T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL MEDYADA DUYGU ANALİZİ VE NİTELİK ÇIKARIMI

TUGAY ÖZGİRGİN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
DOÇ. DR. BANU DİRİ

İSTANBUL, 2016

ÖNSÖZ

Bu çalışma, sosyal medya verileri üzerinde duygu analizi ve nitelik çıkarımı üzerine yapılmıştır. Twitter üzerinden elde edilen verilerin duygu analizi iki sınıfa ayrılarak gerçekleştirilmiş, her bir sınıf için ilgili nitelikler çıkartılmış ve hangi niteliklerin bu sınıfa ait veriyi nitelendirdiği incelenmiştir. Bu çalışma ile sosyal medya platformlarından elde edilen verilerin otomatik bir şekilde anlamlandırılması amaçlanmıştır.

Bu tezde yardım ve emeğini hiç esirgemeyen, bilgisi ve yol göstericiliği ile bana destek olan, kendisinden çok şey öğrendiğim değerli hocam Doç. Dr. Banu Diri'ye teşekkürlerimi sunarım.

Bu çalışmanın yapımı sırasında, veri setlerinin etiketlenmesinde bana yardımcı olan arkadaşlarım Ahmet Kök ve eşi Rabia Kök, Hakan Akhan, Kenan Dönmez'e teşekkürü bir borç bilirim.

Ayrıca, tez ve eğitim sürecim boyunca ihmal ettiğim aileme sonsuz bir sabırla yüksek lisans süreci boyunca bana verdikleri destekten dolayı onlara minnet borçluyum.

Temmuz, 2016

Tugay ÖZGİRGİN

İÇİNDEKİLER

Sayfa

KISALTMA LİSTESİ

API	Application Programming Interface
CRM	Customer Relationship Management
IEEE	Institute of Electrical and Electronics Engineers
JSON	JavaScript Object Notation
LR	Logistic Regression
ME	Maximum Entopy
MLP	MultiLayer Perceptron
NB	Naive Bayes
NoSQL	No Structured Query Language
OLP	Overall Linguistic Performance
ORM	Object Relational Mapper
POS	Part of Speech
RT	Retweet
SQL	Structured Query Language
SVM	Support Vector Machine
VTYS	Veri Tabanı Yönteim Sistemleri

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

SOSYAL MEDYADA DUYGU ANALİZİ VE NİTELİK ÇIKARIMI

Tugay ÖZGİRGİN

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Banu DİRİ

Günümüzde artan sosyal medya kullanımı, bu alandaki veriler üzerinden analiz ihtiyacını doğurmuştur. Bu çalışmada sosyal medya verilerinin duygu analizi yapılarak sınıflandırılmasını ve bu sınıflanmanın altında yatan gerçek nedenin ne olduğunun ortaya çıkarılması ve niteliklerin belirlenmesi amaçlanmıştır. Bu işlem gerçekleştirilirken sosyal medyadaki verilerin genellikle imla kurallarından yoksun olması, çalışmada kullanılan verilerin bir ön işlem yardımıyla düzeltilmesi gerekliliğini de ortaya çıkarmıştır. Çalışma ışığında veriler olumlu ve olumsuz olmak üzere iki sınıfa ayrılmış, bu gruplar üzerinden de kişilerin yaptığı paylaşımların hangi noktaya doğru yoğunlaştığı, bu yoğunlaşmanın hangi unsuru nitelendirdiği bulunmaya çalışılmıştır. Duygu analizi yapılırken öncelikle veriler ön işlemden geçirilmiş ve metinler düzetilmiş, daha sonra makine öğrenmesi teknikleri kullanılarak analiz gerçekleştirilmiştir. Niteliklerin çıkarılması için terim varlığı, frekans analizi ve kelimelerin yapısal olarak incelenmesiyle bir çıkarım elde edilmiştir.

Anahtar Kelimeler: Duygu analizi, nitelik çıkarımı, makine öğrenmesi, metin düzeltme

ABSTRACT

SENTIMENT ANALYSIS AND ATTRIBUTE EXTRACTION OVER SOCIAL MEDIA

Tugay ÖZGİRGİN

Department of Computer Engineering

MSc. Thesis

Adviser: Doç. Dr. Banu DİRİ

The increasing use of social media nowadays has led to the need for analysis on the data in these platforms. In this thesis, we aimed to classify the data by sentiments and clarify real reasons of this classification, by doing this study we assumed that the real attibutes of this classification can be understood. While processing the data, the need to preprocess and correct the data has risen due to the unstructured and ungrammatical text on social media. In the light of this study the data is splitted into two groups, which are positive and negative, then the direction of users' shares and entries are investigated and attribute based condensation of data is tried to be clarified. While analyzing sentiments firstly the data has been preprocessed and spellings has been corrected, then analysis has been done by using machine learning techniques. The result has been taken by using presence of terms, frequency analysis and structure analysis of data.

Keywords: Sentiment analysis, attribute extraction, machine learning, spelling correction

BÖLÜM 1

GİRİŞ

İnternetin yaygınlaşmasıyla ortaya atılan Web 2.0 kavramı kullanıcıların da artık internetin gelişmesine katkı vermesine olanak sağlamıştır. Bu da hayatımıza hatta yeni meslekler ve yeni alanlar girmesine sebep olmuştur. Wikipedia'nın önlenemez büyümesiyle artık kullanıcı tabanlı içeriklerin katkısı yadsınamaz hale gelmiştir. Bunun sonucunda hayatımıza **sosyal medya** yeni bir kavram olarak girmiştir. Sosyal medya kişileri, diğer kişiler, kurumlar ve popüler kişiler ile birbirine doğrudan bağladığı için ve buna ek olarak herkes kişisel fikirlerini özgürce paylaşabildiği için büyük bir hızla büyüme eğilimi göstermiştir. Bu büyüme beraberinde bir çok yenilik getirdiği gibi başka sorunlar da getirmiştir. Sorunların başında verilerin saklanması gelmektedir. O kadar çok veri nasıl ve nerede saklanacaktır, kayıtlar nasıl tutulacaktır? Bu sorun beraberinde yeni teknolojilerin ve yeni çözümlerin doğmasına izin vermiştir. Çözüm olarak **Bulut Mimarisi** (Cloud Architeture) ve **Büyük Veri** (Big Data) kavramları hayatımıza girmiştir. Verilerin büyüklüğü onların ölçeklendirilmesi gerekliliğini doğurmaktadır. Ayrıca, bu gereklilik sağlanırken sistemlerin kesintiye de uğramaması gerekmektedir. Bu sorunun çözümü için Bulut Mimari geliştirilmiştir. Artık dağıtık bir şekilde veriler saklanabilmekte, sunucular çok hızlı bir şekilde yatay ve dikey olarak ölçeklendirilebilmektedir. Günümüzde verilerin saklanması sorunu da ortadan kalkmıştır. Çözülen bu sorundan sonra şimdi de bu verilerin nasıl analiz edilmesi gerektiği problemi karşımıza çıkmıştır. Ancak, bu kadar çok veri nasıl hızlı ve etkili şekilde analiz edilecektir? Bu sorun karşısında bilişim uzmanları bu işlemi gerçekleştirebilmek adına NoSql araçlarını geliştirmişlerdir ve kullanılmaya başlamışlardır. Zira bu işlem yapılırken geleneksel bir yöntem olan ilişkisel veritabanları bunu karşılayamaz durumdadır. NoSql araçları sayesinde verilerin dağıtık saklanabilmesi ve işlenebilmesi mümkün olmuştur.

Sosyal medyada oluşan veriler çok hızlı büyüme göstermektedir. Aynı şekilde kurumlar, ürünler ve kişiler de artık sosyal medyayı etkili bir şekilde kullanmaya başlamıştır. Çünkü bu mecralar son kullanıcı ile kurum ve popüler kişileri doğrudan buluşturabilmektedir. Örneğin; geçmişte bir ürün hakkındaki şikayetinizi direk olarak o kuruma iletmekte zorluk yaşayabilirken, günümüzde sosyal medya sayesinde paylaştığımız şikayetler kurumlarlara daha kolay iletilebilmektedir. Ayrıca, bu şikayetler diğer kullanıcılar tarafında da görüldüğü için daha çok farkındalık yaratmakta ve kurumların bu şikayetler karşısındaki çözüm süresi oldukça hızlanmaktadır. Buna ek olarak sosyal medya, çıkarılan ürün ya da kurum hakkında çok hızlı geri dönüş alma konusunda çok etkili olarak kullanılmaktadır. İnsanlar artık fikirlerini paylaşmaktan çekinmiyor ve bu da gelen geri bildirimlerin, şikayetlerin analiz edilmesi gereksinimini kaçınılmaz hale getirmektedir. Kurumlar ve popüler şahıslar artık bir sosyal medya uzmanıyla çalışmakta ve sosyal medyayı aktif bir şekilde kullanmaya özen göstermektedirler. Bu yüksek kullanım oranının bir sonucu olarak da, bu mecradaki verilerin büyümesi büyük bir hızla artmakta ve bu verilerin analizi için çok büyük kaynaklara ihtiyaç duyulmaktadır.

Sosyal medya platformlarından biri olan Twitter 21 Mart 2006 yılında kurulan bir mikro blog sitesi olup, 140 karakter sınırı ile insanların fikir ve düşüncelerini bu ortamda paylaşmasına izin vermektedir. Twitter'ın kullanımının artmasıyla beraber bu ortamlar artık yeni reklam ve pazarlama mecrası olarak da görülmeye başlamıştır. Bu evrimden sonra artık Twitter'ı kurumlarda kullanmaya başlamış ve ürünleri hakkında bu mecralardan paylaşımlar yapmaya özen göstermektedirler. Kullanıcılar da aynı şekilde bu bildirimlere cevap verebilmekte ve geri bildirim yapabilmektedir. Bu kadar önemli hale gelen Twitter artık çok fazla kişi tarafından kullanılır hale gelmiştir. Twitter kullanım rakamlarına örnek vermek gerekirse;

- Günlük atılan tivit sayısı ortalama 58 milyon.
- Twitter aylık aktif kullanıcı sayısı 115 milyon
- 1 Milyon tivite ula**ş**ması için gereken süre ortalama 5 gün
- Saniyedeki tivit sayısı ortalama 9100

Bu veriler 25 Eylül 2015'te yapılan araştırmadan [1] elde edilmiştir. Twitter daki verilerin bu denli büyük olması ve paylaşılan verilerin herkese açık olması sebebiyle bu çalışmada

hedef platform olarak Twitter verileri kullanılmıştır. Kullanılan veriler kurum ve kişilerin Twitter sayflarından alınmış ve bu veriler üzerinde analiz işlemi gerçekleştirilmiştir.

