farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize etmektir [20]. Veri setinin dağılımına göre doğrusal ve doğrusal olmayan olarak iki gruba ayrılmaktadır.

İki gruba ait verilerin doğrusal olarak yayıldığı sistemlerde kullanılır. Eğitim verisi kullanılarak elde edilen hiper düzlemle ayırmak amaçlanır. Veri setini iki sınıfa bölen bu hiper düzlemin optimal olması hedeflenir. Bu hiper düzlemin yeni katılacak verilere karşı dirençli olabilmesi için sınıfların sınır çizgilerine en yakın noktada olması gerekmektedir. Şekil.3.7 doğrusal DVM ile verilerin 2 sınıfa ayrılışını göstermektedir.



Şekil 3.7. Doğrusal SVM.

Kesikli çizgiler sınır düzlemlerini onlara eşit uzaklıkta bulunan ortadaki çizgi hiper düzlemi ifade eder. Hiper düzlemi H_0 olarak belirtirsek H_0 3.5 ya da 3.6 denklemiyle ifade edilir.

$$H_0=W^TX+b=0 \quad \textbf{(3.5)}$$

$$\sum w_i x_i+b=0 \quad \textbf{(3.6)}$$

burada W ağırlık vektörünü, i niteliklerin herbirini göstermektedir. Yani iki boyutlu bir uzayda düzlemler 3.7 ve 3.8' deki gibi belirtilir.

$$H_1= W^TX+b=1 \quad \textbf{(3.7)}$$

$$H_2= W^TX+b=-1 \quad \textbf{(3.8)}$$

Düzlemin üst alanında kalan verileri 3.9 düzlemin altında yer alan verilerin eşitsizliğini 3.10 formülü izah eder.

$$W^TX+b>0,\ y_1=+1 \quad \textbf{(3.9)}$$

$$W^TX+b<0,\ y_1=-1. \quad \textbf{(3.10)}$$

Veri doğrusal bir düzlem olarak ayrılamadığında destek vektör makineleri en iyi şekilde ayıracak çoklu düzlemi bulmaya çalışır. Bu durumda kullanılacak kernel fonksiyonları doğrusal olmayan dönüşümler yaparak ayrım yapma imkanı sağlayacaktır.
Grupların ikiden fazla sınıfa ayrılması durumunda üç farklı yaklaşım ele alınır:

- Problemin ikili gruplara indirgenmesi (bire bir yaklaşımı, one-to-one)
- Problemin tek gruptan bütün gruplara modellenmesi (bire çok yaklaşım , one-to-many)
- çok sınıf sıralama DVM'leri (multiclass ranking SVM)

Sınıflarımız A, B ve C olsun. Bu sınıflar öncelikle ikili gruplar halinde DVM yöntemi ile eğitilerek sınıfların birbirine göre DVM çarpanları çıkarılır. Her iki sınıf için ayrı bir DVM eğitilmiştir. Yeni bir örneğin sınıflandırılması sırasında ikili olarak önce sorgulama yapılır yani AB, BC ve AC ikilileri sorgulanır ve bu sorgulardan hangisine daha çok cevap alınırsa bu sınıfa konulur. Örneğin aşağıdaki sorgular için çıkan sonuçları ele alalım:

AB →A

BC →B

AC → A

Yukarıdaki 3 sorgudan ikisinde A sonucuna ulaşılmıştır dolayısıyla yeni gelen örnek, A sınıfına daha yakın kabul edilmiştir [21].

Bire çok yaklaşımda alınan bir veri geriye kalan tüm verilerle kıyaslanır. Fakat bu yöntem kıyaslama süresinin çok uzun olması ve sonucun bulunamama ihtimali sebebiyle tercih edilmemektedir.

3.7.3 Karar Ağaçları

Sınıflandırma işlemlerinde en sık başvurulan algoritmalardan biridir. Algoritmaya göre ağaç oluşturularak veri tabanındaki özellikler duruma göre eklenir. Ağaç oluşturulurken tercih edilen algoritmaya göre ağacın şekli farklılık gösterebilir. Kök denilen ilk düğümü oluşturan A_i nin farklı olması, en uçtaki yaprak-a ulaşılırken izlenecek yolu ve dolayısıyla sınıflandırmayı da değiştirecektir. Gerek kök düğümün gerekse de bundan sonraki her bir düğümün belirlenmesinde en büyük kriter o noktadan dallara ayrıldığında veri tabanının geri kalan kısmı belli eşit parçalara ayrılmış olsun. Örneğin veri tabanında bulunan evet/ hayır gibiyse iki eşit parçaya, evet/hayır/belki gibi üç değişkenliyse mümkün olduğunca üç eşit parçaya bölünmesi istenmektedir. Burada amaç, en kısa yoldan istenen yanıta ya da sınıfa ulaşmaktır [9]. Tablo.3.2 örnek bir eğitim verisini Tablo.3.3 bu tablodan çıkartılan kuralları göstermektedir.

Tablo 3.1. Eğitim Verisi[19]

Müşteri	**Borç**	**Gelir**	**Risk**
Ali	Yüksek	Yüksek	Kötü
Ayşe	Yüksek	Yüksek	Kötü
Kenan	Yüksek	Düşük	Kötü
Burak	Düşük	Yüksek	İyi
Begüm	Düşük	Düşük	Kötü

Tablo 3.2. Model Kurma[19]

Sınıflayıcı Model
Eğer Borç =Yüksek ise Risk =Kötü;
Eğer Borç= Düşük ve Gelir =Düşük ise Risk=Kötü;
Eğer Borç= Düşük ve Gelir=Yüksek ise Risk=İyi

Veri tabanından rastgele alınan eğitim verileri algoritma uygulanarak bir model oluşturulur. Bu algoritmalar:

- Entropiye Dayalı (ID3,C4.5)
- Sınıflandırma ağaçları (CART)
- Regresyon ağaçları (Twoing, Gini)
- Bellek tabanlı (K-en yakın komşu) şeklindedir.

Eğitim verileri ve uygun algoritmayla oluşturulan model daha sonra test verilerine uygulanır. Yeni eklenen değerin sınıflandırılacak değişkene göre grubu belirlenir.

3.8. Tweetlerdeki Metrik Değerler

Ağ içerisinde yer alan düğüm ve ilişkilerden yararlanılarak belirli algoritmalar doğrultusunda hesaplanan sayısal değerler grup ya da aktörler hakkında bilgi verir. Bu değerlerden biri de merkezilik ölçüsüdür.

Merkezilik, ağlarda düğümlerin ve bağlantıların ne kadar önemli oldukları sorusunu yanıtlamaya çalışır. Bir düğüm veya bir bağlantı ne kadar bilgi akışı yüklenebiliyorsa ne kadar çok sayıda farklı grup arasında köprü görevi görebiliyorsa ne ölçüde önemli düşüncelerin, bilginin, kararların kaynağı olabiliyorsa o derece önemlidir. Merkeziliğin ölçülmesi ile ağda enformasyonun daha hızlı yayılması sağlanabilir, salgın hastalıklar ve ağdaki bozulmalar önlenebilir [15].

Yerel Merkezilik, bireyin yakın çevresi ile ilgilidir. İç derece, dış derece, ortalama derece ve yoğunluk kavramları grafik teori yardımıyla ele alınmaktadır [23]. Ağdan daha çok birey ve yakın çevresini ilgilendiren çıkarımlarda tercih edilir.

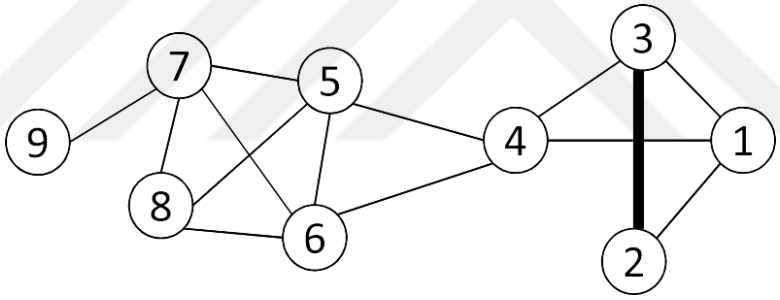
Sosyal ağ bireyleri belirten düğümlerden ve aralarındaki iletişimi belirten kenarlardan oluşmaktadır. Kişiler arasındaki iletişim yönü belirli ise yönlü ağ değil ise yönsüz ağ olarak tanımlanır. Yönlendirilmiş ya da yönlendirilmemiş ağlarda derece aynı şekilde bulunur tek fark yönlendirilmiş ağlarda derecenin iç derece(in-degree) ve dış derece(out-degree) olarak ayrılmasıdır. İç derece gelen bağlantıların sayısını belirtirken dış derece giden bağlantıları gösterir. Düğümler arasındaki karşılıklı bağlantılı olma birliktelik aidiyet durumunu yoğunluk değeri belirtir.

Düğüm ve kenar sayılarını (V,E) şeklinde tanımlarsak yönlendirilmemiş ağlarda yoğunluk 3.11 yönlendirilmiş ağlarda 3.12 deklemleri ile hesaplanır

$$\frac{2*|E|}{(|V|*|V|-1)} \quad \textbf{(3.11)}$$

$$\frac{|E|}{(|V|*|V|-1)} \quad \textbf{(3.12)}$$

Derece Merkeziliği: Bir düğümün komşularıyla tek bir bağla yapmış olduğu bağlantıdır. Yani bir düğümün sahip olduğu link sayısının ölçüsünü belirtir. Çok fazla bilgisine başvurulan ya da popüler birey olup olmadığı konusunda bilgi verir. Mutlak derece merkeziliği kıyaslamalar için uygun olmadığı için normalize edilmiş (0-1) aralığında değerler alan göreceli derece merkeziliği genellikle tercih edilir. Ağdaki i. Düğümün derece merkeziliği 3.13'deki gibi i düğümüyle bağlantısı olan tüm düğüm sayılarının toplanması ile bulunur.



Şekil 3.8. Örnek Ağ[26]

$$C_D(V_i) = di = \sum_j A_{ij} \quad \textbf{(3.13)}$$

$$\text{Normalize Değer} = \frac{di}{n-1} \quad \textbf{(3.14)}$$

Bu formüllere dayanarak Şekil 3.7 de yer alan 7 numaralı düğümün derecesi 4, normalize edilmiş değeri ise 4/(9-1) olarak hesaplanır.

Yakınlık Merkeziliği Yakınlık merkeziliği uzaklığa odaklanır ve dolaylı bağlantı içinde bulunan düğümleri de hesaba katar. Yakınlık i bir düğüm ile çizgedeki diğer bütün düğümler arasındaki en kısa patikaların ortalama uzunluğudur. Erişimin en kısa patikalardan sağlanması koşuluyla yakınlık, ortalama erişim süresi olarak yorumlanabilir [15].

Yakınlık Merkeziliğinde $\sum_i^N d(i,j)$ en kısa yolların toplamını ifade eder. Toplamın tersi alınırsa mutlak yakınlık elde edilir. 3.13 formülü normalize edilmiş sonuçları ifade eder.

$$\frac{[\sum_i^N d(i,j)]-1}{\mathrm{N}-1} \qquad \textbf{(3.15)}$$

Şekil 3.7'deki örnek ağ ele alınıp 3. ve 4. düğümlerin yakınlık merkeziliğini hesaplarsak 4 numaralı düğümün daha merkezi olduğu görülür.

$$\frac{9-1}{1+1+1+2+2+3+3+4} = D_3=0.47$$

$$\frac{9-1}{1+2+1+1+1+2+2+3} = D_4=0.62$$

Özvektör Merkeziliği: Sosyal ağ analizlerinde en önemli bireyi yani önder kişiyi bulmak amacıyla sıklıkla başvurulan algoritmalardan biridir. Özvektör merkeziliği bağların sayısına olduğu kadar niteliğine de bağlıdır,öyle ki bir düğümün az sayıda yüksek kaliteli bağlantısı varsa, çok sayıda daha düşük kalitede bağlantıya sahip düğüme oranla daha fazla katkı sağlar [23]. Google'ın sıralamalarda kullanmış olduğu pagerank algoritması özvektör algoritmasının türevidir.

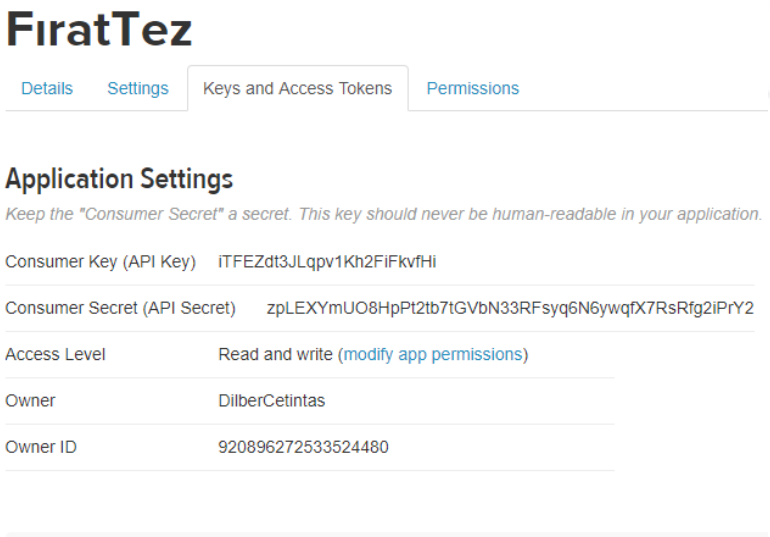
4. YÖNTEM

Bu tez çalışmasında twitterden alınan Türkçe tweetlerin sınıflandırılması için Tweetlerin alınması, Tweetlerin temizlenmesi, Tweetlerin işlenmesi gerçekleştirilmiştir.

4.1. Tweetlerin Alınması

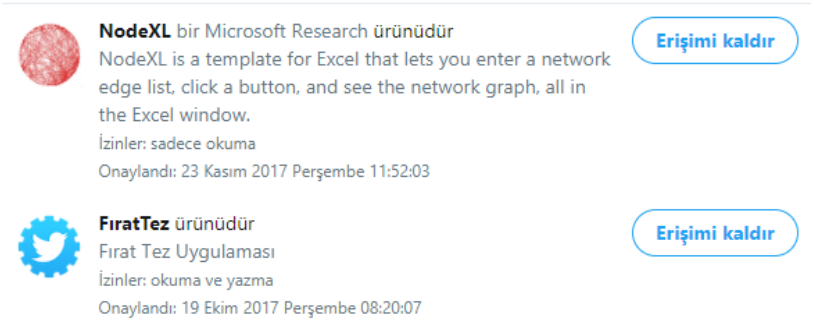
NodeXL farklı veri formatlarındaki dosyaları alıp çizim ve hesaplamalarda kullanabileceği gibi Import özelliği yardımıyla sosyal ağlardan veri alabilmektedir. Bunların yanı sıra Pajek, UCINET gibi sosyal ağ analizi yapan programlardan da dosya alınabilmektedir.

Bu tez çalışması kapsamında ilk adım olarak Türkçe atılan tweetlerin çekilmesi hedeflenmiştir. Verilerin çekilmesi için Twitter verilerine ulaşılmasını sağlayacak Twitter API key leri oluşturulmuştur. Daha sonra NodeXL'in Twitter'a erişimi sağlanmıştır. Şekil.4.1 bu tez çalışması için geliştirilen API'yi göstermektedir.



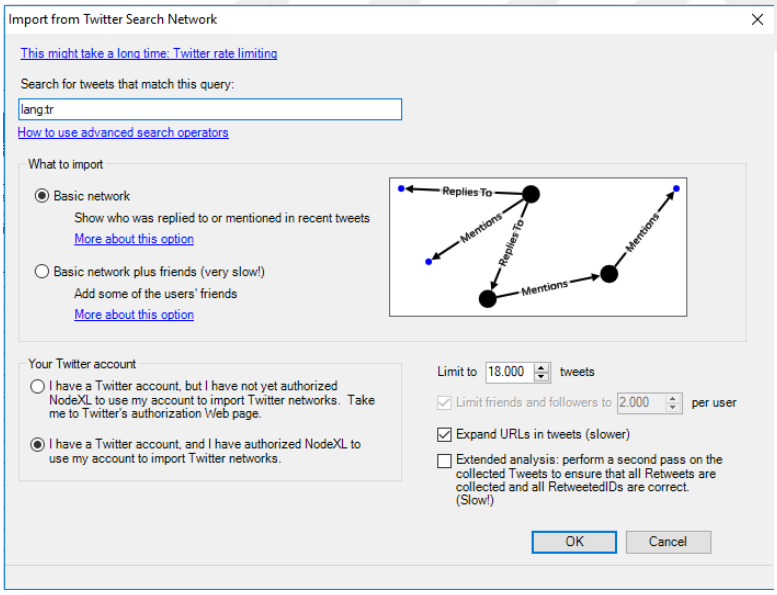
Şekil 4.1. Oluşturulan API.

API oluşturulduktan sonra NodeXL erişimini aktifleştirebilmek için Twitter tarafından gönderilen kod, NodeXL de girilerek Twitter' da yer alan tüm verilere ulaşma imkanı sağlanmıştır. Şekil.4.2 NodeXL'in twitter üzerinden aktifleştirilmesini göstermektedir.

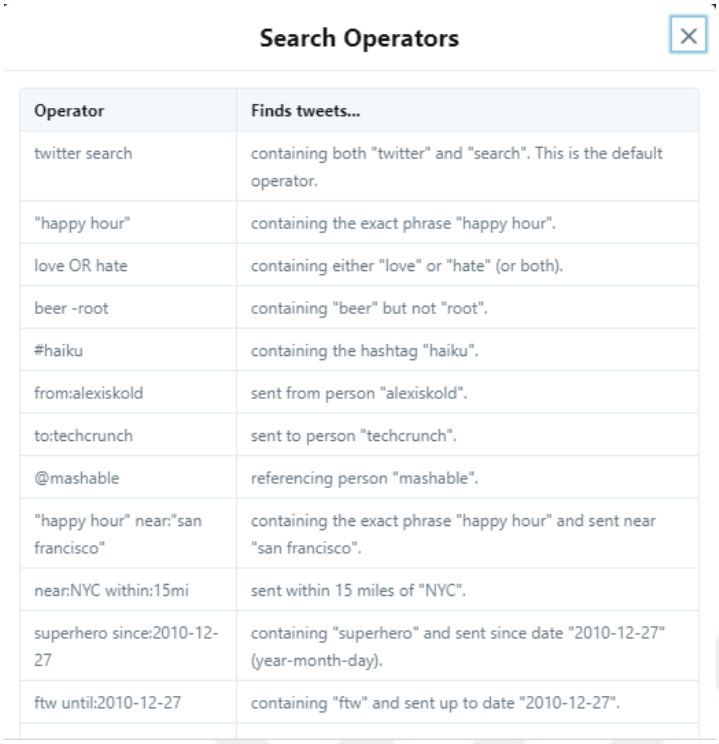


Şekil 4.2. NodeXL'in Twitter Üzerinde Aktifleştirilmesi.

NodeXL ile veriler çekilirken Twitter arama operatörleri kullanılarak kısıtlama yapılabilmektedir. Belirlenen bir düğüm tarafından atılan ya da o düğüme atılan tweetler, aranan kelimeyi içeren tweetler, belirli tarihler arasında atılmış tweetler, son atılan 18000 tweet şeklinde kısıtlama getirilebilmektedir. Bu aramaları yaparken dikkat edilmesi gereken Türkçe karakter girilmemesidir. Çünkü NodeXL Türkçe karakter konusunda destek vermemektedir. Şekil.4.3 Twitter'den Türkçe verileri çekme modülünü Şekil.4.4 arama operatörlerini göstermektedir.



Şekil 4.3. Twitter Türkçe verileri çekme modülü.



Search Operators

Operator	Finds tweets...
twitter search	containing both "twitter" and "search". This is the default operator.
"happy hour"	containing the exact phrase "happy hour".
love OR hate	containing either "love" or "hate" (or both).
beer -root	containing "beer" but not "root".
#haiku	containing the hashtag "haiku".
from:alexiskold	sent from person "alexiskold".
to:techcrunch	sent to person "techcrunch".
@mashable	referencing person "mashable".
"happy hour" near:"san francisco"	containing the exact phrase "happy hour" and sent near "san francisco".
near:NYC within:15mi	sent within 15 miles of "NYC".
superhero since:2010-12-27	containing "superhero" and sent since date "2010-12-27" (year-month-day).
ftw until:2010-12-27	containing "ftw" and sent up to date "2010-12-27".

Şekil 4.4. Arama Operatörleri.

Tablo 4.1. Twitter'dan Çekilen Tweet Örnekleri

No	Tweet
1	Günaydın https://t.co/UZBqeuoz6w
2	RT @Yedikçe zayıflatan süper besinler https://t.co/owrnD36viK https://t.co/pBgJdt93EL
3	Temel Mantık Devreleri https://t.co/Jje7HxVjjl https://t.co/s4jMtYhic0
4	RT @LafTReni: Allah senden razı olsun, Orhan abi.. https://t.co/qfmgfMr2sw
5	3.Olağan kongre sonrası ilk toplantımızı gerçekleştirdik @omrfarukay https://t.co/D8fDlb6a2F
6	RT @neminooo: Gazi Üniversitesi profesöründen YÖK'e mektup: Vazgeçin https://t.co/c80w3t2c0D #gaziyesahipcık #GaziyeDokunma
7	RT @GokHan_Turk6: Türklerin en büyük düşmanı arap dini, arap coğrafyası, araplasmış ve sapıklaşmış halktır. https://t.co/sdZ1063Qxu
8	@blackroseulku94 Günaydın meleğim ❤❤❤
9	RT @HMZZDEMIR: Şükret, geceyi gündüze bağlayana .
10	ne gece ,ne gündüz uyuyamıyorum yaa
11	Bugün Seçim Takvimi Açıklanıyor...Allah'ın izni ve Milletimizin takdiri ile ilk turda işi bitireceğiz ve #ReisiBaşkanYapacağız...
12	RT @universitecim: Yüksek lisans için ortalaman yetiyor mu dediklerinde ben

	https://t.co/H6cd51Aap5
13	RT @Sedaca1903_BJK: Dervişe sordular: " huzura neyle ulaşılabilir " Cevap verdi"Uzaklaşarak " https://t.co/B35BiWvQob
14	Derbinin hakemi belli oldu! https://t.co/MnhPQuILOe
15	Bu antilere de gün doğuyor hemen havlayın havlayın açılırsınız biz hedefimize odağız AdemYaz Gönder1890
16	Günlük Rapor: 5 kişi profilimi gezdi, 2 tweet attım via https://t.co/DQeNGkUGIK
17	Siz sıcacık yatağınızda uyuyun ben sıkıcı bir derste uyuklıyım
18	RT @krprckkaymakam: Kaymakamımız @EnurYmn İnönü Mahallesinde Hanife Vural teyzemizi ziyaret etti. https://t.co/znYIOXySGc
19	RT @themarginale: Konu Abdullah Gül..Tuğrul Selmanoğlu meseleyi muhteşem özetlemiş.. "İnsanda biraz vefa olacak.." https://t.co/mMKmm9Dwco
20	tuylerim diken diken oldu ya HARİKASINIZ HARİKA https://t.co/0Q8bsrdd9M
21	Tavrımızı CHP,den 15 millet vekilinin İYİ Parti'ye katılması değil, bunun yanı sıra Gerçek bir Atatürk,cü CB aday belirlemeleri ve Parlamenter sisteme dönmeleridir. ikinci bir erdoğan vakasına bu milletin tahammülü yoktur. https://t.co/TxKPiMLAn8

4.2. Tweetlerin Temizlenmesi

Twitter'ın kullanıcılara sunduğu esnek yapı sayesinde kullanıcılar "#" ile etiket oluşturup onun altında gruplanabilir ya da attığı tweetlerde "@" işaretiyle bahsettiği kişileri belirtebilir. Tweet içerisinde herhangi bir paylaşımın linkini paylaşarak şahısların ulaşmasını sağlayabilir. Beğenilen tweetlerin kendi sayfalarında yayınlayarak RT yapabilir. Bu durum verilerin gürültülü olmasına ve sınıflandırma sorunlarına sebebiyet vermektedir. Bunun giderilmesine yönelik kullanılan makro ile yapılan iyileştirmeler Tablo 4.2'de verilmiştir.

Tablo 4.2. Makro ile Yapılan İyileştirmeler

@ içeren karakterler	Silindi
# içeren karakterler	Silindi
RT karakterleri	Silindi
http ya da https ile başlayan ifadeler	Silindi
Noktalama İşaretleri	Silindi

Bu işlemlerin gerçekleştirilmesi sonucunda tamamen hashtag ya da tamamen mention içeren ifadeler boş tweet olarak algılandığı için bu satırlar silinmiştir.

Her dilde olduğu gibi Türkçede de bir cümlenin anlamını olumluluk ya da olumsuzluk açısından etkilemeyen durak sözcükler (stop words) vardır. Durak sözcük sözlüğünün içerdiği kelime tiplerine örnek olarak aşağıdakiler verilebilir[8]:

- Sayılar
- Zamirler
- Edatlar

Türkçenin sondan eklemeli bir dil olmasından da kaynaklanan kelimelerin birden fazla ek alarak kullanılması da ayrı bir karmaşıklık oluşturmaktadır. Kelime köküne eklenen çekim ekleri anlamı değiştirmeden farklı kelime olarak algılanmasına neden olmaktadır.

Bir başka problem olan kelimelerin büyük ya da küçük yazılmış olması aynı kelimelerin farklı kelimeler gibi algılanmasına sebebiyet verdiği için knime programı yardımıyla stop words, case sensitive ve kelime eklerinin elimine edilmesi sağlanmıştır. Tablo. 4.1'de verilen Türkçe tweetlerin temizlenmesi sonucunda elde edilen tweetler Tablo. 4.3' te verilmiştir.