1.1 Literatür Özeti

Sosyal medyadaki verinin büyüklüğü ve kullanım yoğunluğu doğal dil işleme üzerine çalışan kişilerin en popüler çalışma alanlarından biri olmuştur. Twitter'ın da bu verilere erişmede sağladığı kolaylık (API) ve verinin miktarının oldukça büyük olması sebebiyle biz de çalışmamızda Twitter üzerindeki verileri analiz etmek için kullandık. Benzer çalışmalara bakıldığında doğal dil işleme ve veri madenciliği alanında Twitter'ın ön plana çıktığı görülmektedir. Verinin bu kadar kolay elde edilmesi ve herkes tarafından bu verilerin girilmesi beraberinde bazı problemler de getirmektedir. Problemlerden birincisi kullanıcıların fikirlerini 140 karakterle anlatması gerekliği yüzünden kullanıcıların kuralsız olarak kelimeleri kısaltmaları, dil bilgisi kurallarından yoksun ve imla kurallarına uymadan girilen bu verilerin analizinin zorluğu veri analizi yapan kişilerin önündeki en büyük sorunlar olarak ön plana çıkmaktadır. Bu sorun öncelikle veriler üzerinde ön işlem yapılması gerekliliğini doğurtmaktadır. Bu alanda yapılan çalışmalarda kullanılan teknik iki grupta incelenebilir. Birincisi sözlük tabanlı teknikler, ikincisi ise makine öğrenmesi yöntemleri kullanılarak yapılan çalışmalardır. Biz bu çalışmada iki yöntemi birleştirerek sonuç elde etmeye çalıştık.

Makine öğrenmesi teknikleriyle duygu analizi yapılırken, eğiticili öğrenme teknikleri kullanılmaktadır. Etiketlenen veriler yardımıyla sistem eğitilir. Böylece sistem yeni gelen verilerin sınıfına karar verir duruma gelmektedir. Bu işlem yapılmadan önce verilere bir normalizasyon ve ön işlem uygulamak gerekir. Uygulanan normalizasyon içinde metin düzeltme amacıyla bazı teknikler kullanılmıştır.

"Speech and Language Processing 2nd Edition" isimli kitapta [2] anlatılan kelime bazlı n-gram yöntemleri bu çalışmada denenmiş ancak, Türkçe metinler için istenilen başarı elde edilememiştir. Metin düzeltme için gerçekleştirilen yöntemin başarısı "Evaluating Evaluation Metrics for Spelling Checker Evaluations" isimli çalışmada [3] otomatik bir metin düzelticinin başarısının nasıl ölçüleceği anlatılmaktadır.

Ayrıca metin düzeltme amacıyla geliştirilen yöntem, Gülşen Eryiğit ve Dilara Torunoğlu tarafından "A Cascaded Approach for Social Media Text Normalization of Turkish" isimi

makaleden [4] incelenmiş ve ilgili çalışmada sistemin testi için kullanılan veri seti ile de test edilerek, ilgili çalışma ile kıyaslama yapılmıştır.

2012 yılında "Sentiment Analysis and Opinion Mining: A Survey" isimli çalışmada [5] farklı konularda yapılan eleştiriler Naive Bayes (NB), Maximum Entropy (ME) ve Support Vector Machine (SVM) algoritmaları kullanılarak sınıflandırılmaya çalışılmıştır. Bu çalışmada özellik vektörü olarak terim varlığı, terim frekansı, olumsuzluk, n-gramlar ve POS etiketleri kullanılmış ve eleştiriler olumlu ve olumsuz olarak 2 sınıfa ayrılarak incelenmiştir.

"On the Optimality of the Simple Bayesian Classifier under Zero-One Loss" isimli başka bir araştırmada [6] duygu analizinde Naive Bayes'in (NB) özelliklerin biribine bağlımlı olduğu durumda daha iyi çalıştığı gösteren bir çalışmadır. Naive Bayes'in özelliklerin temel yaklaşımı olan bağımsız özellik yaklaşımına karşın alınmış sonucu bir sürpriz olarak algılanmış ve bu çalışma üzerine Zhen Niu "Sentiment classification for microblog by machine learning" isimli başka bir çalışma [7] yapmış ve özellik seçimi için yeni bir model ortaya koymuştur. Bu model ağırlık hesaplamaları ve sınıflandırmayı içermektedir. Bu çalışmada ağırlıkları kullanarak NB algoritmasını daha efektif bir şekilde kullanmıştır.

"Robust sentiment detection on twitter from biased and noisy data" isimli bir başka çalışmada [8] tivitleri sınıflandırmak için iki adımlı bir sınıflandırma metodu geliştirilmiştir. Bu çalışmada sınıflandırma yöntemi geliştirilirken gürültülü bir eğitim seti kullanılmış, böylelikle etiketleme süresi azaltılmıştır. Tivitler nesnel ve öznel olarak 2 gruba ayrılmış daha sonra nesnel olan tivitler pozitif ve negatif olarak işaretlenmiştir.

Aslı Çelikyılmaz tarafından yapılan "Probabilistic model-based sentiment analysis of twitter messages" isimli çalışmada [9] telaffuza dayalı kelime sınıflandırma sistemi geliştirilmiş, bu sayede gürültülü tivitlerde normalizasyon sağlanmıştır. Ayrıca, bu çalışmada benzer jetonların (token) atanması, html linkler, kullanıcı betimlemeler, hedef organizasyon isimleri gibi metin işleme teknikleri kullanılmıştır. Normalizasyon adımından sonra gerçekleştirilen sınıflandırma için olasılıksal bir model tercih edilmiştir. BoosTexter sınıflandırıcısı yardımıyla elde edilen "polar lexicon" dağılımı ile özellikler seçilmiş ve hata oranı azaltılmıştır.

“Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model“ isimli bir diğer çalışmada [10] emojilerin metnin duygusunu nasıl etkilediği araştırılmıştır. Bu çalışmada Twitter API yardımıyla otomatik olarak tivitleri toplayan ve bu tivitleri emojilerine göre otomatik sınıflandıran bir model geliştirilmiştir. Tivitlerin N-gram ve POS özellikleri çıkarılmış ve etiketleme için Multinominal Naive Bayes (MNB) yöntemi kullanılmıştır. İlgili çalışmada tivitler sadece emojiler yardımıyla etiketlendiği için hata olasılığı yüksek olmuştur.

“Ensemble of feature sets and classification algorithms for sentiment classification“ isimli başka bir çalışmada [11] ise ensemble çatısı yardımıyla bir sınıflandırma gerçekleştirilmiştir. Ensemble çatısı, birden fazla algoritmanın birleştirilmesiyle oluşturulan bir sınıflandırma modelidir. Bu model sayesinde farklı veriler üzerinde farklı başarıları olan yöntemler bir araya getirilerek daha etkili bir yöntem oluşturulması sağlanmıştır. Ensemble yöntemi kullanılan bu çalışmada iki farklı özellik ve üç farklı algoritma kullanılarak sınıflandırma yapılmıştır. Kullanılan özellikler POS etiketleri ve kelime ilişkileri olarak seçilmiştir. Algoritma olarak ise ME, NB ve SVM kullanılmıştır.

“Domain adaptation in sentiment analysis of twitter” isimli bir diğer çalışmada [12] Twitter’daki haber ve filmler hakkındaki kullanıcı yorumları üzerinden bir duygu analizi gerçekleştirilmiş ve bu işlem sonucu IMDB ve Blippr ile karşılaştırılmıştır.

2013 yılında yapılmış “Sentiment Analysis in Twitter using Machine Learning Techniques” ismindeki bir çalışmada [13] Twitter’daki verilerin makine öğrenmesi teknikleriyle sınıflandırılması yapılmış ve bu çalışmada hem sözlüksel yöntem hem de makine öğrenmesi teknikleri birleştirilerek 4 farklı algoritma ile başarısı ölçülmüştür. Kullanılan yöntemler NB, ME ve SVM iken bunun yanında bu üç yöntemin birleşmesi olan Ensemble yöntemi de denenmiş ve en yüksek başarı bu yöntem ile elde edilmiş olup, %90 civarında olmuştur.

2009 yılında Yelena Mejanova tarafından yazılan “Sentiment analysis: An overview” isimli bir makalede [14] bir metnin ana fikrinin terim varlığı ve terim frekansı yardımıyla çıkartılabileceği belirtilmiştir. Ayrıca, Stemler tarafından 2001’de yazılan “An Overview of Content Analysis” isimli bir makalede [15] bir metnin ana fikrinin kelime yoğunluğu ile çıkarılabileceği vurgulanmıştır.

Bilkent üniversitesinde yapılan “Generic Text Summarization for Turkish” isimli makalede [16], metnin özetlenmesinde hangi tür anahtar kelimelerin etkili olacağı tartışılmış ve kısa isim tamlamalarını özetlemede çok önemli bir yer tuttuğu gösterilmiştir.

Farklı çalışmalar da incelendiğinde bir metnin hangi yazara ait olduğunun tespitiyle ilgili bilgi içeren “An Overview of Content Analysis” isimli makalede [15] kullanılan terim varlığı ve frekans analizi gibi yöntemlerin bizim çalışmamızda da başarılı olacağı varsayımını ortaya koymuştur.

Ayrıca, “Analysis of the Relation Between Turkish Twitter Messages and Stock Market Index” isimli [17] çalışma ile borsadaki veriler ile twitter girdileri birlikte incelenmiş ve aralarında %45 lik bir ilişki olduğu gözlenmiştir. Bu da twitter verilerinin analizi sonucu bir yargıya varılabileceğini gösteren bir çalışma olarak öne çıkmaktadır.

“Sentiment Analysis on Social Media” isimli çalışmada [18] iki şirket hakkında atılan tivitler incelenmiş ve algı karşılaştırılması yapılmıştır. Bu çalışmada isim, sıfat, fiil gibi kelime türleri incelenerek %96’lık bir başarı elde edilmiştir.

1.2 Tezin Amacı

Twitter her gün milyonlarca insanın her konuda attığı tivitlerle gittikçe ve çok büyük bir hızla büyüyen sosyal medya platformu olduğu için bu alandaki veriler değerli ve anlamlıdır. Bu sebeple bu ortamdan elde edilen kullanıcı fikirlerinin önemi gittikçe daha fazla önem arz eder hale gelmektedir. Nitekim bu tür sosyal medya platformları kurum, ürün veya ünlü kişileri son kullanıcıyla buluşturduğu için hem kurumlar için hem de şahıslar için oldukça önemlidir. Twitter’daki verilerin kullanılmasının bir diğer sebebi ise herkese açık olarak bir API aracılığı ile bu verilerin erişilebilir olmasıdır. Bu sayede verilerin analiz edilmesine olanak sağlanmaktadır.