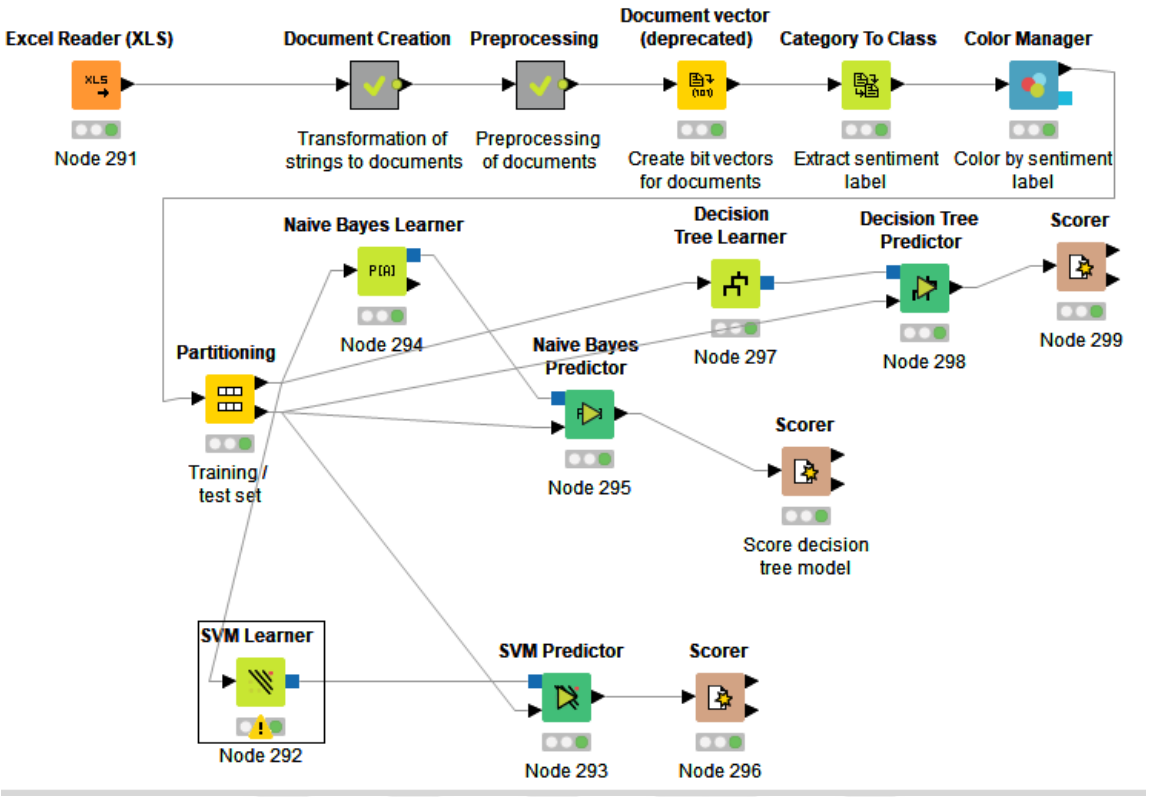
Tablo 4.3. Temizlenmiş Tweet Örnekleri

No	Tweet
1	Günaydın
2	zayıflatan süper besinler
3	Temel Mantık Devreleri
4	Allah senden razı olsun Orhan abi
5	3 Olağan kongre sonrası ilk toplantımızı gerçekleştirdik
6	Gazi Üniversitesi profesöründen YÖK'e mektup Vazgeçin
7	Türklerin en büyük düşmanı arap dini, arap coğrafyası, araplasmış ve sapıklaşmış halktır
8	Günaydın meleğim
9	Şükret geceyi gündüze bağlayana
10	ne gece ne gündüz uyuyamıyorum yaa
11	Bugün Seçim Takvimi Açıklanıyor Allah ın izni ve Milletimizin takdiri ile ilk turda işi bitireceğiz ve
12	Yüksek lisans için ortalaman yetiyor mu dediklerinde ben
13	Dervişe sordular huzura neyle ulaşılabilir Cevap verdi
14	Derbinin hakemi belli oldu

15	Bu antilere de gün doğuyor hemen havlayın havlayın açılırsınız biz hedefimize odağız AdemYaz Gönder1890
16	Günlük Rapor 5 kişi profilimi gezdi 2 tweet attım via
17	Siz sıcacık yatağınızda uyuyun ben sıkıcı bir derste uyuklıyım
18	İnönü Mahallesinde Hanife Vural teyzemizi ziyaret etti
19	Konu Abdullah Gül Tuğrul Selmanoğlu meseleyi muhteşem özetlemiş İnsanda biraz vefa olacak
20	tuylerim diken diken oldu ya HARİKASINIZ HARİKA
21	Tavrımızı CHP, den 15 millet vekilinin İYİ Parti ye katılması değil bunun yanı sıra Gerçek bir Atatürkcü CB aday belirlemeleri ve Parlamenter sisteme dönmeleridir ikinci bir erdoğan vakasına bu milletin tahammülü yoktur
22	Kor olmağdan daha betər şey varsa o da görmə qabiliyyətin olub da ətrafındakı heç nəyi heç kimi görməməkdi

4.3. Tweetlerin İşlenmesi

Bu tez çalışmasında Twitterdan alınan 20000 veri sözlük tabanlı yönteme göre yapılandırılmıştır. Gürültüden arındırılan her bir tweet 5 farklı kullanıcı tarafından olumlu, olumsuz, nötr olarak etiketlenmesi istenmiştir. Anlam kargaşasını önlemek amaçlı 5 farklı kişiye yaptırılmış olup değerlendirilen tweette nicel fazlalığı olan sınıf tercih edilmiştir. Etiketlenen 14471 tweet Knime ile sınıflandırılmıştır. Bu alanda bahsedilen aşamaların Knime ile gerçekleştirilmesi aşağıda adım adım verilmiştir. Şekil 4.5. Analiz için Hazırlanan Knime Programını göstermektedir.

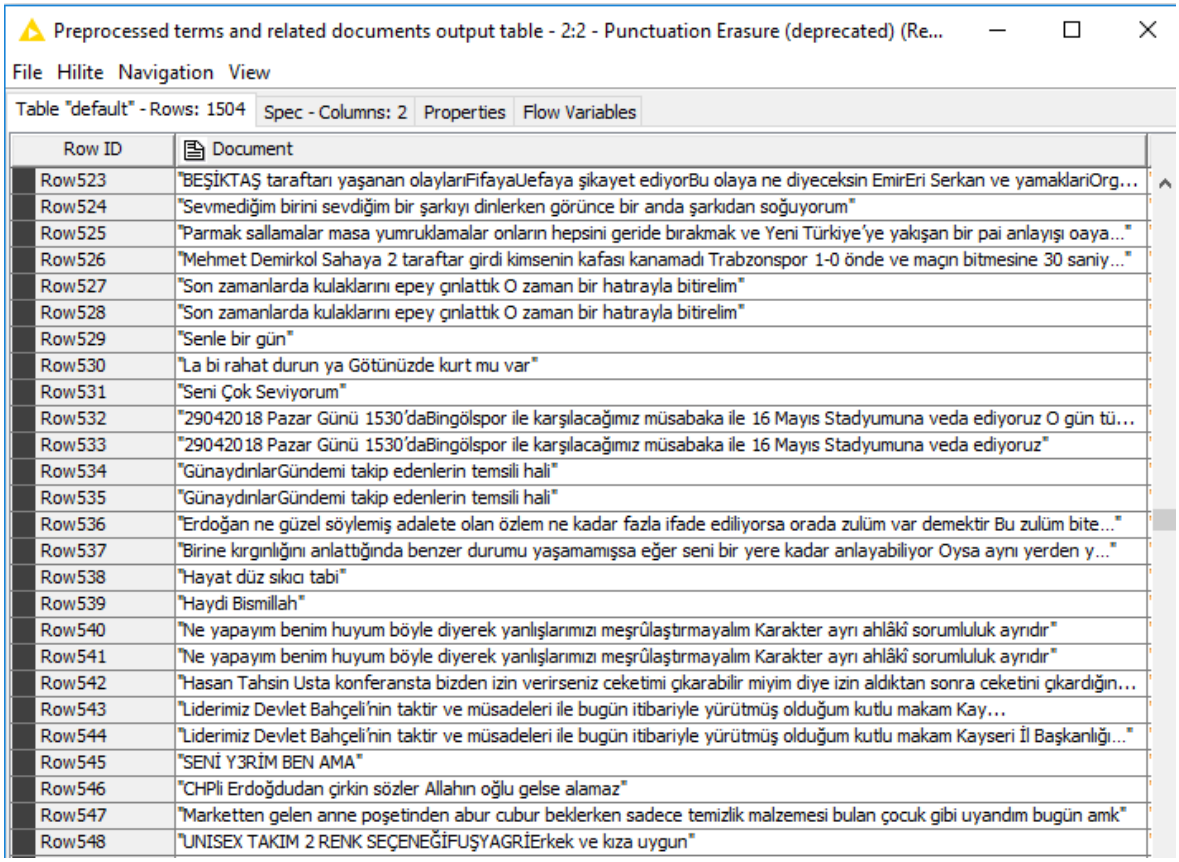


Şekil 4.5. Analiz için Hazırlanan Knime Programı.

Excel Reader(XLS): NodeXL ile import edilen tweetler excel dosyasında tutulduğu için ilk olarak excel dosyasının yüklenmesi gereklidir. Excel reader modülü aldığı verileri okuyarak çıkış portuna iletir. Bu işlemi gerçekleştirirken numeric, date, boolean ve string dataları okuyabilmektedir. Bunun yanı sıra diyagram ve resim içeriklerini okuyamamaktadır. Çok fazla veri içeren belgeler mevcut olduğunda çok fazla zaman ve çok fazla hafıza gerektiği için Apache POI kütüphanesiyle okuma performansı sınırlıdır.

String to Document: Belirtilen kolonların dizelerini belgelere dönüştürür. Her satır için bir belge oluşturulacak ve oluşturulan yeni belge oluşturulan yeni kolona eklenmektedir. Sınıflandırma algoritmalarının giriş olarak string yerine doküman kullanması modülü zorunlu kılmıştır.

Punctuation Erasure: Duygu analiz aşamasında, anlam ifade etmeyen noktalama işaretleri daha önce makro ile temizlenmiş ve işlenebilir veriler elde edilmişti. Punctuation erasure modülü kullanılarak unutulan noktalama işaretlerinin silinmesi ve yeni eklemelerinde kullanılabilmesi amacıyla kullanılmıştır. Şekil.4.6 Noktalama işaretlerinden temizlenmiş tweetleri göstermektedir.

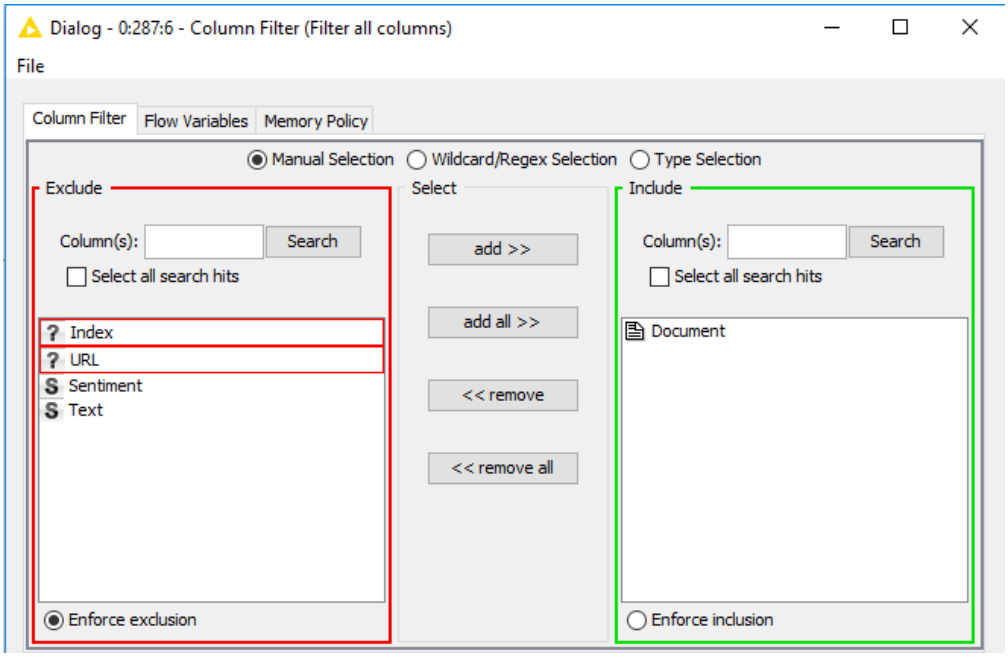


Row ID	Document
Row523	"BEŞİKTAŞ taraftarı yaşanan olaylarıFifayaUefaya şikayet ediyorBu olaya ne diyeceksin EmirEri Serkan ve yamaklariOrg...
Row524	"Sevmediğim birini sevdiğim bir şarkıyı dinlerken görünce bir anda şarkıdan soğuyorum"
Row525	"Parmak sallamalar masa yumruklamalar onların hepsini geride bırakmak ve Yeni Türkiye'ye yakışan bir pai anlayışı oaya..."
Row526	"Mehmet Demirkol Sahaya 2 taraftar girdi kimsenin kafası kanamadı Trabzonspor 1-0 önde ve maçın bitmesine 30 saniy..."
Row527	"Son zamanlarda kulaklarını epey çınlattık O zaman bir hatırayla bitirelim"
Row528	"Son zamanlarda kulaklarını epey çınlattık O zaman bir hatırayla bitirelim"
Row529	"Senle bir gün"
Row530	"La bi rahat durun ya Götünüzde kurt mu var"
Row531	"Seni Çok Seviyorum"
Row532	"29042018 Pazar Günü 1530'daBingölspor ile karşılacağımız müsabaka ile 16 Mayıs Stadyumuna veda ediyoruz O gün tü...
Row533	"29042018 Pazar Günü 1530'daBingölspor ile karşılacağımız müsabaka ile 16 Mayıs Stadyumuna veda ediyoruz"
Row534	"GünaydınlarGündemi takip edenlerin temsili hali"
Row535	"GünaydınlarGündemi takip edenlerin temsili hali"
Row536	"Erdoğan ne güzel söylemiş adalete olan özlem ne kadar fazla ifade ediliyorsa orada zulüm var demektir Bu zulüm bite..."
Row537	"Birine kırgınlığını anlattığında benzer durumu yaşamamışsa eğer seni bir yere kadar anlayabiliyor Oysa aynı yerden y..."
Row538	"Hayat düz sıkıcı tabi"
Row539	"Haydi Bismillah"
Row540	"Ne yapayım benim huyum böyle diyerek yanlışlarımızı meşrûlaştırmayalım Karakter ayrı ahlâkî sorumluluk ayrıdır"
Row541	"Ne yapayım benim huyum böyle diyerek yanlışlarımızı meşrûlaştırmayalım Karakter ayrı ahlâkî sorumluluk ayrıdır"
Row542	"Hasan Tahsin Usta konferansta bizden izin verirseniz ceketimi çıkarabilir miyim diye izin aldıktan sonra ceketini çıkardığın...
Row543	"Liderimiz Devlet Bahçeli'nin taktir ve müsadeleri ile bugün itibariyle yürütmüş olduğum kutlu makam Kay...
Row544	"Liderimiz Devlet Bahçeli'nin taktir ve müsadeleri ile bugün itibariyle yürütmüş olduğum kutlu makam Kayseri İl Başkanlığı..."
Row545	"SENİ Y3RİM BEN AMA"
Row546	"CHPli Erdoğdudan çirkin sözler Allahın oğlu gelse alamaz"
Row547	"Marketten gelen anne poşetinden abur cubur beklerken sadece temizlik malzemesi bulan çocuk gibi uyandım bugün amk"
Row548	"UNISEX TAKIM 2 RENK SEÇENEĞİFUŞYAGRİErkek ve kıza uygun"

Şekil 4.6. Noktalama işaretlerinden temizlenen veri.

Number Filter: Alınan tweet örneklerinde çok sayıda sayısal değer içeren paylaşımların olduğu gözlemlenmiştir. Kullanılan rakamsal ifadelerin duygu betimlemesi açısından anlam ifade etmediği için bu modülle hepsi elimine edilmiştir.

Column Filter: NodeXL ile alınan excel dosyasında birden fazla sütun bulunmaktadır. Analiz aşamasında gerek duyulmayacak içerikleri barındıran kolonları diskalifiye etmek amacıyla eklenmiştir. Böylece karmaşık yapı indirgenerek verim oranı arttırılmaktadır. Şekil. 4.7 Column filter konfigürasyonunu göstermektedir.



Şekil 4.7. Column Filter Konfigürasyonu.

N Chars Filter: Belirli bir karakter sayısından az sayıda karakter içeren kelimeleri kaldırmak için kullanılır. Bu tez çalışmasında yer alan işaret zamirlerinin (bu, şu, o) ve bunların yanısıra "ya", "ya da" ,"ve" , vs. kelimeleri bu modül yardımıyla filtrelenmiştir.

Case Converter: Kullanılan algoritmaların case-sensitive olması aynı kelimelerin farklı yazılmaları durumunda farklı kelime olarak algılayıp sonuç vermemesi için tercih edilmiştir. Tweetler içerisinde yer alan tüm kelimeler küçük harflerle ifade edilmiştir. Şekil.4.8 Tweetlerin küçük harflere dönüştürülmesi sonucundaki tweetleri göstermektedir.

Filtered terms and related documents output table - 2:28 - Case converter (deprecated) (To lower case)

File Hilite Navigation View

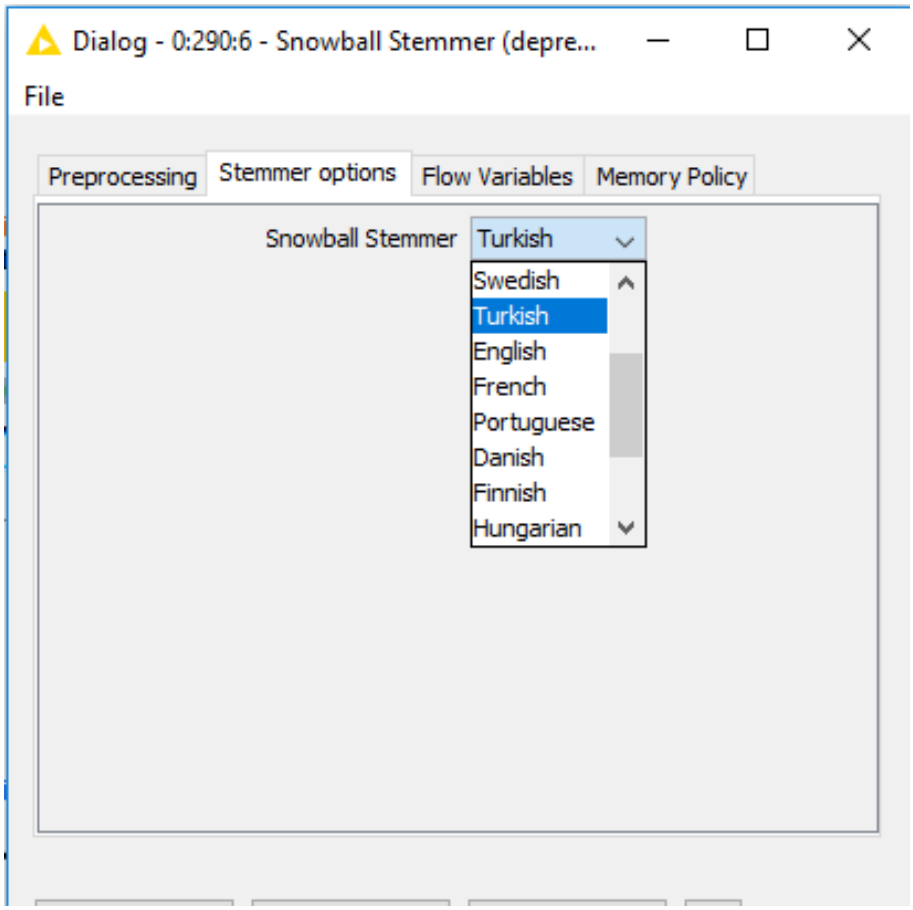
Table "default" - Rows: 1504 Spec - Columns: 2 Properties Flow Variables

Row ID	Document	Orig Document
Row477	"bak sonradan hesap meydanında bu yazı size soruluncabizim haberimiz yoktu hesabımız otomatik beğenm...	"Bak sonradan hesap meydanında bu yazı size sorulunca,bizim haberimiz yoktu hesabımız otomatik beğenmişti deyip beni su...
Row478	"az önce iki yakın arkadaşımın artık sıradan arkadaşlarım olduğunu kararlaştırdım şimdilik haberleri yok"	"az önce iki yakın arkadaşımın artık sıradan arkadaşlarım olduğunu kararlaştırdım şimdilik haberleri yok"
Row479	"ne gece ne gündüz uyuyamıyorum yaa"	"ne gece ,ne gündüz uyuyamıyorum yaa"
Row480	"arkadaşlar"	"ARKADAŞLAR"
Row481	"kayserili bir çobanın oğlu ile dünyanın en büyük imparatorluğunun kraliçesi aynı karede buluşabiliyorsa her...	"Kayserili bir çobanın oğlu ile dünyanın en büyük imparatorluğunun kraliçesi aynı karede buluşabiliyorsa, herkes bunu cumhu..."
Row482	"rahman ve rahim olan allahın adıylaandolsun biz insanoğlunu şerefli kıldıkısrâ"	"Rahman ve Rahim olan Allah'ın adıyla'Andolsun, biz insanoğlunu şerefli kıldık"İsrâ 70'"
Row483	"rahman ve rahim olan allahın adıylaandolsun biz insanoğlunu şerefli kıldık"	"Rahman ve Rahim olan Allah'ın adıyla'Andolsun, biz insanoğlunu şerefli kıldık'"
Row484	"bugün seçim takvimi açıklanıyorallah'ın izni ve milletimizin takdiri ile ilk turda işi bitireceğiz"	"Bugün Seçim Takvimi AçıklanıyorAllah'ın izni ve Milletimizin takdiri ile ilk turda işi bitireceğiz"

Şekil 4.8. Case Converter Sonuçları.

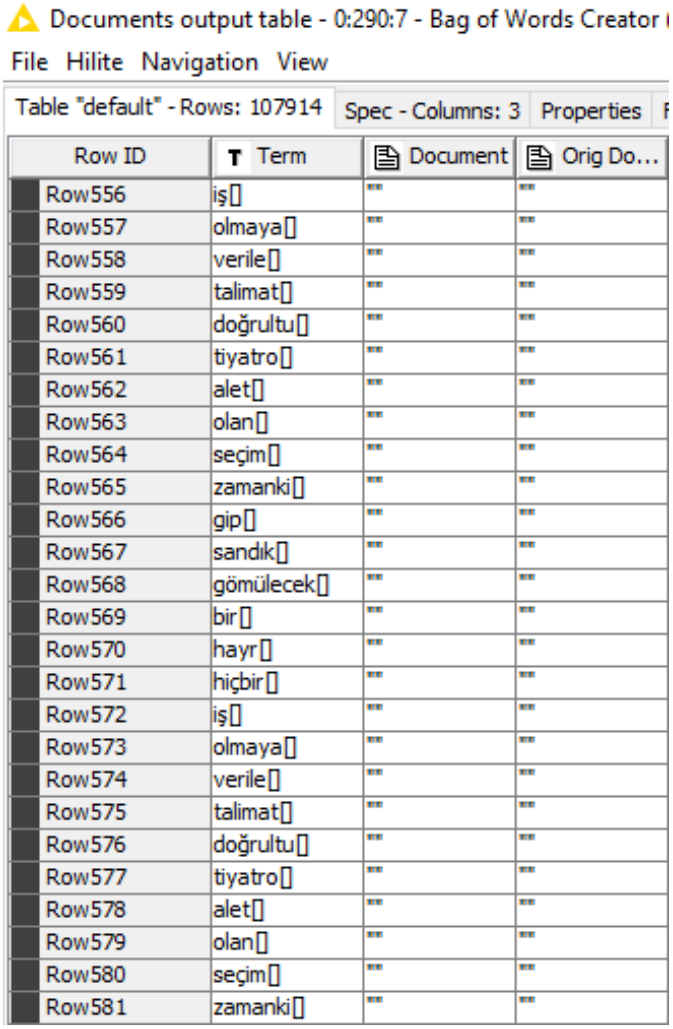
Snowball Stemmer: Bu modül Snowball kütüphanesini kullanarak kelimeleri köklerine ayırmak için tercih edilmektedir. Knime içerisinde yer alan Kuhlen modülü sadece İngilizce kök bulma konusunda destek verdiği için Snowball kullanılmıştır.

Snowball Şekil.4.9'da gösterildiği gibi İngilizce, Almanca ve Fransızca gibi dillere de destek vermektedir.



Şekil 4.9. Snowball Stemmer Desteklediği Diller.

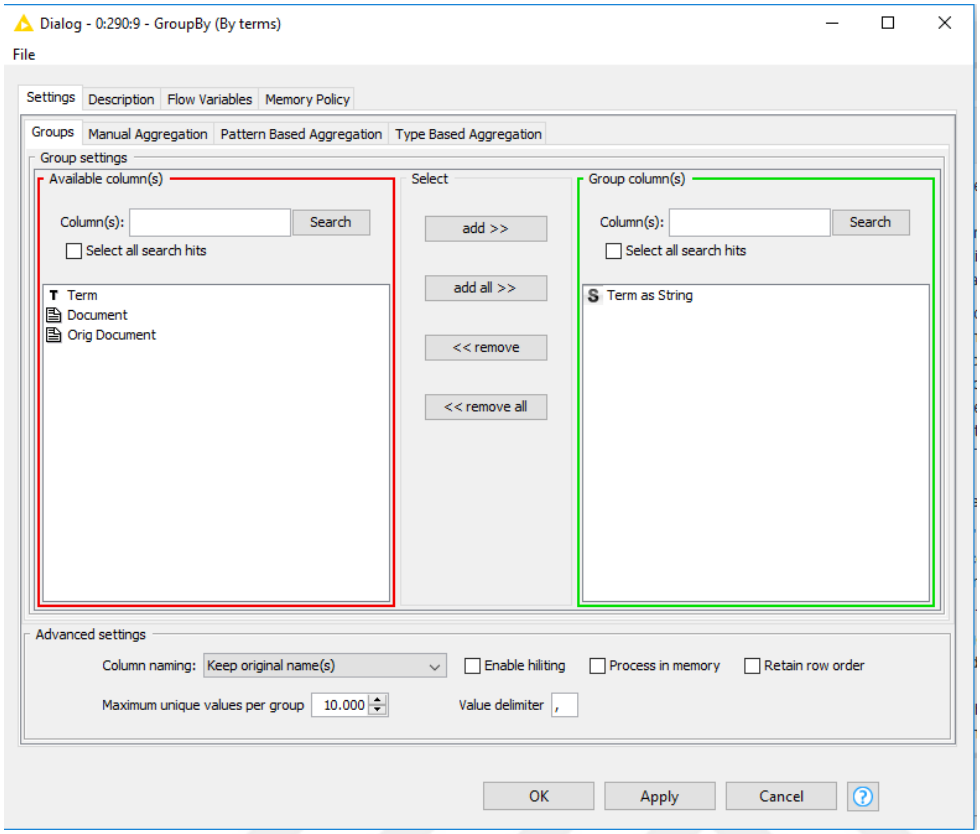
Bag of Words Creator: Snowball Stemmer modülü ile kök haline getirilen bütün kelimeleri toplayarak RowID'lerini oluşturur. Şekil. 4.10 tweetlerden elde edilen kelimelerin bir bölümünü göstermektedir.