Bu tez ile amaçlanan yöntem, bu verilerin otomatik olarak işlenmesi ve anlamlandırılmasıdır. Çünkü çok büyük bir kitle tarafından kullanılan bu platformlar, kurumların, ürünlerin veya ünlü kişilerin kendileri hakkında son kullanıcının ne düşündüğünü analiz etmesi ve bu doğrultuda strateji belirlemesi için gereklidir. Buradaki verinin büyüklüğü ise insan eliyle bu analizin gerçekleştirilmesinin oldukça zor hatta

imkansız olduğu gerçeğidir. Çok hızlı büyüyen bu veri ancak dağıtık bir şekilde işlenerek ve otomatik bir şekilde analiz edilerek kullanılabilir. Bu sebepledir ki yapılan bu çalışma ile son kullanıcı verileri duygusal açıdan iki sınıfa ayrılmış ve bu sınıflar kendi içinde analiz edilerek neden o gruba ait olduğunu belirleyen niteliklerin dağılımı incelenmiştir. Bu inceleme sonucu elde edilen çıkarım ilgili kurum, ürün veya şahıs için hızla değişen piyasada nasıl bir strateji belirlemesi gerektiğini hızlı bir şekilde öngörüp, ilerlemesine olanak sağlamaktır. Elde edilen sonuçlar bu niteliklerin olumlu ve olumsuz sınıftaki oranları çıkartılarak, son kullanıcı fikirlerinin sınıflandırmadaki dağılımlarının gösterilmesi amaçlanmıştır. Yani, bir sınıf içinde yoğun olarak geçen bir niteliğin aynı zamanda diğer grup içerisindeki dağılımına da bakılarak o niteliğin her iki grup içindeki dağılımının gösterilmesi amaçlanmıştır.

1.3 Hipotez

Son yıllarda gelişen teknolojiler sayesinde artan hesaplama hızlarıyla artık büyük verilerin analiz edilmesi kolaylaşmıştır. Bu sebeple kullanıcı yorumları büyük bir hızla işlenerek analiz edilebilmektedir.

Bu gelişmeler ışığında film yorumları, ürün yorumları ve borsa değişimleri kullanıcıların yorumları üzerinden analiz edilebilmekte ve bir tahmin çıkarılabilmektedir. Hatta son yıllardaki bu analizler geçmiş verilere dayanarak ortalama gerçekleşecek satış miktarlarını dahi tahmin edebilmektedir. Örneğin, geçtiğimiz yıl piyasaya çıkan yeni bir telefon modelinin ortalama satış miktarı lokasyon tabanlı bir sosyal medya platformu tarafından açıklanmıştır. Bu rakamlar ilk zamanlar ses getirmezken, ilk gün satışları tamamlandığında açıklanan rakamlar arasındaki ilişkinin birbirine çok yakın olması verinin analizi ve sonuç çıkarımının aslında nelere imkan sağladığı herkes tarafından görülmüştür.

Bu tezde ortaya atılan hipotez bir kurum, ürün veya şahıs hakkında kullanıcıların yorumu, aslında o şeyin niteliklerini belirler varsayımıdır. Yani, bir kurum hakkında kullanıcıların şikayeti veya talebi aslında o kurumun spesifik olarak o konu hakkında ki niteliklerini göstermekte olduğu varsayılmıştır. Örneğin, bir kullanıcının bir şirket hakkındaki servis kalitesini gösteren bir şikayeti, o kurumun servis niteliğinin kapasitesini ve kalitesini belirleyen en önemli faktördür. Lakin bu bağlamdaki kullanıcı yorumları olumlu veya

olumsuz olmasının belirlenmesi o niteliğin aslında hangi kategoride değerlendirileceği açısından önemlidir. Bu sebepledir ki bu tezde öncelikle verilerin doğruluğu ve sınıflandırılması, daha sonraki aşamada ise bu gruplanmanın altında yatan gerçek nedenin ne olduğunun araştırılması gelmektedir. Bu doğrultuda yapılan hipotez, Twitter verilerinin aslında bir kurum, ürün veya şahıs hakkında çıkarım yapabileceği varsayımına dayanmaktadır.

BÖLÜM 2

TWITTER VERİLERİ

Twitter 2006 yılında kurulmuş bir micro blog sitesidir ve yaklaşık 320 milyon aylık aktif kullanıcısı vardır. SMS'in mantığından yola çıkarak kullanıcıların fikirlerini 140 karakterle paylaşmasına izin veren bir platformdur.SMS 160 karakterden oluştuğu için Twitter'da 140 karaktermesaja, 20 karakter de kullanıcı adına ayrılmıştır. Son günlerde alınan bir karar ile 140 karakteri 10000 karaktere çıkaracağını duyurmuştur [19]. Bu gerçekleşirse önümüzdeki günlerde kullanıcılar daha ayrıntılı ve daha düzgün bir şekilde fikirlerini dile getirebilecekler. Twitter'in 140 karater sınırlandırması Twitter verilerinde karşımıza yeni bir problem getirmektedir. Bu problem 140 karaktere kullanıcılar fikirlerini sığdıramadıklarından bazı kısaltmalarınkaçınılmaz olarak kullanılmasıdır. **Selam** kelimesi yerine **slm** kelimesinin kullanılması buna bir örnektir. Buradaki kısaltmalar belli bir düzene göre yapılmamakta, tamamen kullanıcının insiyatifindedir. Bu yüzden paylaşılan verilerde kısaltmalar birbirinden çok farklı olmaktadır. Bu da önümüze bu verilerin düzeltilmesi problemini çıkarmaktadır. Ayrıca, Twitter'da paylaşılan website adreslerinin de 140 karakter sınırına takılmaması için kısaltılarak yazılması gerekmektedir.

Twitter'daki karakter kısıtlaması sebebiyle kullanıcılar daha az karakter sayısıyla duygularını anlatabilmek amacıyla emojiler kullanmaktadır. Bu emojiler duygunun daha kolay anlaşılmasına yardımcı olmaktadır. Örneğin ;), :) gibi emojiler olumlu bir duyguyu ifade ederken :(, :`(gibi emojiler olumsuz duyguları anlatmaktadır. Bu tip emojiler tivitlerde çok yoğun bir şekilde kullanılmaktadır.

Twitter'daki verilerde 3 önemli kavram vardır. Bunlardan birincisi hashtag, ikincisi kullanıcı adı ve üçüncüsü de retweet kavramıdır. Hashtag, girilen verinin belli bir konuda etiketlenmesini sağlayan "#" sembolü kullanılarak etiketlenirler. Kullanıcı adları ise "@"

sembolü ile başlamaktadır. Belli bir kullanıcıya mesaj göndermek istendiğinde, başında "@" sembolünü kullanarak o kullanıcıya herkese açık olacak şekilde mesaj göndermek mümkündür. Retweet ise bir kişinin mesajını beğenmeniz durumunda onu kendi takipçileriniz ile paylaşmanıza o tivitin retweet edilmesi denilmektedir. Bu durumda mesajın başına "RT" ön eki eklenerek, aynı mesaj hiç değiştirilmeden takipçilerle paylaşılır.

Twitter verilerine API aracılığı ile erişilebilinir. Belli bir hashtag veya kullanıcı için atılmış herkese açık tivitler API aracılığı ile elde edilebilmektedir. Bu API, JSON formatında sonuç geri döndürmektedir. Bu formatta elde edilen veri içerisinde kullanıcı adı, oluşturulma tarihi, lokasyon bilgisi, dili, mesaj ve id'si gibi bir çok veri bulunmaktadır. Twitter API'sinin olumsuz yönü belli kısıtlara sahip olmasıdır.. Belli bir süre aralığında sorgu sayısı ve her istekte dönen tivit sayısı kısıtlanmıştır. Bu limitlere takılmamız durumunda 15 dakika bekleyerek yeni bir sorgu gerçekleştirmemize olanak tanınmaktadır. Ayrıca, her bir sorguda ancak 100 adet tivit geri dönmektedir. Twitter API kullanabilmek için Twitter üzerinden bir uygulama oluşturarak belli erişim parametrelerinin elde edilmesi gerekmektedir. Bu parametreleri elde ettikten sonra herkese açık olan API metotları aracılığı ile yine herkese açık olan tivitlere erişme olanağı tanınmaktadır.

2.1 Verilerin Toplanması

Metin düzeltme verileri için **köşe yazılarından** oluşan bir **derlem** kullanılmıştır. Bu derlem yaklaşık **18 köşe** yazarına ait **630 köşe** yazısından oluşmaktadır. Bu derlem Yıldız Teknik Üniversitesi Kemik grubu web sitesinden temin edilmiştir [21].

Bu çalışmada duygu analizi ve nitelik çıkarımı çalışmasının gerçekleştirilmesi için veriler Twitter'in API aracılığı ile elde edilmiştir. Bu işlem için toplamda **5.990** adet tivit eğitim ve **2.509 adet** tivit de test ve nitelik çıkarımı için kullanılmıştır. Bu tivitlerin elde edilmesi için Pyhton dili ile bir betik yazılmış ve veriler Excel formatında kaydedilmiştir. Bu işlemi gerçekleştirmek için Twitter API'sinin **search** methodu kullanılmıştır. Ayrıca, nitelik analizi için Türkiye'nin GSM opeartörlerinin hesaplarına gönderilen güncel tivitler ve bir futbolcunun hesabına atılmış tivitler de kullanılmıştır. Bu veriler yazılan betik aracılığı ile elde edilmiştir. Veriler direk olarak kullanılmamış içinde "**@Operatör Hesabı**" şeklinde

geçen, direk o hesaba atılmış tivitler kullanılmıştır. Ayrıca, RT'ler ve lokasyon paylaşımları bu tivitler içerisinden ayıklanmıştır.

Twitter API'sinde bulunan kısıtlar yüzünden tivitler 100 parçalık gruplar halinde çekilmiştir. Şekil 2-1 ve Şekil 2-2'de Twitter API'sinin örnek istek ve cevabı görülmektedir. Bu işlem yapılırken Twitter'in API'sinden [24] güncelden en eskiye doğru veriler sorgulanmıştır.

```
GET /1.1/search/tweets.json?q=turkcell&count=1
```

Şekil 2.1 Örnek İstek


```
{
  "statuses": [
    {
      "metadata": {
        "result_type": "recent",
        "iso_language_code": "tr"
      },
      "created_at": "Mon Feb 29 22:54:02 +0000 2016",
      "id": 704439514727649300,
      "id_str": "704439514727649282",
      "text": "Turkcell durup dururken neden böyle bisi yaptı anlam veremedim, glb bana asık :sss
https://t.co/xqeeOaLUdk",
      "source": "<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>",
      "truncated": false,
      "in_reply_to_status_id": null,
      "in_reply_to_status_id_str": null,
      "in_reply_to_user_id": null,
      "in_reply_to_user_id_str": null,
      "in_reply_to_screen_name": null,
      "user": {},
      "geo": null,
      "coordinates": null,
      "place": null,
      "contributors": null,
      "is_quote_status": false,
      "retweet_count": 0,
      "favorite_count": 0,
      "entities": {},
      "favorited": false,
      "retweeted": false,
      "possibly_sensitive": false,
      "lang": "tr"
    }
  ],
  "search_metadata": {
    "completed_in": 0.092,
    "max_id": 704439514727649300,
    "max_id_str": "704439514727649282",
    "next_results": "?max_id=704439514727649281&q=turkcell&count=1&include_entities=1",
    "query": "turkcell",
    "refresh_url": "?since_id=704439514727649282&q=turkcell&include_entities=1",
    "count": 1,
    "since_id": 0,
    "since_id_str": "0"
  }
}
```


Şekil 2.2: Örnek Cevap

2.2 Verilerin Saklanması

Twitter'dan betik aracılığı ile çekilen veriler Excel formatında saklanmış olup, duygu analizi modülüne ve nitelik çıkarımı modülüne veriler aktarılırken bu dosyalar kullanılarak gerçekleştirilmiştir.

Duygu analizi ve nitelik çıkarımı için geliştirilen uygulamanın çalışması şöyledir; Elde edilen tivit verileri uygulama çalıştığı esnada içeri aktarılarak analiz işlemi gerçekleştirilir. Verilerin ana uygulamada saklanmasında bir **NoSQL** veritabanı olan **Apache Cassandra** kullanılmıştır.

Apache Cassandra [25] ilk olarak Facebook tarafından mesajların aranması için geliştirilen bir araçtır. Bu geliştirmeyi Facebook adına yapan Prashant Malik ve Amazon Dynomo'yu geliştiren Avinash Lakshman'dır. Haziran 2008'de Google Code'da açık kaynak proje olarak yayınlanmıştır [20]. 2009'un Mart ayında ise Apache Incubator'a dahil olmuştur. Cassandra'nın diğer NoSQL veritabanlarından farkı, normal SQL veritabanları gibi kolon tabanlı bir yapısının olmasıdır. SQL diliyle veri çekilebilmesine olanak tanır. Bu sebeple veritabanındaki veri modelinin belli bir yapısının olması gerekmektedir. Diğer NoSQL veri tabanları (MongoDB, CouchDB …) gibi tamamen yapısal bir serbestlik söz konusu değildir. Verilerin mutlaka belli bir yapıya uygun olması gerekir. Ayrıca, Cassandara yatay olarak ölçeklendirilebilir ve veri dağıtık bir şekilde saklanıp işlenebilir. Diğer klasik VTYS'leri gibi ölçeklendirildiği zaman hız logaritmik olarak artmak yerine, doğru orantılı olarak artmaktadır. Bu da Cassandra'nın en büyük avantajıdır.

Çizelge 2.1 Cassandra Tablo Yapısı

id	hashtag	keywords	tweet	type

Ana uygulamada içeri aktarılan veriler ilgili hastag'a göre Cassandra veri tabanına aktarılmaktadır. Bu veriler Çizelge 2.1'deki gibi bir tablo yapısında saklanmaktadır. Bu tabloya ait sql oluşturma betiği Şekil 2.3'deki gibidir. Duygu analizi ve nitelik çıkarımında kullanılan veriler Cassandra veritabanından çekilerek gerçekleştirilmiştir.

```
CREATE TABLE sociolog.tweets (
    id uuid PRIMARY KEY,
    hashtag text,
    keywords list<text>,
    tweet text,
    type int
) WITH bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy',
'max_threshold': '32', 'min_threshold': '4'}
    AND compression = {'chunk_length_in_kb': '64', 'class':
'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND crc_check_chance = 1.0
    AND dclocal_read_repair_chance = 0.1
    AND default_time_to_live = 0
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair_chance = 0.0
    AND speculative_retry = '99PERCENTILE';
CREATE INDEX tweets_hashtag_idx ON sociolog.tweets (hashtag);
CREATE INDEX tweets_type_idx ON sociolog.tweets (type);
```

Şekil 2.3 Cassandar Veri Modeli

Uygulama her çalıştırıldığında ilgili hesaba ait veriler tablolardan boşaltılarak analizin yeniden en baştan steril bir şekilde gerçekleştirilmesi sağlanmıştır. Her çalışma esnasında uygulama Excel'den verileri ilgili veritabanına aktarmaktadır. Uygulamanın

doğrudan Twitter'dan veri çekerek analiz edebilecek modül alt yapısı da olmasına rağmen uygulamanın çalışması Excel üzerinden olacak şekilde kurgulanmıştır. Çünkü bu şekilde büyük verilerin analizi için gereken yapılar, anlık veriler üzerinden çalışmak yerine veri ambarları üzerinden çalışan asenkron işler (asynchronous jobs) ile yapılır. Bu sebeple doğrudan Twitter üzerinden erişim yapılması yerine Excel ile içeri aktarım yapacak şekilde kurgulanıp geliştirilmiştir.

2.3 Verilerin Etiketlenmesi

Verilerin etiketlenmesi insan eliyle olmuştur. Eğitim verisi en az iki kişi tarafından etiketlenerek oluşturulmuş ve test verisi olarak kullanılan 3 operatöre ait tivitler de en az bir kişi tarafından etiketlenmiştir. Test verisinin etiketlenme sonucu sistem üzerinden sınıflandırma gerçekleştirilerek test edilmiş. Farklı sınıf etiketine sahip tivitler tekrar gözden geçirilmiştir. Futbolcu ile ilgili tivitler ise sistem tarafından otomatik olarak sınıflandırılmıştır.

2.3.1 Eğitim Verisinin Etiketlenmesi

Eğitim veri seti **5.990** adet tivitten oluşmaktadır. **2.990** adet tivit olumsuz iken, **3.000** adet tivit olumlu bilgi içermektedir. Bu tivitler **telekominikasyon**, **sağlık**, **finans**, **spor**, **sigorta**, **gıda**, **otomotiv**, **gayrimenkul** ve **siyasi** içeriklere sahip tivitlerden oluşmaktadır. Eğitim setinin etiketlenmesi rastgele olarak 4 kişi tarafından etiketlenmiş, bir kişi tarafından etiketlenen veri başka bir kişiye daha gösterilerek doğrulama sağlanmıştır. Aynı tivitin en az 2 kişi tarafından etikenlenmesiye bu veriler eğitim setine dahil edilmiştir. Buradaki veriler kişilerin ruh hali ve düşünce yapısına göre farklılık gösterebilmektedir. Eğitim için kullanılan veri setinin farklı kategorilerden seçilmiş tivitler olması eğitim setinin bir çok alan için ortak olarak kullanılabilmesine olanak sağlaması amacıyla yapılmıştır. Ancak, test için kullanılan verilerin çoğunluğunun telekominikasyon sektörüne ait tivitler olması, sınıflandırma başarısını olumsuz yönde etkilediği görülmüştür. Bu da bize genel bir eğitim seti kullanılmasının, spesifik bir alan için sınıflandırma başarısını düşürebildiğini göstermiştir. Zira o kategorilerde geçen terimlerin ancak bir kısmı eğitim setinde olacağından böyle bir olumsuzluk ile karşı

karşıya kalınabilmektedir. Bunu aşmak için eğitim veri setinin çok fazla veri içermesi durumunda sorunun üstesinden gelinebileceği düşünülmektedir.

2.3.2 Nitelik Çıkarım Verisinin Etiketlenmesi

Nitelik çıkarımı için kullanılan veri ise Türkiye'nin 3 GSM operatörüne ve bir futbolcu hesabına atılmış tivitlerden oluşmaktadır. Bu tivitler en az bir kişi tarafından etiketlenmiş olup, sadece futbolcuya ait tivitler sistem tarafından otomatik olarak etiketlenmiştir. Çıkarılan tivitlerden lokasyon paylaşımı olanlar temizlenerek etiketleme işlemi yapılmıştır. Etiketleyen kişiler birbirlerinden habersiz olarak etiketle işlemini yapmış ve sonuçta iki kişi farklı görüş belirtmiş ise üçüncü bir kişinin görüşü alınarak o tivit için sınıf etiketi belirlenmiştir. Elde edilen bu veriler eğitim setiyle eğitilen uygulamaya test verisi olarak verilmiş ve bu sınıflar üzerinden nitelikler çıkartılmıştır. Nitelik çıkarımı için kullanılan verilerin eğitim setiyle eğitildikten sonra sınıflandırma başarısı da ayrıca ölçülmüştür.

BÖLÜM 3

VERİLERİN İŞLENMESİ

Verilerin işlenmesi 4 adımda gerçekleştirilmiştir. Bunlar sırasıyla içeri aktarım, ön işlem ve metin düzeltme, duygu analizi ve nitelik çıkarımıdır. Uygulamadaki işlem adımları ve iş akışı Şekil 3.1'de gösterilmiştir.

Şekil 3.1 Uygulama Adımları

3.1 İçeri Aktarım

Python programlama diliyle yazılmış olan bir betik yardımıyla tivitler çekilmiştir ve bu veriler Excel (xlsx) formatında kaydedilmiştir. Betik aracılığı ile tivitler çekilirken Twitter API'sinin arama özelliğinde bir kullanıcıya gönderilmiş tivitler çekilemediğinden gelen tivitler içinde "@Kullanıcı Adı" ifadesinin geçtiği tivitler seçilerek bir Excel dosyasına kaydedilmektedir. Java EE kullanılarak yazılmış programa girdi olarak bu dosyalar

verilmiş ve dosyaların içeriği uygulama tarafından okunarak veritabanına aktarılmıştır. Okunan tivitler bir **Tweet** sınıfına atanarak Cassandra veritabanına kaydedilmesi sağlanmıştır.

3.2 Önişlem ve Metin Düzeltme

Önişlem ve metin düzeltme modülü verileri düzgün bir formata getirmeyi amaçlamaktadır. Zira bir sonraki analiz ve sınıflandırma adımlarında başarıya etki edeceği varsayımı yapılmıştır. Bu modülde tivitler üzerinde gerçekleştirilen ön işlem adımları şöyledir;

- Web adreslerinin temizlenmesi
- Tek başına bir anlam ifade etmeyen edat, bağlaç vb. kelimelerin temizlenmesi (bir, daha, kadar gibi…)
- Hashtaglerin temizlenmesi
- Noktalama işaretlerinin temizlenmesi (emojiler hariç)
- Gereksiz boşlukların temizlenmesi
- Kelimelerin düzeltilmesi
- Kelimelerin parçalara ayrılması

Bu işlemler yapıldıktan sonra veriler duygu analizi ve nitelik çıkarımı için hazır hale gelmiştir. Kelimelerin düzeltilmesi sırasında Leveinstein ve Jaro Winkler algoritmalarından yararlanılmıştır. Bu işlem sırasında verilerin düzeltilmesi adına köşe yazılarının bulunduğu derlem üzerinden kontrol sağlanmıştır. Derlem içerisinde aynısının bulunması durumunda kelime doğru kabul edilmiştir. Aksi halde kelime sözlükteki kelimeler ile karşılaştırılmış ve Leveinstein algoritmasına göre uzaklığı % 51'den küçük olan kelimeler bir listede tutulmuştur. Bu liste daha sonra Jaro Winkler algoritmasıyla karşılaştırılmış ve frekansları da göz önüne alınarak en yakın kelimeye evrilmiştir. Ayrıca, kelimelerin olumsuzluk denetimi yapılıp, ona göre işaretlenmiştir. Olumsuz anlamını bulmak için değil veya me/ma gibi olumsuzluk eki içeren fiiler ile karşılaştığımızda kendisinden önce gelen 3 kelimenin başına "not_" eklenerek güncellenmiştir. Kendisinden önce gelen kaç kelimenin bu şekilde yapılması gerektiği yapılan testler

sonucunda en iyi sonucu veren 3 değeri alınarak kullanılmıştır. Detaylarından 4. Bölümde bahsedilecektir. Bu işlem sonunda özelliklerin çıkarılması için kelimeler 3-gramlarına ayrılan kelime parçacıklarının başına da eğer olumsuzluk içeren bir kelimenin n-gramları ise onlara da "not_" ön eki eklenmiştir.