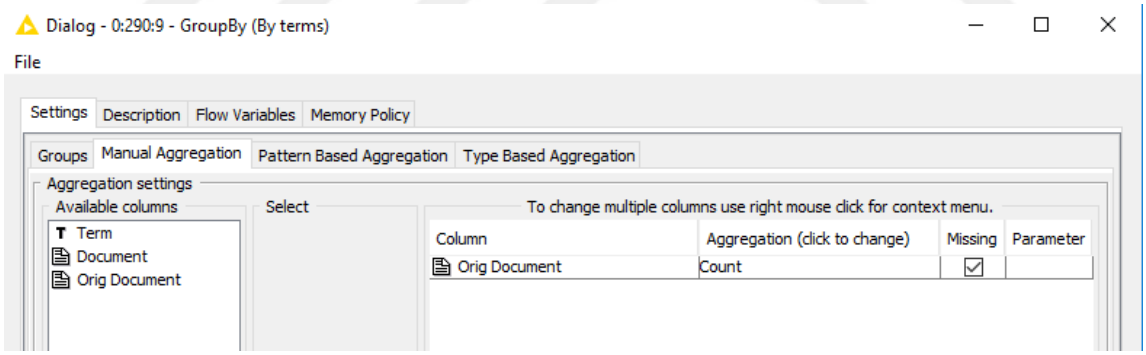
Şekil 4.10. Toplanan Kelimeler .

Term to String: Terimlerin içerdiği tagler atılarak sadece kelime alınır. String olarak alınan kelimeler yeni sütun oluşturularak oraya aktarılır. Bu adımdan sonra yapılacak terim frekansı işlemlerinde string türünde veri girişlerine ihtiyaç duyulacağı için dönüştürme işlemi uygulanmıştır.

Group By: Bir tablonun satırlarını, seçilen grup sütunlarındaki benzersiz değerlerle gruplandırır. Seçilen grup sütununun her benzersiz değeri için bir satır oluşturulur. Kalan sütunlar, belirtilen toplama ayarlarına göre toplanır. Çıkış tablosu, seçilen grup sütunlarının her bir benzersiz değer kombinasyonu için bir satır içerir. Aggregation birden fazla sütunun toplama yöntemini değiştirebilmemize izin verir. Şekil 4.11. Group By Modulü ile sütün seçme ayarlarını, Şekil 4.12. Aggretion (Toplama) Yönteminin Belirlenmesi göstermektedir.

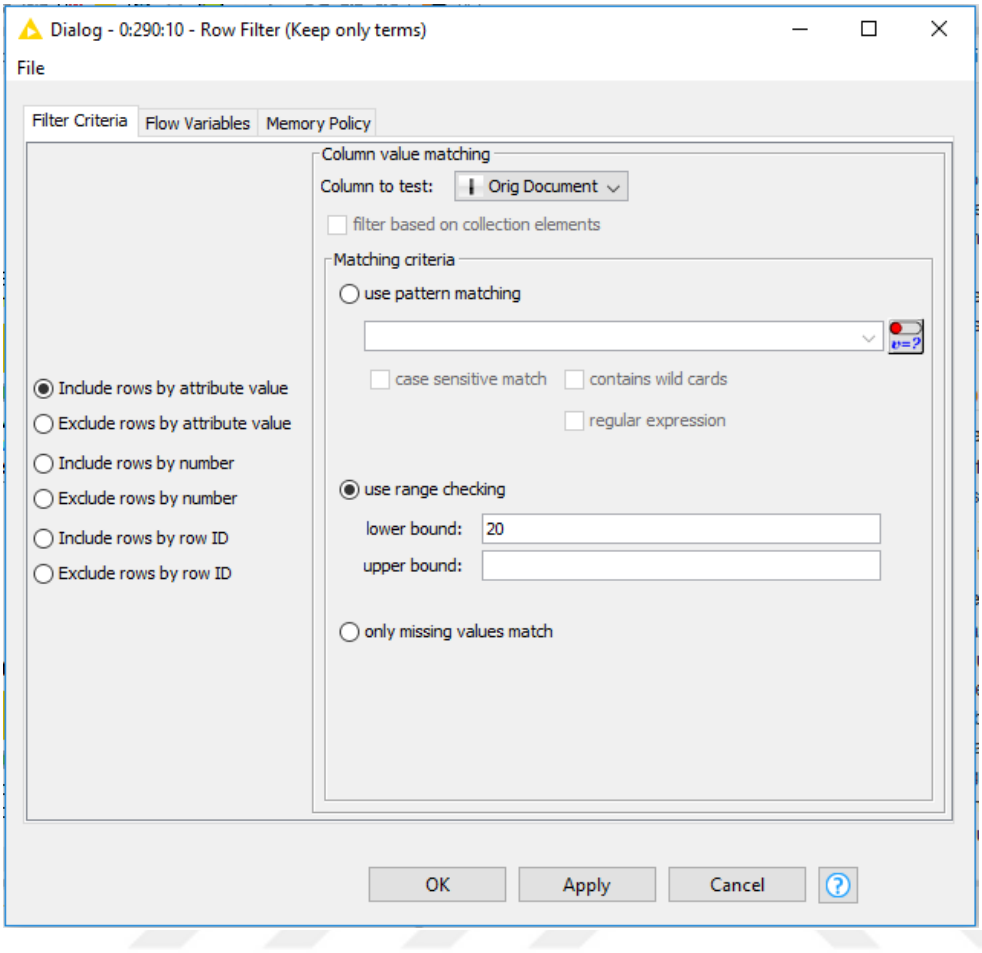


Şekil 4.11. Group By Modulü Sütün Seçme Ayarları.



Şekil 4.12. Aggretion (Toplama) Yönteminin Belirlenmesi.

Row Filter: Belirtilen kritere göre satırın dahil ya da hariç yapılmasına karar verir. Kriterler eşleşme olarak veya belirli bir aralık belirterek seçilebilir. Bu tez çalışmasında daha doğru sonuçlara ulaşabilmek adına 20 defadan az kullanılan kelimeler analize dahil edilmemiştir. Şekil. 4.13 Row Filter Ayarlarını göstermektedir.



Şekil 4.13. Row Filter Ayarları.

Term Frequency: Metin içinde geçen kelimelerin kaç defa geçtiğini hesaplar. Kelimelerin ağırlıklandırılması amacıyla yapılan bu işlemlerde sık kullanılan kelimelerin ağırlığı daha fazla olarak belirlenir. Şekil 4.14. tweetlerdeki bazı kelimelerin hesaplanan TF Değerlerini göstermektedir.

Terms and documents output table - 0:290:12 - TF (term frequency)

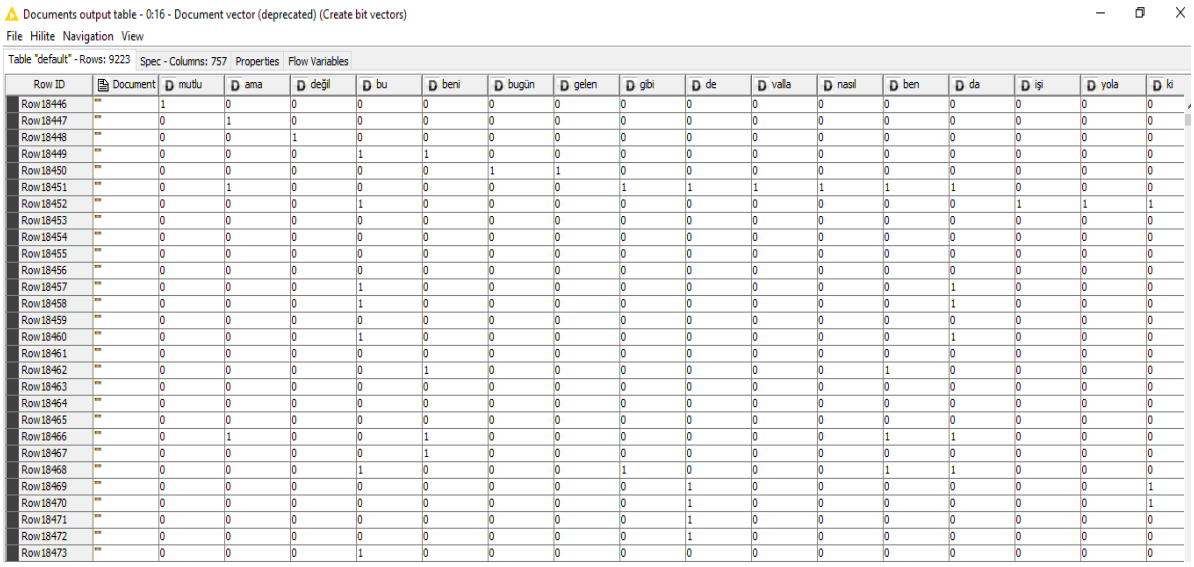
File Hilite Navigation View

Table "default" - Rows: 109599 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Term	Document	Orig Do...	Term a...	TF rel
Row1	çıkarın[]	""	""	çıkarın	0.143
Row2	artık[]	""	""	artık	0.071
Row3	şu[]	""	""	şu	0.143
Row4	bedelliyi[]	""	""	bedelliyi	0.143
Row5	hayatımız[]	""	""	hayatımız	0.143
Row6	mahvoldu[]	""	""	mahvoldu	0.143
Row7	anlasanıza[]	""	""	anlasanıza	0.143
Row8	aık[]	""	""	aık	0.071
Row9	çıkarın[]	""	""	çıkarın	0.143
Row10	artık[]	""	""	artık	0.143
Row11	şu[]	""	""	şu	0.143
Row12	bedelliyi[]	""	""	bedelliyi	0.143
Row13	hayatımız[]	""	""	hayatımız	0.143
Row14	mahvoldu[]	""	""	mahvoldu	0.143
Row15	anlasanıza[]	""	""	anlasanıza	0.143
Row16	çıkarın[]	""	""	çıkarın	0.143
Row17	artık[]	""	""	artık	0.143
Row18	şu[]	""	""	şu	0.143
Row19	bedelliyi[]	""	""	bedelliyi	0.143
Row20	hayatımız[]	""	""	hayatımız	0.143
Row21	mahvoldu[]	""	""	mahvoldu	0.143
Row22	anlasanıza[]	""	""	anlasanıza	0.143
Row23	fettah[]	""	""	fettah	0.067
Row24	medya[]	""	""	medya	0.067
Row25	sponsorluğu...	""	""	sponsorluğu...	0.067
Row26	geliyor[]	""	""	geliyor	0.067
Row27	pop[]	""	""	pop	0.067
Row28	müziğin[]	""	""	müziğin	0.067
Row29	sevilen[]	""	""	sevilen	0.067
Row30	ismi[]	""	""	ismi	0.067
Row31	nisan[]	""	""	nisan	0.067
Row32	cumartesi[]	""	""	cumartesi	0.067
Row33	1500te[]	""	""	1500te	0.067
Row34	carrefoursa[]	""	""	carrefoursa	0.067
Row35	bursa[]	""	""	bursa	0.067
Row36	alışveriş[]	""	""	alışveriş	0.067

Şekil 4.14. TF Değerleri.

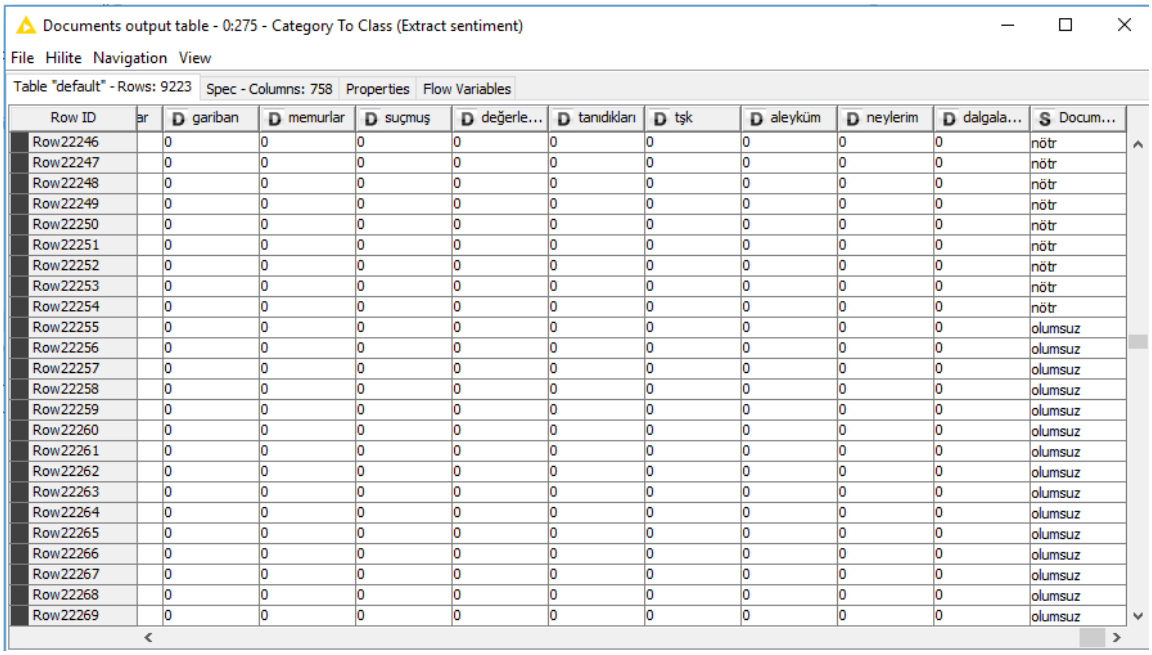
Document Vector: Bu düğüm, her bir belge terimlerini temsil eden bir belge vektörü oluşturur. Özellik vektörlerinin değerleri, boole değerleri olarak veya belirtilen bir sütunun değerleri olarak, yani bir tf * idf sütunu olarak belirtilebilir. Vektörlerin boyutu, Bag of Word'daki farklı terimlerin sayısı olacaktır. Şekil 4.15. Terimlerin vektöre dönüştürülmüş halini göstermektedir.