Örneğin; "***Bu hizmeti veren iyi bir operatör olamaz***" ve "***Bu tarife hiç de kötü değil***" cümlelerini inceleyelim. Bu cümlelerdeki kelimeler denetlenerek Zemberek [22] yardımıyla olumsuzluk eki veya değil kelimelerini içerip içermediği kontrol edilmektedir. Olamaz kelimesi içindeki olumsuzluk eki tespit edildiği durumda iyi, bir ve operatör kelimeleri ***not_iyi***, ***not_bir*** ve ***not_operatör*** olacak şekilde değiştirilmiştir. Aynı benzer işlem diğer cümle için de yapıldığında ***not_hiç***, ***not_de***, ***not_kötü*** olarak değiştirilmiştir. Bu işlem sonucunda ***not_iyi*** aslında ***kötüyü***, ***not_kötü*** kelimesi de ***iyi*** kelimesini temsil etmektedir. 3-gramlar çıkarılırken de aynı şekilde not_ope, not_per, not_era … gibi çevrilerek özellikler çıkarılmıştır.

Ön işlem adımında yapılan işlemlerin tamamı Şekil 3.2'de gösterilmiştir.

Şekil 3.2 Ön işlem ve Kelime Düzeltme Adımı İşlemleri

3.3 Duygu Analizi

Duygu analizi modülünde 2 aşamalı eğiticili makine öğrenmesi tekniği kullanılmıştır. Birinici adımda Multi Nominal Naive Bayes algoritması kullanılırken, ikinci adımda farklı sınıflandırma algoritmaları kullanılmıştır. Eğiticili makine öğrenmesi tekniklerini kullanılırken bir özellik vektörüne ihtiyaç duyulmaktadır. Bu sebeple duygu analizi modülünde sözlük tabanlı yöntem ve n-gram yöntemi bir arada kullanılarak özellikler elde edilmiştir. Bu iki yöntemin bir arada kullanılması hem özellik vektörünün daha kapsayıcı olmasını hem de başarının arttırılmasına yönelik tercih edilmiştir. Lakin bu yöntemlerden herhangi biri tek başına sonuca direk etki edememektedir. Bu yöntemlerden elde edilen değerler makine öğrenmesi teknikleriyle birleştirilerek kompleks bir duygu analizi sistemi tasarlanmıştır.

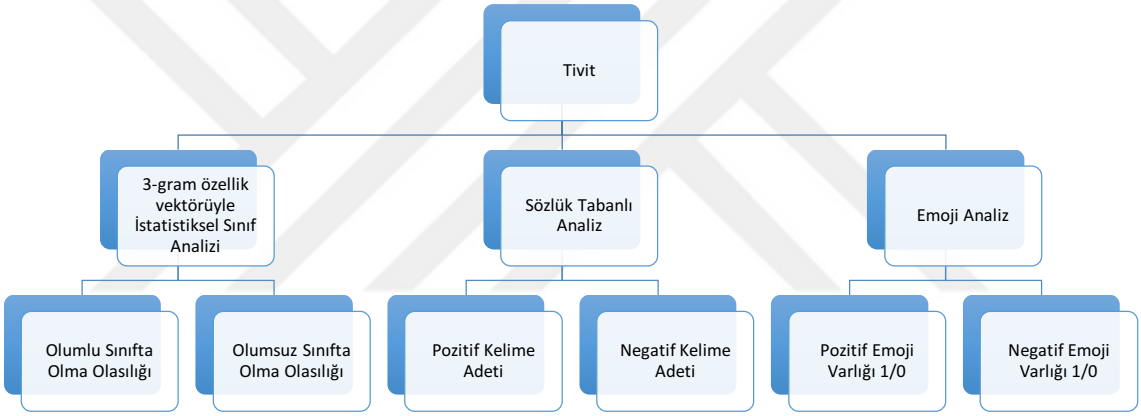
Ön işlem modülünde kelimelere ayrılan tivitler öncelikle 3-gramlarına ayrılarak özellik vektörü elde edilmiştir. Yapılan denemelerde en yüksek başarının 3-gram olarak ayrılmış özellik vektöründen alınmış olmasından dolayı 3-gramlar kullanılmıştır. Kelimeleri 3-gramlarına ayırmak için "Apache Lucene" kütüphanesinden [27] yararlanılmıştır. 3-gramlarına ayrılan kelimeler ile özellik vektörü çıkarılmış ve bu özellik vektörüne ayrıca kelimenin parçalanmamış orijinal hali de bir özellik olarak eklenmiştir. Bunun sebebi sözlük tabanlı yöntemler uygulanırken kelimenin orjinalinin bulunması gerekliliğidir. Bu özellik vektörü Multinominal Naive Bayes sınıflandırıcısıyla eğitilmiştir. Bu aşama için elde edilen özellik vektörü içinde toplam **28.293** adet özellik bulunmaktadır. Duygu analizi adımında elde edilen özellik vektörüne göre sınıflandırma yapılmış ve bir tivit için olumlu ve olumsuz sınıfta olma olasılığı elde edilmiştir. Bu olasılıklar bir sonraki adımda kullanılmak üzere yeni özellik vektörünün iki farklı özelliğini oluşturmaktadır.

Olumlu ve olumsuz kelime sayılarını elde etmek amacıyla sözlük tabanlı bir yöntem kullanılmıştır. Bunun için argo ve olumsuzluk içeren negatif kelimeler sözlüğü oluşturulmuş ve yine aynı şekilde bir de pozitif anlam içeren olumlu kelimelerden oluşan bir pozitif kelime sözlüğü daha oluşturulmuştur. Oluşturlan sözlük, düzenli ifadeler de barındırabilmektedir. Sözlükte bulunan **güzel** kelimesi tivitler içinde farklı varyasyonlarda bulunabilir. Örneğin; **güzeller**, **güzellik** ve **güzelim** gibi kelimeler de güzel kelimesinden türeyen kelimelerdir. Dolayısıyla tarama işlemi gerçekleştirilirken çekim eki almış hallerinin de göz önünde bulundurulması gerekmektedir. Bu sebeple sözlükte **güzel$** ifadesi bulunarak, güzel kelimesinin tüm varyasonları kapsanmış olmaktadır. Böylece tam eşleşen (exact match) ifadeler ile beraber bu kelime ile başlayan diğer türev kelimeler de bulunmuş olur. Oluşturulan sözlük iki şekilde de kelimeler içermektedir ve sistem düzenli ifadelerle oluşturulan sözlüğe izin vermektedir. Bu adımda tivitler her bir kelime için sözlükte taranmakta, sözlükte bulunan kelime sayıları elde edilmektedir. Elde edilen pozitif ve negatif kelimeler sözlüğünde bulunma sayıları ile birlikte, iki özellik daha elde edilmiş olacaktır.

Son iki özelliği elde etmek amacıyla emojiler incelenmiştir. Emojilerin bir metnin ana duygusuna etkisi [7] ilgili çalışmada belirtilmiştir. Bu özelliklerin elde edilmesinde yapılan varsayım olumlu veya olumsuz emojinin sayısının çokluğunun ana duyguya herhangi bir etkisi olmamasıdır. Bu sebeple emoji sayısı yerine emoji varlığı özellik olarak

kullanılmıştır. Bu sebeple elde edilen bu iki özellik ikilik (binary) düzende kodlanmıştır. Tarama işlemi yapılırken bir listede tutulan olumlu ve olumsuz duygu ifade eden bu emoijler, tivit içindeki varlığı kontrol edilerek bu iki özellik elde edilmiştir. Bunun iki ayrı özellik olarak kullanılması bazı tivitlerin hem olumlu hem olumsuz duygu barındıran emojiler içermesinden kaynaklanmaktadır.

Birinci ve ikinci adımda yapılan çalışmalar sonucunda toplam 6 özellik elde edilmiştir. Birinci adımdan elde edilen 2 özelliğin yanına ikinci adımda da ek olarak yeni 4 özellik daha elde edilmiştir. Bu özellikler Şekil 3.3'de görüldüğü üzere **n-gramlara göre olumlu sınıfta olma olasılığı**, **n-gramlara göre olumsuz sınıfta olma olasılığı**, **olumlu emoji varlığı**, **olumsuz emoji varlığı**, **pozitif kelime sayısı**, **negatif kelime sayısı** olmak üzere toplam 6 özellik üzerinden eğiticili makine öğrenmesi tekniği kullanılmıştır.



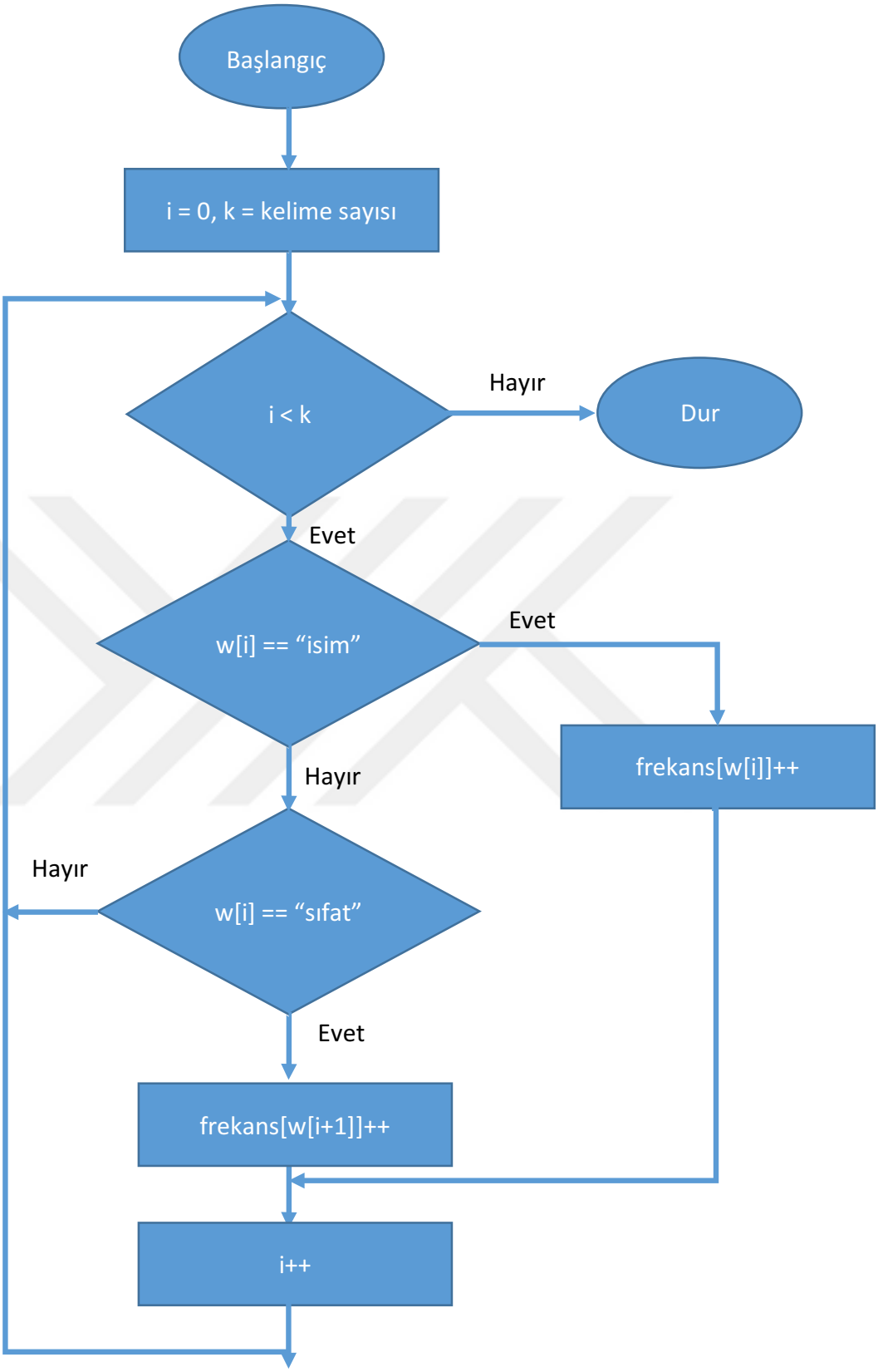
Şekil 3.3 Duygu Analizi Özellik Vektörünün Elde Edilmesi

3.4 Nitelik Çıkarımı

Nitelik çıkarımı modülünde duygu analizi gerçekleştirilecek olan tivitlerin **terim varlığı** ve **frekans analizi** yapılmıştır. Bu modüle veriler, sınıflandırılması yapılmış halde iletilmektedir. Olumlu ve olumsuz sınıfa ait bu tivitler içinde bulunan kelimeler analiz edilmektedir. Yapılan analiz neticesinde kelimenin türlerine bakılarak sıfat ve isim olan kelimelerin frekansları çıkarılarak, ilgili sınıfta o kelimenin tekrar sayısı bulunarak frekans analizi yapılmaktadır. Bu adımda sıfat olan kelimeler, sıfatların kendisinden sonra gelen ismi niteler prensibi doğrultusunda, bir sonraki kelimenin frekansına etki etmektedir. Örneğin; güzel tarife kelimesi tarifeyi nitelendirdiğinden tarife kelimesi önem arz edecek şekilde frekansı artırılmıştır. Şekil 3.4'de bu işlem adımları detaylı olarak gösterilmiştir.

Bunun yapılmasındaki amaç, sıfatların tek başına bir anlam ifade etmediği ancak bir şeyin olumlu ve olumsuz yönünü nitelendirmesinde önemli rol oynamasından kaynaklanmaktadır. Örneğin; insan tek başına kesin bir yargı ifade etmezken aynı şekilde iyi kelimesi tek neyi nitelendirdiği bilinmeden bir yargı yapılması söz konusu değildir. Ancak, "**iyi insan**" veya "**kötü insan"** şeklinde bir ifade, sıfatın insanı nitelediği için insanın hakkında bir yargı bildirmekte ve o insanın niteliğini belirtmektedir. Bu bağlamdan yola çıkarak bir şeyin niteliklerinin hangi özelliklerde toplandığını, olumluluk ya da olumsuzluğu niteleyen kısımların tespit edilmesinde niteliklerin önemli bir rolü bulunmaktadır. Yapılan frekans analizi sonucunda bu nitelikler belirlenmeye çalışılmıştır. Her iki sınıf için de aynı işlem gerçekleştirilmiş ve elde edilen frekansı en yüksek 10 kelime, karşı sınıftaki frekansları da incelenerek kendi bulunduğu sınıftaki konumu ve karşı sınıftaki konumu analiz edilmiştir. Bu işlem sonucunda elde edilen veri tivitlerin ait olduğu kullanıcı için hangi taraflarının beğenildiği veya beğenilmediğini göstermektedir.

Bu çalışmada tasarlanan sistem elde edilen Twitter verilerinin günlük olarak analiz edilmesine dayanan bir arkaplan uygulaması olarak tasarlanmıştır. Analiz uygulamaları tasarımı gereği veri ambarları üzerinden çalışmaktadırlar, çünkü anlık veriler sürekli değişebilen kaynaklar olduğu için ancak veriler sabit ve kararlı bir hale geldiğinde analiz edilebilmektedir. Bu sebeple kurgulanan tasarım, verilerin bir Excel veya başka bir formatta sisteme aktarılması ön görülerek tasarlanmıştır. Şekil 4.1 de görüldüğü üzere sistem iki temel uygulama ve bu uygulamaların içindeki modüllerden oluşmaktadır.



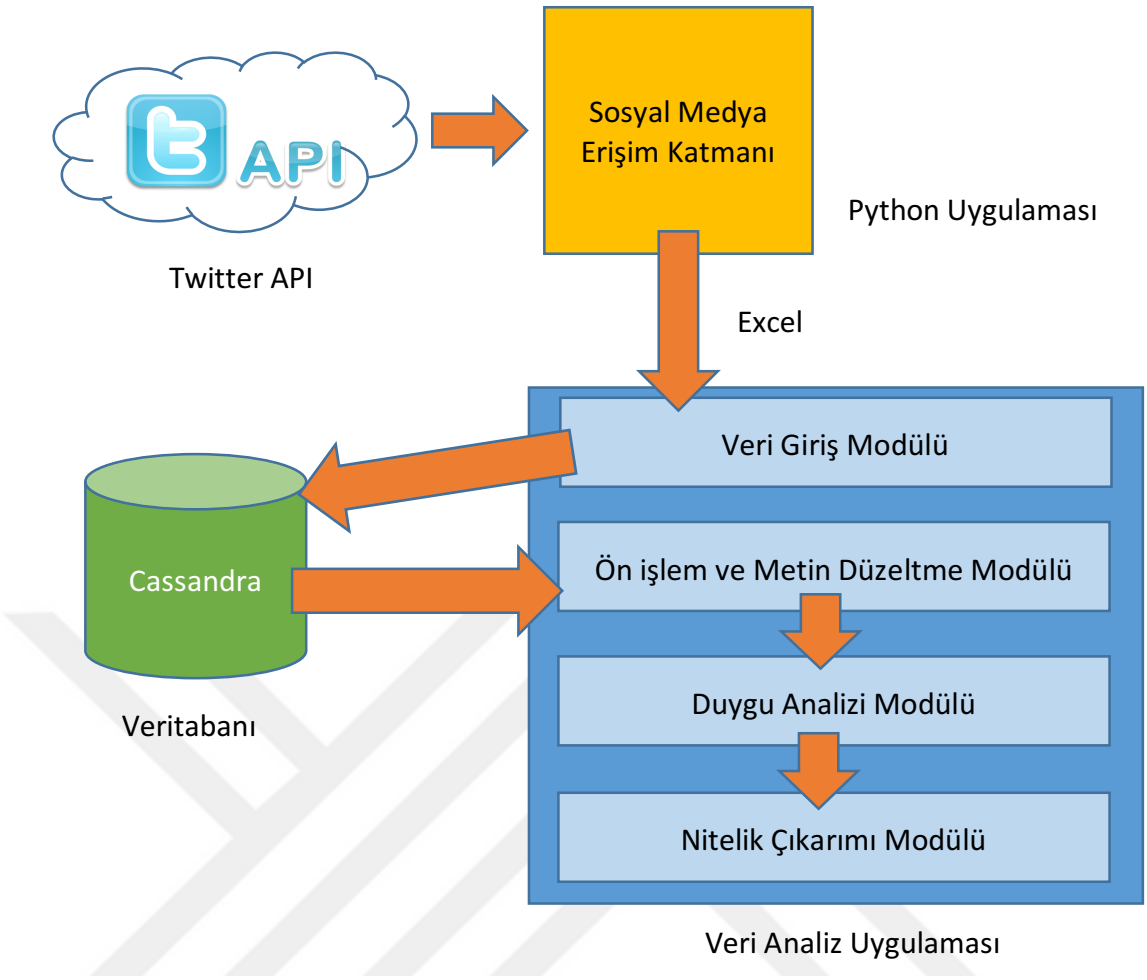
Şekil 3.4 Nitelik Analizi Algoritma Akışı

Sistem için gereken girdi ve bu girdilerin sonucunda çıktılar bu şekilde elde edilmiştir.

BÖLÜM 4

SİSTEMİN TASARIMI VE UYGULAMA

Bu çalışmada tasarlanan sistem elde edilen Twitter verilerinin günlük olarak analiz edilmesine dayanan bir arkaplan uygulaması olarak tasarlanmıştır. Analiz uygulamaları tasarımı gereği veri ambarları üzerinden çalışmaktadırlar, çünkü anlık veriler sürekli değişebilen kaynaklar olduğu için ancak veriler sabit ve kararlı bir hale geldiğinde analiz edilebilmektedir. Bu sebeple kurgulanan tasarım, verilerin bir Excel veya başka bir formatta sisteme aktarılması ön görülerek tasarlanmıştır. Şekil 4.1 de görüldüğü üzere sistem iki temel uygulama ve bu uygulamaların içindeki modüllerden oluşmaktadır.



Şekil 4.1 Uygulama Katmanları ve Modülleri

Sistemin bu şekilde tasarlanması her türlü veri formatında işlem yapılmasına olanak vermektedir. Veri giriş modülünde gerekli düzenleme yapılarak uygun veri formatında çalışması sağlanabileceği gibi bu modül düzenlenerek direk Twitter API üzerinden canlı veri alabilecek duruma da getirilebilmektedir. İlgili modül soyutlaştırma da (abstraction) bu aşamada tasarlanmıştır.