Row ID	Document	mutlu	ama	değil	bu	beni	bugün	gelen	gibi	de	valla	nasıl	ben	da	işi	yola	ki
Row18446	""	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18447	""	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18448	""	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18449	""	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Row18450	""	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
Row18451	""	0	1	0	0	0	0	0	1	1	1	1	1	1	0	0	0
Row18452	""	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
Row18453	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18454	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18455	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18456	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18457	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18458	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18459	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18460	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18461	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18462	""	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
Row18463	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18464	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18465	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18466	""	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0
Row18467	""	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Row18468	""	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0
Row18469	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Row18470	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Row18471	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Row18472	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Row18473	""	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Şekil 4.15. Terimlerin Vektöre Dönüştürülmesi.

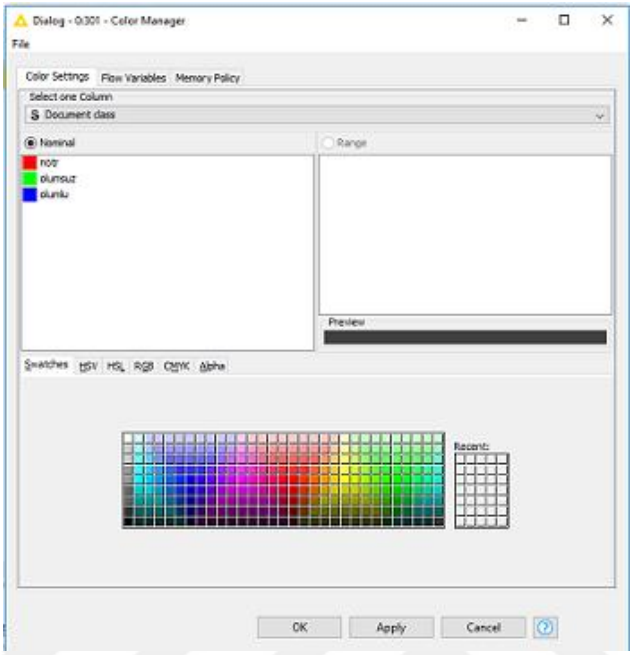
Category to Class: Her satıra bir sınıf sütunu eklenebilmesini sağlar. Sınıfın değeri belgenin dizge olarak kategorisidir. Kategori tanımlanmamışsa, sınıf değeri "tanımsız" olarak ayarlanır. Birden fazla kategori tanımlanmışsa, alfabetik sıralı kategoriler listesinin ilk kategorisi sınıf değeri olarak kullanılır. Şekil 4.16. Satırların nötr, olumlu ve olumsuz olarak kategorilendirilmesini göstermektedir.



Row ID	ar	gariban	memurlar	suçmuş	değerle...	tanıdıkları	tşk	aleyküm	neylerim	dalgala...	Docum...
Row22246		0	0	0	0	0	0	0	0	0	nötr
Row22247		0	0	0	0	0	0	0	0	0	nötr
Row22248		0	0	0	0	0	0	0	0	0	nötr
Row22249		0	0	0	0	0	0	0	0	0	nötr
Row22250		0	0	0	0	0	0	0	0	0	nötr
Row22251		0	0	0	0	0	0	0	0	0	nötr
Row22252		0	0	0	0	0	0	0	0	0	nötr
Row22253		0	0	0	0	0	0	0	0	0	nötr
Row22254		0	0	0	0	0	0	0	0	0	nötr
Row22255		0	0	0	0	0	0	0	0	0	olumsuz
Row22256		0	0	0	0	0	0	0	0	0	olumsuz
Row22257		0	0	0	0	0	0	0	0	0	olumsuz
Row22258		0	0	0	0	0	0	0	0	0	olumsuz
Row22259		0	0	0	0	0	0	0	0	0	olumsuz
Row22260		0	0	0	0	0	0	0	0	0	olumsuz
Row22261		0	0	0	0	0	0	0	0	0	olumsuz
Row22262		0	0	0	0	0	0	0	0	0	olumsuz
Row22263		0	0	0	0	0	0	0	0	0	olumsuz
Row22264		0	0	0	0	0	0	0	0	0	olumsuz
Row22265		0	0	0	0	0	0	0	0	0	olumsuz
Row22266		0	0	0	0	0	0	0	0	0	olumsuz
Row22267		0	0	0	0	0	0	0	0	0	olumsuz
Row22268		0	0	0	0	0	0	0	0	0	olumsuz
Row22269		0	0	0	0	0	0	0	0	0	olumsuz

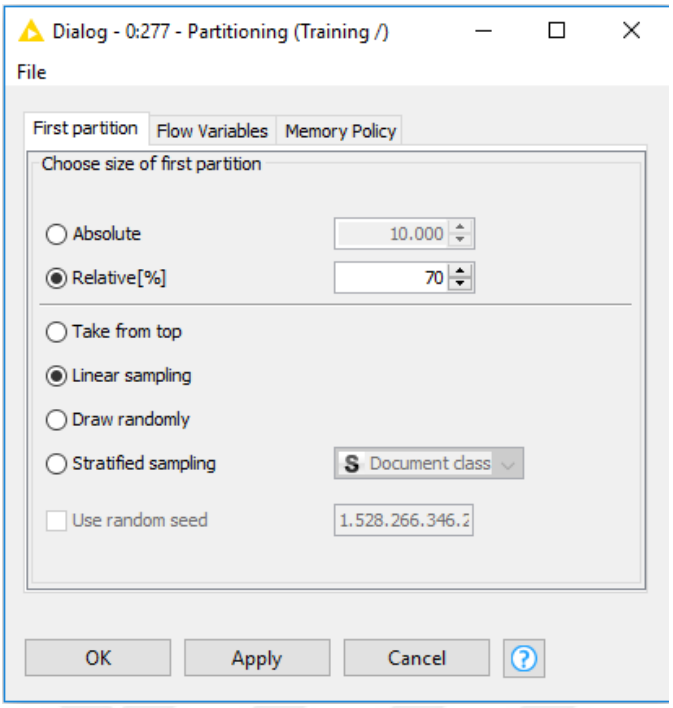
Şekil 4.16. Satırların Kategorilendirilmesi.

Color Manager: Şekil.4.17'de gösterildiği gibi görsel değerlendirme yapabilmek adına renk atamalarının yapılmasını sağlar.

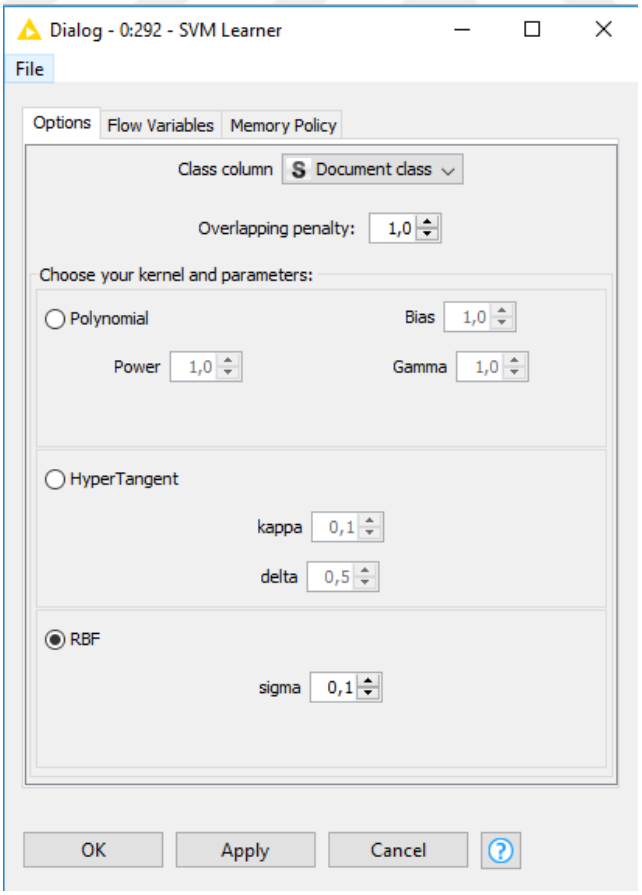


Şekil 4.17. Metin Gruplarına Renk Ataması.

Partitioning: Verilerin eğitim ve test verileri olarak ayırabilmek için tercih edilir. Bu tez çalışmasında %70 eğitim %30 test verisi kullanılmıştır. Bu verilerin seçimi rasgele yapılmıştır. Şekil 4.18. kontrol ve test verilerinin oluşturulmasını göstermektedir. Sınıflandırma için kullanılan DVM algoritmasının parametre seçim ayarları da Şekil.4.18'de verilmiştir.



Şekil 4.18. Kontrol ve Test Verilerinin Oluşturulması.



Şekil 4.19. DVM Modül Ayarları.

Overlapping penalty: giriş verilerinin ayrıştırılmaması durumunda faydalıdır. Sınıflandırılan her bir noktaya ne kadar ceza verildiğini belirler. Bunun için iyi değer 1'dir. Doğrusal olarak ayrılamayan verilerimiz için daha büyük boyuta taşınması amacıyla kullanılan çekirdek fonksiyonlardan RBF fonksiyonu tercih edilmiştir.

Scorer: Decision Learner, Naive Bayes, DVM sınıflandırma algoritmalarının sonuçlarını yansıtabilmek amacıyla kullanılır. Giriş olarak gerçek değerleri çıkış olarak tahmin edilen değerleri yansıtır. Sonuçları aktarırken İnsertion Order yada Lexial Order şeklinde dizer.

5. DEĞERLENDİRME

Literatürde, sınıflandırma başarımlarını belirlemek için Tablo.5.1'de gösterilen karışıklık matrisi (Confusion Matrix) kullanılmaktadır. Karışıklık matrisinde satırlar tweetlerin gerçek sınıflarını sütunlar ise sınıflandırma ile tespit edilen tweet sınıfını göstermektedir. Matriste verilen TP,TN, FP, FN değerleri tweet sayısını göstermektedir.

Tablo.5.1 Karışıklık Matrisi

		Tahmin	
		Pozitif	Negatif
Gerçek	Pozitif	TP	FN
	Negatif	FP	TN

Gerçek pozitif (TP): Doğru pozitif tahmin
Yanlış pozitif (FP): yanlış pozitif tahmin
Doğru negatif (TN): Doğru negatif tahmin
Yanlış negatif (FN): yanlış negatif tahmin
Denklem. 5.1 'de verilen hata oranı (ERR), tüm yanlış tahminlerin sayısı, veri kümesinin toplam sayısına bölünerek hesaplanır. En iyi hata oranı 0.0, en kötü ise 1.0'dır.

$$ERR=(FP+FN))/(TP+FN+FP+TN) \quad \textbf{(5.1)}$$

Denklem. 5.2 'de verilen Doğruluk (ACC), tüm doğru tahminlerin sayısı olarak, veri kümesinin toplam sayısına bölünerek hesaplanır. En iyi doğruluk 1,0, en kötü ise 0,0'dır. Ayrıca 1 - ERR ile hesaplanabilir.

$$ACC=(TP+TN)/(TP+TN+FN+FP) \quad \textbf{(5.2)}$$

Denklem. 5.3 'de verilen Duyarlılık (SN), doğru pozitif tahminlerin sayısı olarak toplam pozitif sayısına bölünür. Ayrıca Recall (REC) veya gerçek pozitif oran (TPR) olarak da adlandırılır. En iyi duyarlılık 1.0, en kötü ise 0.0'dır.

SN=TP/(TP+FN) **(5.3)**

Denklem. 5.4 ‘de verilen Özgünlük (SP), toplam negatif sayıyla bölünen doğru negatif tahminlerin sayısı olarak hesaplanır. Aynı zamanda gerçek negatif oranı (TNR) olarak da adlandırılır. En iyi özgüllük 1.0, en kötü ise 0.0'dır.

SP=TN/TN+FP **(5.4)**

Denklem. 5.5 ‘de verilen Kesinlik (PREC), pozitif tahminlerin toplam sayısına bölünmesiyle doğru pozitif tahminlerin sayısı olarak hesaplanır. Pozitif kestirim değeri (PPV) olarak da adlandırılır. En iyi kesinlik 1.0, en kötü ise 0.0'dır.

PREC=TP/(TP+FP) **(5.5)**

Denklem. 5.6 ‘da verilen Yanlış pozitif oran (FPR), toplam negatif sayıya bölünen yanlış pozitif tahminlerin sayısı olarak hesaplanır. En iyi yanlış pozitif oran 0.0, en kötü ise 1.0'dır. Ayrıca 1 özgüllük olarak hesaplanabilir.