Gerçekleştirilen uygulama bir web uygulaması değildir ve bu yüzden herhangi bir web sunucuya veya konteynıra ihtiyaç duymadan çalışabilmektedir.

Geliştirilen uygulama Cassandra veritabanını kullandığı için yatay olarak sunucu eklenerek ölçeklendirilebilir ve dağıtık hesaplama yapılmasına da olanak sağlar. J2EE teknolojileri ile gerçekleştirildiği için platform bağımsız bir uygulama elde edilmiştir. Herhangi bir işletim sistemi üzerinde analiz gerçekleştirebilecek alt yapıya sahiptir.

4.1 Sosyal Medya Katmanı

Sosyal Medya Katmanı Python kullanılarak yazılmıştır. Python küçük uygulamalar için oldukça hızlı ve basit geliştirilebildiğinden tercih edilmiştir. Twitter API'sinin arama (search) metoduna ilgili hesap ismi üzerinden istek yapılarak veriler çekilmiştir. Ancak, alınan veriler tamamen bu hesaba ait veriler olmamaktadır. Twitter API arama metodu kullanıldığından, ilgili kelimenin geçtiği tüm tivitlerdönmektedir. Bu yüzden elde edilen tivitlerde "@hesap ismi" nin varlığı kontrol edilerek bu tivit veri setine eklenmiş, ilgili hesap ismi barındırmayan tivitler çıkartılmıştır. Ayrıca, sorgu yapılırken retweet'ler çıkartılarak sorgu yapılmıştır. Çünkü RT içeren tivitler, aynı veri setinden aynı tivitin birden fazla bulunmasına sebep olmaktadır. Twitter API limitleri dolayısıyla tivitler 100 adetlik parçalar halinde çekilmiştir. Bir sonraki sorgu da son tivitin id si "max_id" parametresine verilerek, o tivitten itibaren veriler elde edilerek her bir hesap için yaklaşık 850 civarında tivit veri setinde eklenmiştir. Çekilen bu tivitler Excel formatında dosyaya aktarılarak kaydedilmiştir. Excel içinde bulunan örnek veriler Çizelge 4.1 deki gibidir. Bu veriler veri analizi uygulamasına aktarılarak analiz ile elde edilmiştir.

Çizelge 4.1 Örnek Tivit Verisi Çıktısı

id	created_at	text	hashtag
676362586540126000	Mon Dec 14 11:26:21 +0000 2015	Halen Türkcell kullanmayanlar için @Turkcell https://t.co/HU4TjXAPfR	turkcell-filter:retweets
676357960411558000	Mon Dec 14 11:07:58 +0000 2015	@duetduello OMBUDSMAN 'a EVRAK da iletebiliyorsun derdini de .. Seninkine benzer bizim haklı olayımızda dümdüz ettiler bu @Turkcell 'İ..	turkcell-filter:retweets
676356357663101000	Mon Dec 14 11:01:36 +0000 2015	@duetduello Kardeşim sen 2 ayda pes etmişin.. 26 aydan beri mücadele edenler var.. Hem de mahkeme kararına rağmen @Turkcell @TurkcellHizmet	turkcell-filter:retweets

4.2 Veri Analiz Uygulaması

Veri Analiz uygulaması **J2EE** teknolojileri kullanılarak yazılmıştır. Bağımlılıkların çözümü için **Maven** paket yöneticisi kullanılmıştır. Eklenen kütüphaneler Maven aracılığı ile eklendiği için başka bir paket bağımlılığı varsa onların otomatik olarak eklenmesi sağlanarak ileride uygulamanın daha kompleks bir yapıya getirilmesi kolaylaştırılmıştır.

Yapılan uygulama için herhangi bir arayüz tasarlanmamış, direk olarak dosya üzerinden analiz yapılarak, konsol ekranına analiz sonucunun gösterilmesi sağlanmıştır.

4.2.1 Veri Giriş Modülü

Veri Giriş modülü Excel olarak kaydedilmiş verilerin okunarak veritabanına aktarılması görevini yerine getiren modüldür. Bu modül "Apache POI" kütüphanesi [26] kullanılarak gerçekleştirilmiştir. Bu kütüphane ile Excel dosyasındaki tivitler okunmuş ve sistemde tanımlanmış olan Tweet sınıfına bu tivitler atanmıştır. Elde edilen bu tivit objeleri daha sonra ORM aracı yardımıyla veri tabanına kaydedilmiştir. ORM [28] verilerin objelerle eşleştirilmesini ve veri tabanından objelerle sorgulanmasını sağlayan bir yöntem ve tasarım prensibidir. Bunun için Cassandra ORM olarak DataSlax firmasının geliştirmiş olduğu Java veritabanı sürücüsü [29] kullanılmıştır. Veri tabanına veriler aktarılırken otomatik tekil bir id üretilerek aktarılmış, ayrıca veri tabanında her bir tivit için sınıfını (olumlu / olumsuz) gösteren bir tip alanı da tutulmaktadır. Bu alan ikilik (binary) değer yani 1 veya 0 değerini almaktadır. Eğer tivitlerin sınıfları belli değil ise boş olarak kaydedilip, analiz modülü tarafından sınıflandırılması sağlanmıştır.

4.2.2 Ön işlem ve Metin Düzeltme Modülü

Ön işlem modülünde sistemdeki tivitler duygu analizi yapılmadan önce ön işlemden geçirilmiş ve metin düzeltme sağlanmıştır. Bu adımda Şekil 3.2'de bahsedilen ön işlemadımları uygulanmıştır.

Metin düzeltme adımında **2.105.089** adet kelimeden oluşan derlem üzerinden tarama yapılarak düzeltme işlemi gerçekleştirilmiştir. Bu modülde uzaklık analizi için iki önemli algoritma kullanılmıştır. Bunlar **Leveinstein** ve **Jaro Winkler** algoritmalarıdır. Metin düzeltme için gerçekleştirilen algoritma adımları şöyledir;

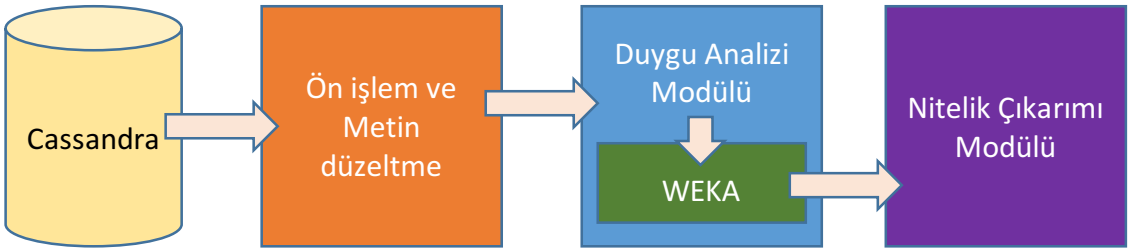
- Kelime iki harften küçük ise işlem yapma
- Kelimenin ilk harfi büyük harf ise işlem yapma(özel isim varsayımı)
- Kelime sayılardan oluşuyorsa işlem yapma
- Kelimeyi derlemdeki tüm kelimelerle karşılaştır, Levenshtein uzaklığı kelimenin uzunluğundan %51 daha büyük ise o kelimeyi benzer olarak algılama

- Kelimenin ve bir sonra gelen kelimenin derlemdeki yerine bak frekansa göre sıralı listeye ekle
- Eklenen kelimeleri Jaro Winkler uzaklığına göre karşılaştır uzaklığı **0.2** den küçük ise çıkar ve kalanlardan frekansı en yüksek olan ile benzer olarak değiştir

Bu algoritma adımlarında kullanılan parametreler testler sonucunda elde edilmiş, en yüksek başarının elde edildiği parametrelerdir. Uygulanan bu algoritma sonucunda kelimeler en yakın benzerlikteki kelimeler ile değiştirilmiştir. Bu işlem derlemdeki kelimeler çok fazla olduğu için uzun sürmektedir. Ortalama bir tivit için 30 saniye kadar süreye ihtiyaç duyulmaktadır. Bu işlemin getirdiği olumsuz taraf kullanılan derlemin köşe yazılarından oluşmasından kaynaklı olarak argo kelimeler ve sosyal medya jargonunu içermediğinden başarıyı az da olsa olumsuz yönde etkilemesidir.

4.2.3 Duygu Analizi Modülü

Duygu Anlizi Modülü verilerin ön işlem ve metin düzeltme modülünden geçtikten sonra iletildiği modüldür. Bu modülde eğitim seti verileri ile sistemin doğruluğunu test etmek amacıyla 10 katlı çapraz doğrulama yapılarak test edilmiştir. Bu doğrulama işlemi için WEKA kütüphanesi [23] kullanılmış, WEKA kütüphanesinin [23] üzerinde bulunan çeşitli algoritmalar ile denemeler yapılmış ve en başarılı olan algoritmalar seçilmiştir. Verilerin nasıl analiz edilği ve özellik setinin çıkarılmasına ait bilgiler Bölüm 3'te detaylı bir şekilde anlatılmıştır. Eğitim seti ile eğiticili öğrenme gerçekleştirildikten sonra test için üç telefon operatörüne ait tivitler ile test işlemi gerçekleştirilmiştir. Bu modülde duygu analizi iki sınfa göre yapılarak analiz gerçekleştirilmiştir. **İki sınıf kullanılmasının amacı bir konu hakkında beyan edilen fikrin ancak olumlu ya da olumsuz bir yargı barındıracağı varsayımıdır**. Nötr bir yaklaşım için herhangi bir yorum beyan edilmeyeceği düşünülmüştür.



Şekil 4.2 Duygu Analizi Modül Yapısı

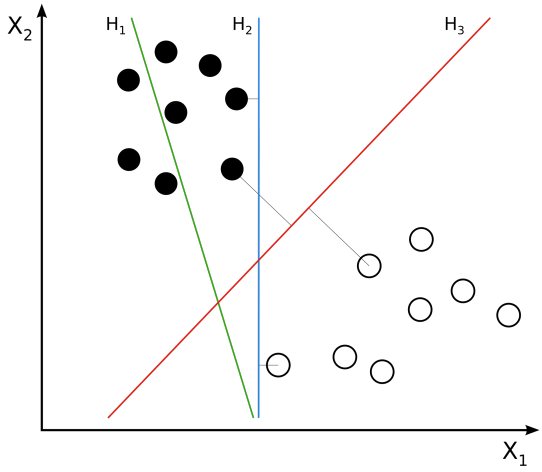
Duygu Analizi Modülünün diğer modüller ile iletişimi Şekil 4.2'de gösterilmiştir. Duygu Analizi için yapılan ön denemeler sonucunda en yüksek başarıyı veren 3 algoritma ile analiz işlemi yapılmıştır. Bu algoritmalar SVM, Logistic Regression ve Multi Layer Perceptron'dur.