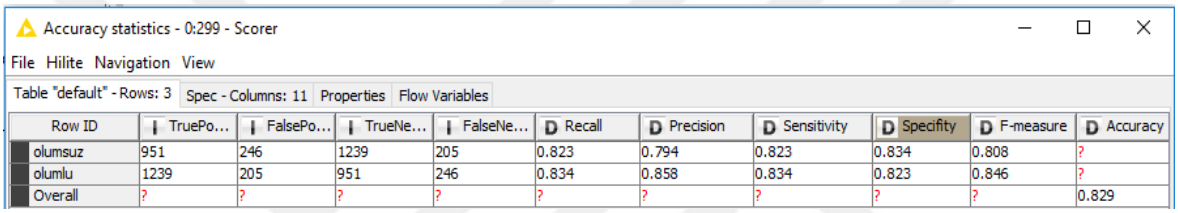
FPR=FP/(TN+FP) **(5.6)**

Bu tez çalışmasında 21/04/2018 ile 26/04/2018 tarihleri arasında yayınlanmış olan Türkçe tweetler ele alındı. Yakın tarihden geriye doğru alınan tweetlerde 26 Nisan tarihli tweet sayısı daha fazladır. Alınan Tweetlerin 18.000 adedi NodeXL aracılığı ile 2.000 adedi Twitter API ve PHP dili yardımıyla toplam 20000 tweet import edildi. Import edilen bu tweetler 12768 düğüm tarafından atıldığı belirlendi.

Tüm tweet ve tweetlere ait özelliklerin tutulduğu excel dosyası gönüllü kişilere yayılıp olumlu, olumsuz, nötr şeklinde anlamsal kutuplama yapmaları istendi. Gönüllülere sunulan 30 günlük zaman diliminde 14471 tane tweet etiketlenmiş bunlardan 4106 adedi olumsuz, 5244 adedi olumlu, geriye kalan 5121 adedi nötr olarak belirtildi. Tweetlerin alındığı tarihin Türkiyede erken seçim ilanının belirtildiği döneme denk gelmesi sebebiyle talep içeren tweetler gönüllüler tarafından nötr olarak etikenlenmiş olup bu durumun nötr sayısını yükselttiği saptandı.

Yinelenen tweetler çıkarıldığında 3592 tweet eksildi. Bunlardan 969 adedi olumsuz, 1350 adedi olumlu, 1276 adedi nötr olduğu saptandı. Bu durum toplum olarak pozitif duyguları daha çok beğenip yaydığımız sonucunu doğurdu.

Sınıflandırma aşamasında ikişerli gruplar halinde alınan(olumlu-olumsuz, olumlu-nötr, olumsuz-nötr) tweetlerin değerlendirilmesi yapılarak herbir etiketin başarıyı etkileme düzeyi araştırıldı. İlk olarak alınan olumlu ve olumsuz veri örneklerinin yer aldığı kümede karar ağacı algoritması ile %82,9 başarı oranı elde edildi. Olumlu verileri doğru tahmin oranı %85,8 olurken olumsuz verilerin doğru tahmin değerinin %79,4'de kaldığı şekil 5.1'de gözlemlendi. Bu veri setini Naive Bayes %73,4, DVM algoritması %80,9 doğruluk oranı ile hesapladı.


Accuracy statistics - 0:299 - Scorer

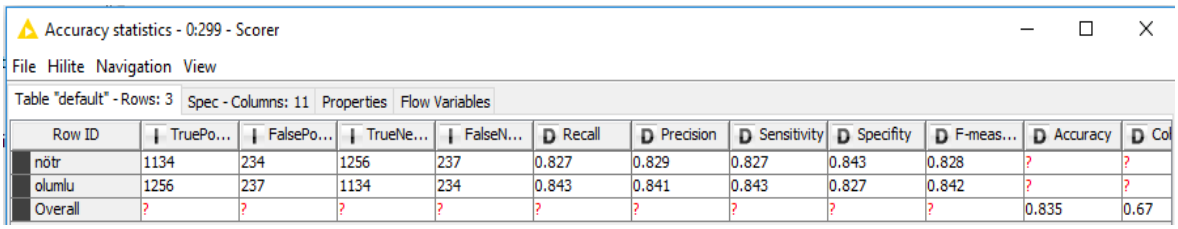
File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
olumsuz	951	246	1239	205	0.823	0.794	0.823	0.834	0.808	?
olumlu	1239	205	951	246	0.834	0.858	0.834	0.823	0.846	?
Overall	?	?	?	?	?	?	?	?	?	0.829

Şekil 5.1. Olumlu-Olumsuz En Yüksek Başarı Sonuçları.

Aynı işlemler olumlu-nötr şeklinde alınan veri setine uygulandığında şekil 5.2'de verilen karar ağacı algoritması ile %83,5 başarı oranına ulaşıldı. Naive Bayes ile %72,8 DVM ile %80,4 değerleri elde edildi. Bu aşamada nötr ve olumlu değerlerin ayrı ayrı tahminlenme başarısının birbirine yakın değerler aldığı saptandı.


Accuracy statistics - 0:299 - Scorer

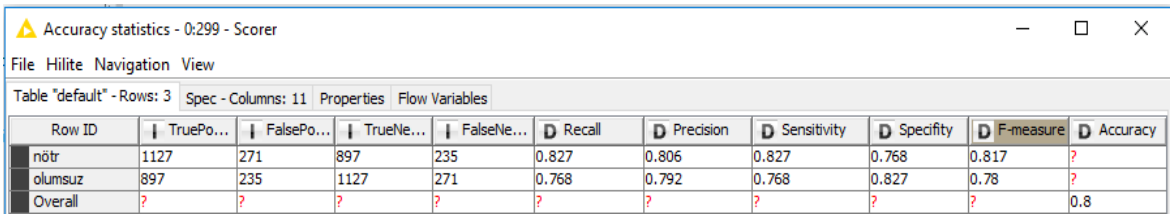
File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Col
nötr	1134	234	1256	237	0.827	0.829	0.827	0.843	0.828	?	?
olumlu	1256	237	1134	234	0.843	0.841	0.843	0.827	0.842	?	?
Overall	?	?	?	?	?	?	?	?	?	0.835	0.67

Şekil 5.2. Olumlu-Nötr En Yüksek Başarı Sonuçları.

Son olarak ele alınan olumsuz-nötr veri setinde yine karar ağacı ile en yüksek başarı değerine ulaşılmış olup %80 başarı sağlandı. Nötr etiketlerin doğru bulunma oranının olumsuz etiketlerden daha yüksek olduğu şekil 5.3'de farkedildi. Bu kategoride Naive Bayes %68,5, DVM %75,5 başarı ile sonuçlandı.


Accuracy statistics - 0:299 - Scorer

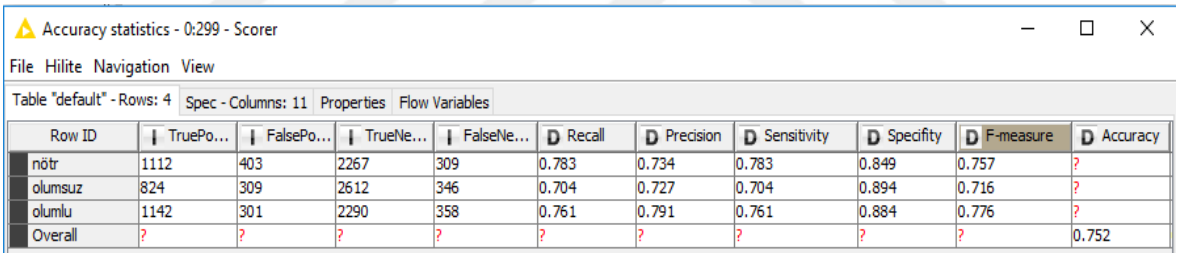
File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	1127	271	897	235	0.827	0.806	0.827	0.768	0.817	?
olumsuz	897	235	1127	271	0.768	0.792	0.768	0.827	0.78	?
Overall	?	?	?	?	?	?	?	?	?	0.8

Şekil 5.3. Olumsuz-Nötr En Yüksek Başarı Sonuçları.

Yapılan ikili kombinasyonların sonucunda en yüksek başarının karar ağacı algoritmasıyla en düşük başarının naive bayes algoritmasıyla elde edildiği gözlemlendi. Seçilen veri setinin ve veri setindeki etiket dağılımlarının doğruluk oranını etkilediği saptandı. Olumlu etiket değerine sahip verilerin yer aldığı sınıflandırmalarda olumlu değerlerin doğru tahmin başarısından kaynaklı doğruluk yüzdelerinin arttığı farkedildi. Olumlu etiketlerde yüksek başarının net ifade kullanımından ve toplumun olumlu verileri tekrarlayarak kullanılan kelime ağırlıklarını artırmasından kaynaklandığı; yinelemenin az olması ve mecazi anlatımların, imaların yoğun olarak kullanılması olumsuz etiketlerde başarının en düşük çıkmasında etkili olduğu sonucuna varıldı.

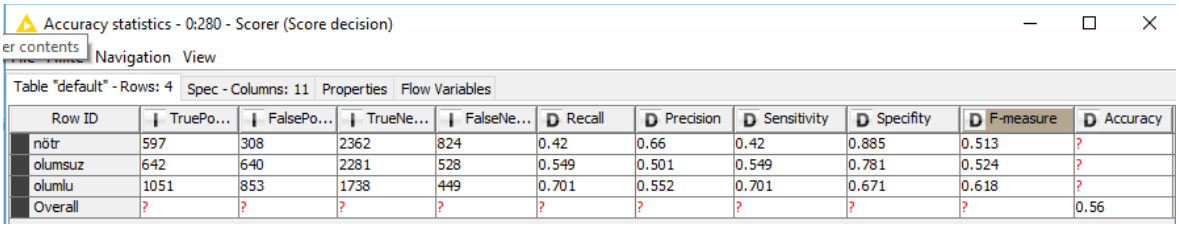

Accuracy statistics - 0:299 - Scorer

File Hilite Navigation View

Table "default" - Rows: 4 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	1112	403	2267	309	0.783	0.734	0.783	0.849	0.757	?
olumsuz	824	309	2612	346	0.704	0.727	0.704	0.894	0.716	?
olumlu	1142	301	2290	358	0.761	0.791	0.761	0.884	0.776	?
Overall	?	?	?	?	?	?	?	?	?	0.752

Şekil 5.4. Decision Tree Sonuçları.


Accuracy statistics - 0:280 - Scorer (Score decision)

er contents Navigation View

Table "default" - Rows: 4 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	597	308	2362	824	0.42	0.66	0.42	0.885	0.513	?
olumsuz	642	640	2281	528	0.549	0.501	0.549	0.781	0.524	?
olumlu	1051	853	1738	449	0.701	0.552	0.701	0.671	0.618	?
Overall	?	?	?	?	?	?	?	?	?	0.56

Şekil 5.5. Naive Bayes Sonuçları.

Tüm veri seti sınıflandırma algoritmalarına tabi tutulduğunda en yüksek başarı Şekil 5.4' de verildiği gibi decision tree ile %75,2 oranında, en düşük başarı şekil 5.5 de görüntülenen naive bayes ile %56 oranında elde edildi. Değişken grubunun artmasının başarıyı olumsuz yönde etkilediği farkedildi.

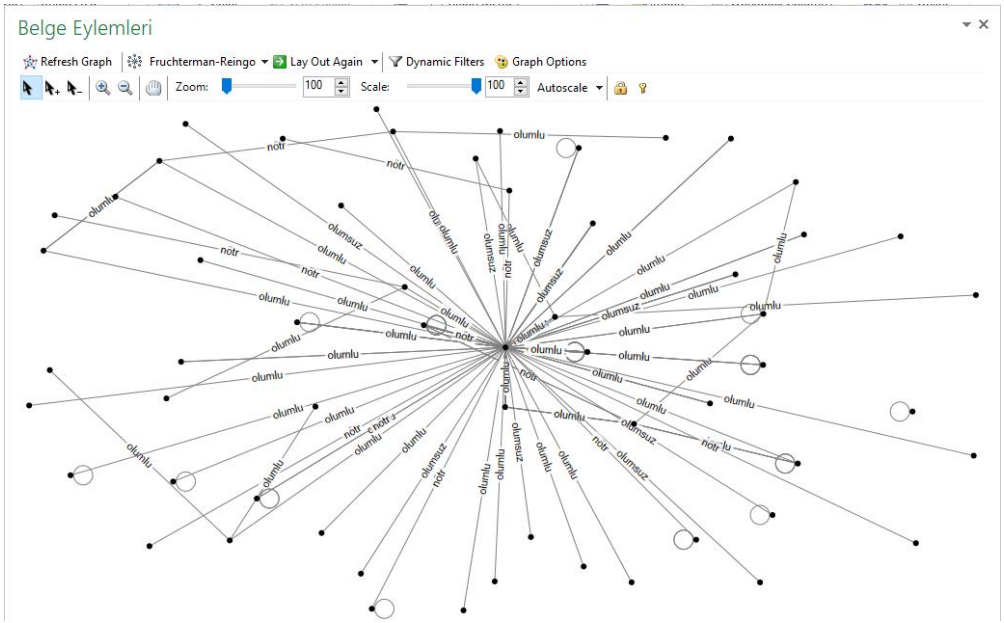
Sosyal Ağ İlişkilerinin Anlamsal Kutupluk Üzerindeki Etkisi

Sosyal ağda yer alan kullanıcıların ağ içerisindeki konumunun anlamsal polarizasyona etkisini bulabilmek amacıyla ilk olarak NodeXL de yer alan verilerin metrik değerleri hesaplandı. Şekil.5.6 tweetleri değerlendirilen bazı düğümlerin metrik değerlerini göstermektedir.

In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	Clustering Coefficient	Followed	Followers	Tweets	Favorites
1	10	93203,377	0,000	0,003	0,009	3043	1322	23683	23033
1	0	0,000	0,000	0,000	0,000	1204	49716	8948	15754
1	0	0,000	0,000	0,000	0,000	257	187403	1087	331
1	0	0,000	0,000	0,000	0,000	1333	244234	5740	3522
1	3	10,000	0,250	0,000	0,083	183	48087	19512	567
1	0	0,000	0,143	0,000	0,000	113	1138	2851	1216
1	0	0,000	0,143	0,000	0,000	1053	30987	3735	2566
1	6	52794,596	0,000	0,000	0,000	3807	3555	17905	18745
1	0	0,000	0,000	0,000	0,000	10410	11506	130085	128501
1	0	0,000	0,000	0,000	0,000	7065	46813	69506	84903
1	0	0,000	0,000	0,000	0,000	4617	4760	29776	29518
1	6	22302,886	0,000	0,003	0,024	904	1232	14372	530
1	0	0,000	0,000	0,000	0,000	459	70263	7956	824
1	2	6,000	0,333	0,000	0,000	115	1311	876	146
1	0	0,000	0,200	0,000	0,000	179	5896	2412	558
1	0	0,000	0,200	0,000	0,000	1439	7409	15316	22182
1	8	157585,281	0,000	0,005	0,069	125	198	768	572
1	0	0,000	0,000	0,000	0,000	9	20673	459	0
1	9	129768,385	0,000	0,002	0,000	702	2203	13728	16727
1	0	0,000	0,000	0,000	0,000	1732	4495	5299	799

Şekil 5.6. Sosyal Ağdaki Metrik Değerler.

Kullanıcıların ağdaki konumu, etkinlik değerleri, takipçi ve takip eden vs. sayılarının yer aldığı tablodan yararlanarak gruptaki önder kişinin(Eigenvector Centrality, Out Degree) diğerleri üzerine etkisi incelendi.



Şekil 5.7. En Yüksek Out Degree Değerine Sahip Kullanıcı.

Şekil 5.7'de en yüksek out degree değerine sahip kullanıcı ele alındığında ağırlıklı olarak olumlu ya da nötr tweetler paylaştığı farkedildi. Bunun ağ üzerindeki etkisi sorgulanıp birinci derece alt grafları incelendiğinde onların da genellikle olumlu veya nötr paylaşımlar yaptığı saptandı. Hatta lider durumdaki kullanıcının tweetlerinin alt-graflar tarafından aynen paylaşıldığı gözlemlendi. Bu durum bize herhangi bir ağ hakkında fikir sahibi olunabilmesi için out degree değeri yüksek kullanıcıların değerlendirilmesi ile olabileceğini gösterdi.

Seçim dönemine rastlayan tweetler alındığı için in-degree, eigen vectör ve betwenness centrality değerlerinde en yüksek değer dönemin cumhurbaşkanı Recep Tayyip Erdoğan'a ait olduğu saptandı. Sonraki dereceler de diğer parti yöneticileri yer aldı. Bu durum gönüllüler tarafından nötr olarak etiketlenen talep tweetlerinin ne kadar fazla sayıda olduğunu bir kez daha kanıtlamış oldu.

6. SONUÇ

Bu tez çalışmasında günümüzde çok fazla ilgi duyulan doğal dil işleme alanında Türkçe dilinin işlenmesine yönelik bir sistem geliştirilmiştir. Bu sistemin gerçekçi olabilmesi için veriler Twitter kullanıcılarının yayınlamış olduğu metinlerden alınmıştır.

Sistem geliştirilirken Türkçenin sondan eklemeli bir dil olduğu göz önüne alınarak kelimenin köklerine ayrıştırılması sağlanmıştır. Köklerine ayrılan kelimelerden en sık kullanılan kelime daha anlamlıdır mantığıyla kelime sayısından yararlanarak metinler vektörlere dönüştürülmüştür. Her kelime anlam ifade etmeyeceğinden bazı kelimeler elenip 20 sayısının altında kullanılan kelimeler de sınıflandırmaya dahil edilmemiştir.

Sınıflandırma algoritmalarının sonucu işlenecek veri kümesinin ve verilerdeki etiket dağılımının başarı üzerinde ciddi rol oynadığı farkedilmiştir. Olumlu, olumsuz, nötr etiketlerinin dengeli dağılması durumlarında başarı oranının arttığı gözlemlenmiştir. Özellikle nötr etiketinin karar aşamasında etkili olduğu anlaşılmış olup nötr sayının azalmasıyla belirsizliğin azaldığı ve başarının arttığı saptanmıştır. Bu tez çalışmasına katılan gönüllülerin borsa terimlerini içeren ve siyasi görüşlerini bildiren her veriyi nötr olarak etiketlemesi başarı oranını zedelemiştir.

Olumlu etiketine sahip verilerde başarı oranı hep en yüksek olup %80'in altına hiç düşmemiştir. Bu durum geliştirilen sistemin net ifadelerde daha doğru tahminlerde bulunup mecazi anlatım içeren olumsuz ifadelerde aynı başarıyı gösteremediğini kanıtlamıştır.

Başarı durumunu etkileyen bir diğer etmen sadece sözlük tabanlı yöntemin kullanılmış olmasıdır. Örneğin; kullanıcıların ingilizce klavye kullanması durumunda "sınıf" kelimesi "sinif" olarak yazılıp sistem tarafından farklı iki kelime olarak algılanması kelime ağırlıklarını değiştirerek başarı oranını düşürmüştür.

Anlamsal kutuplaşmada sosyal ağ ilişkilerinin etkisi ölçülmeye çalışılmış ağdaki önder kişilerin yaymış olduğu pozitiflik ya da negatiflik alt çizgileride etkilediği gözlemlenmiştir. Bana arkadaşını söyle sana kim olduğunu söyleyeyim atasözü birkez daha doğrulanmış olup kişi faktörünün ağdaki etiket sayılarını değiştirdiği belirlenmiştir.

Yapılan benzer çalışmalarda hakaret içeren verileri bulma, mutlu ya da üzgün ayrımına gitme gibi daha keskin ayrımlar içeren veye sözlük tabanlı ile n-gram yöntemini birlikte kullanma gibi çalışmalar yapılarak elde edilmiş başarı oranı ile yaklaşık aynı değerlerde başarı elde edilmiştir.

Tezde konu alanı belirlenerek ya da belirli hashtagler kullanılarak yapılan çalışmalardan farklı olarak hiçbir kısıtlama olmaksızın türkçe tüm veriler alınmıştır.Bu durum başarı oranını negatif yönde etkilemiştir.

Yaygın olarak metin madenciliği alanında kullanılan doğal dil işleme çalışmalarında Türkçenin eksik olduğu sözlük-anlam çalışmalarında katkı sağlanmaya çalışılmıştır.

Dil birçok bilimin temelinde yer alan bir kavram olduğu için yapılan çalışma ile her alanda yapılacak çalışmaya kavram-anlam ilişkisinde olumlu destek verecektir.

Alınan olumsuz veriler küfür, istismar vs. şeklinde derecelendirilerek güvenlik alanında yapılacak çalışma için önemli rol oynayacaktır.

KAYNAKÇA

[1] **Güler**, **E.**, **Mutlu**, **M. E.**, Akademik Personelin Akademik Sosyal Ağları Kullanım Düzeyi, Journal of Research in Education and Teaching, **2**, 72-76.

[2] https://www.frmtr.com/sosyal-ag-haberleri/7041152-sosyal-medyanin-toplum-uzerindeki-etkisi.html, Sosyal Medyanın Toplum Üzerindeki Etkisi, 03/03/2017.

[3] **Akarsu**, **C.** , 2016. Twitter Verileri ile Türk Televizyonları İzlenme Oranı Sıralamaları Tahmini, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[4] **Yıldırım**, **S.**, 2018. Twitter Verileriyle Duygu Analizi ve Türkçe Duygu Kütüphanesi, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[5] **Çoban**, **Ö.** , 2016. Metin Sınıflandırma Teknikleri ile Türkçe Twitter Duygu Analizi, Yüksek Lisans Tezi, Atatürk Üniversitesi Fen Bilimleri Enstitüsü, Erzurum.

[6] **Erçolak**, **N.**, 2017. Twitter'da Türkçe Veriler Üzerinde Hakaret Suçu Analizi, Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

[7] **Adak Kaplan**, **B.** , 2016. Twitter Üzerindeki Türkçe Mesajlarda Veri Madenciliği ile Duygu Analizi, Yüksek Lisans Tezi, Beykent Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[8] **Meral**, **M.**, 2014. Twitter Verilerini Anlamsal Sınıflandırma, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[9] **Silahtaroğlu**, **G.**, 2013. Veri Madenciliği Kavram ve Algoritmaları, Papatya Yayıncılık, İstanbul.

[10] https://www.socialbakers.com/statistics/twitter/ , Twitter İstatistiği,11/05/2018.

[11] **Şeker** , **S. E.** , 2016. Duygu Analizi, YBS Ansiklopedi,Yönetim Bilişim Sistemleri Ansiklopedisi, **3**, 22-30.

[12] **Çelik**, **U.**, **Akçetin**, **E.**, **Gök**, **M.**, 2017. Rapidminer ile Uygulamalı Veri Madenciliği, Pusula Yayıncılık, İstanbul.

[13] http://bilgisayarkavramlari.sadievrenseker.com/2008/12/01/ozellik-cikarimi-feature-extraction/ , Özellik Çıkarımı, 14/05/2018.

[14] **Sert**, **F.** ,**Tüzüntürk**, **S.**,**Gürsakal**, **N**., 2014. NodeXL ile Sosyal Ağ Analizi: #AkademikZam Örneği, Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu, Isparta, Volume: 15.

[15] **Gürsakal ,N.,** 2016. Sosyal Ağ Analizi, Anadolu Üniversitesi Yayınları, Eskişehir.

[16] https://tr.wikipedia.org/wiki/KNIME , Knime Tanım, 15/05/2018.

[17] https://ceaksan.com/tr/knime-nedir/ , , Knime Tanım, 15/05/2018.

[18] **Türkmenoğlu**, **C.** , 2015. Türkçe Metinlerde Duygu Analizi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[19] **Özkan**, **Y.** , 2013. Veri Madenciliği Yöntemleri, Papatya Yayıncılık, İstanbul.

[20] **Ayhan, S**., ve **Erdoğmuş**, **Ş**. ,2014. Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü için Çekirdek Fonksiyonu Seçimi, Eskişehir Osmangazi Üniversitesi İİBF Dergisi, **28**, 176-179.

[21] http://bilgisayarkavramlari.sadievrenseker.com/2008/12/01/cok-sinifli-dvm-multiclass-svm/ , SVM Multiclass Sınıflandırma, 15/05/2018.

[22] https://www.e-adys.com/makine_ogrenmesi/naive-bayes-classifier/ , Naive Bayes, 15/05/2018.

[23] **Seçkin Codal**, **K.** , **Çoşkun**, **E.,** 2016. Sosyal Ağ Türlerinin Karşılaştırılmasına İlişkin Bir Ağ Analizi, Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, ISSN (Online): 2147-3064.

[24] **Beğenilmis**, **E.**, 2017. Detected of Organized Behaviours on Twitter, Yüksek Lisans Tezi, Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

[25] **Şeker** , **S. E.** ,**Yeşilyurt**, **A.,** 2017. Metin madenciliği Yöntemleri ile Duygu Analizi, YBS Ansiklopedi,Yönetim Bilişim Sistemleri Ansiklopedisi, **4**, 27-28.

[26] **Shah**, **N.**, https://blog.twitter.com/marketing/en_us/a/2016/introducing-emoji-targeting.html, , Introducing emoji targeting , 30/05/2018

[27] **Abdullah, M., Hadzikadic, M. ,** 2017. Emotions Revealed Regarding Donald Trump during the 2015-16 Primary Debates, International Conference on Tools with Artificial Intelligence, North Carolina, U.S.

[28] **Yan, J., Matthews, S. J.,** 2016. Applying Clustering Algorithms to Determine Authorship of Chinese Twitter Messages.

[29] **Ahuja, S., Dubey, G.,** 2017. Clustering and Sentiment Analysis On Twitter Data, International Conference on Telecommunication and Networks.

ÖZGEÇMİŞ

2006 yılında Elazığ Balakgazi Lisesi'nden mezun olduktan sonra 2008 yılında Fırat Üniversitesi Bilgisayar Mühendisliği bölümünü kazandım. 2012 yılında Fırat Üniversitesinden mezun oldum ve yine aynı yıl içerisinde Teknokent Sonra Bilgi Teknolojileri Ltd. Şti. firmasında İOS uygulama geliştirme alanında çalışmaya başladım. 2014 yılında atandığım Muş Alparslan Üniversitesi Bilgi İşlem Daire Başkanlığı Yazılım biriminde çalışmaya devam etmekteyim. Orta düzeyde İngilizce bilgisine sahibim. Teknik anlamda ilgi alanlarım veri madenciliği, sosyal ağlar, görüntü işleme konularından oluşmaktadır.

Dilber ÇETİNTAŞ
d.cetintas@alparslan.edu.tr