Bildindiği üzere makine öğrenmesi eğiticili ve eğiticisiz olmak üzere ikiye ayrılmaktadır. Bu çalışmada kullanılan algoritmalar eğiticili öğrenme tekniklerinde kullanılan algoritmalardır.

4.2.3.1 Destek Vektör Makinesi (SVM)

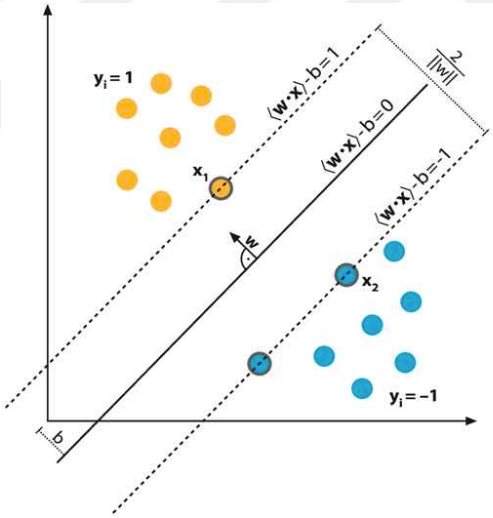
Destek Vektör Makineleri istatistiksel temele dayalı, oldukça yüksek başarı sağlayan bir eğiticili öğrenme algoritmasıdır. Doğrusal ve dağıtık olan verilerin her ikisinde de başarısı oldukça yüksek olmaktadır. Bunun en büyük nedeni farklı kernel tiplerine göre sınıflandırma yapabilmesinden kaynaklanmaktadır. Bu sebeple oldukça yaygın kullanım alanına sahiptir. Görüntü ve ses işleme alanlarından, veri madenciliğine kadar bir çok alanda çok yaygın bir şekilde kullanım alanına sahiptir. Bizim uygulamamızda da polinom kernel kullanılmıştır.

Destek Vektör Makinelerinin temel prensibi; sınıf sayısına göre sınıf elemanları arasında, bu sınıfları birbirinden ayıran sonsuz sayıdaki doğru içinden marjini en yüksek olan doğruyu seçerek sınıflandırma gerçekleştirme prensibine dayanır. Oluşan marjin doğrusu sınıf verilerinin bu doğruya en yakın örneklerine paralel olacak şekilde oluşur.



Şekil 4.3 SVM Sınıflandırma Örneği

Şekil 4.3'de görüleceği üzere H3 doğrusu sınıf verilerine eşit uzaklıktadır. Bu doğru, sınıf verilerinin birbirlerine eşit uzaklıkta olmaktadır ve bu doğru sınıf uc noktalarından geçen doğruya paralel olarak çizilen bir doğrudur. Bu sayede sınıf örneklerini bir birinden ayırır ve marjin olarak adlandırılır.



Şekil 4.4 SVM Marjinlerin Oluşumu ve Sınıflandırma

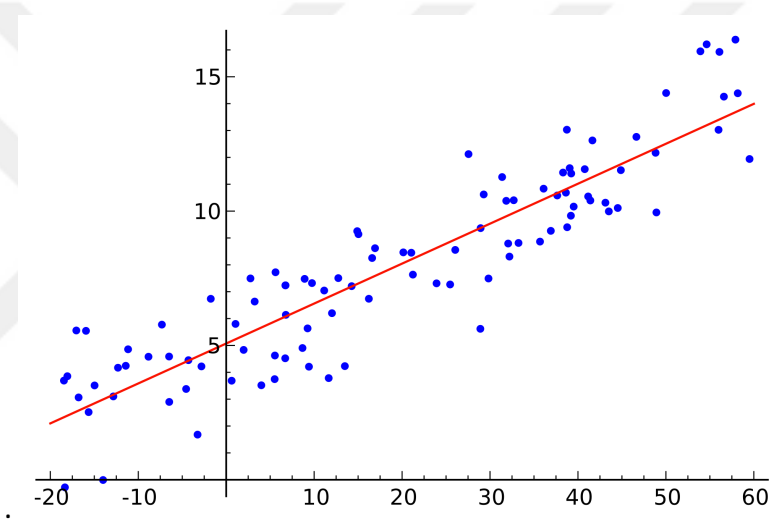
Şekil 4.4'de görüldüğü üzere sınıf elemanları arasındaki uzunluk $\frac{2}{\|w\|}$ şeklinde olurken sınıf elemanlarının karşı sınıfa yaklaştığı en yakın noktalarda oluşan doğrular $(w \cdot x) -$

$b = 1$ ve $(w \cdot x) - b = -1$ formüleriyle elde edilirken margin doğrusu $(w \cdot x) - b = 0$ denklemiyle elde edimektedir.

4.2.3.2 Lojistik Regresyon (Logistic Regression)

Lojistik Regresyon istatistikde kullanılan bir regresyon modelidir. İkili bağımlılığı bulunan bir bağımlı değişkeni tahmin etmek amacıyla birden fazla faktörden oluşan bir model yardımıyla yapılan istatistiksel bir analiz algoritmasıdır.

İki ya da daha fazla birbirine bağımlı parametrenin durumlarına bağlı bir şekilde oluşturulan bu model, bilinmeyen verilerin önceki verilere göre tahmin edilmesi prensibine dayanır.



Şekil 4.5 Lojistik Regresyon Modeli

Şekil 4.5'te olduğu gibi mavi ile gösterilen eğitim verileridir. Bu eğitim verileriyle eğitilen algoritma sonucunda, şekildeki gibi analitik düzlemdeki herhangi bir verinin x1,x2, xk değerleri için karşılık gelen y değerinin için oluşturulur. Bu veriler (4.1)'deki denkem ile elde edilir.

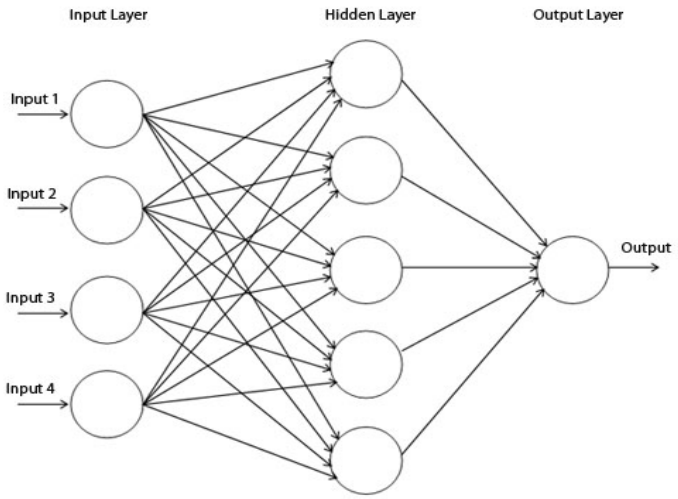
$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (4.1)$$

Tahmin işlemi bu denklem üzerinden yapılarak, Xn değerine karşılık gelen y değeri bulunur.

4.2.3.3 Çok Katmanlı Yapay Sinir Ağları (MultiLayer Perceptron)

MultiLayer Perceptron (MLP) ileri güdümlü yapay sinir ağı modelidir. Bu modelde girdilerin belli ağırlıklara göre bir ya da birden çok çıktı üretmesine göre sınıflandırma gerçekleştiren bir prensibe dayanan makine öğrenmesi algoritmasıdır.

MLP algoritmasında bir çok giriş ucu ve bir çok çıkış ucu bulunabilir. Girdi sayısına ve istenen çıktı sayısına göre arada gizli katmanlar bulunabilir. Buna Hidden Layer denir.



Şekil 4.6 Multilayer Perceptron Modeli

Şekil 4.6'da görüleceği üzere 4 girdi değeri için oluşacak tek çıktı için belli ağırlıklara göre bir ara gizli katman oluşturarak hesaplanır.

4.2.4 Nitelik Çıkarım Modülü

Nitelik Çıkarımı Modülü duygu analizi gerçekleştirilmiş verilerin frekans ve terim varlığı analizinin gerçekleştirildiği modüldür. Bu modül içinde sınıf içerisinde bulunan kelimelerin frekansları hesaplanır ve türlerine bakılır. Detaylı bilgi verilerin elde edilmesi aşamasında 3.bölümde bahsedilmiştir. Bu modülde kelimelerin türlerinin saptanmasında Zemberek [22] kullanılmıştır. Duygu analizinden gelen tivitlerin ait olduğu kullanıcının nitelikleri bu modülde analiz edildikten sonra çıktısı o hesabın niteliklerinin olumlu ve olumsuz sınıftaki dağılım şeklinden oluşur. Örnek çıktı Çizelge 4.2'de gösterilmiştir.

Çizelge 4.2 Örnek Nitelik Çıkarım Sonucu

Kelime	Olumlu Sınıftaki Oranı (%)	Olumsuz Sınıftaki Oranı (%)
kampanya	81	9
fiyat	70	30
merkez	19	71
dünya	18	82

Bu çıktıya göre kampanya kelimesinin frekansı olumlu sınıfta, olumsuz sınıftakinin 9 katı olmuştur. Bu da bize olumlu sınıftaki tivitlerde kampanyanın daha çok bulunduğunu göstermektedir. Analiz işlemi bu değerler göze alınarak yapılmıştır.

BÖLÜM 5

DENEYSEL SONUÇLAR

İşlenen verilerden elde edilen sonuçlar 3 başlık halinde sunulmuştur. Bunlar Metin Düzeltme, Duygu Analizi ve Nitelik Çıkarımı sonuçlarıdır.

5.1 Metin Düzeltme

Metin Düzeltme aşamasında Twitter'dan elde edilen veriler üzerinde metin düzeltme yapılmıştır. Oluşturulan sistemin başarısını ölçümlemede "Evaluating Evaluation Metrics for Spelling Checker Evaluations" adlı çalışmadaki [3] yöntem referans alınmıştır. Bir metin düzeltme uygulamasındaki başarı fm_o (Toplam F Ölçüşü) ve Overall Linguistic Performance (OLP) değerlerine göre ölçümlenmekte ve eşitlik (5.1) ve (5.2)'deki gibi hesaplanmaktadır.

$$fm_o = \frac{4}{\left(\frac{1}{R_c}\right) + \left(\frac{1}{P_c}\right) + \left(\frac{1}{R_i}\right) + \left(\frac{1}{P_i}\right)} \tag{5.1}$$

$$OLP = (fm_o * 0.667) + (SA * 0333) \tag{5.2}$$

(5.1) Formüldeki R_c ve R_i değerleri doğru (correct) ve hatalı (incorrect) sınıflandırılan veriler için **Recall** değeridir. Aynı şekilde P_c ve P_i değerleri sınıflandırmanın **Precision** değerleridir.

(5.2) denklemindeki OLP'nin hesabında kullanılan SA değeri ise (5.3)'deki gibi hesaplanmaktadır.

$$SA = \frac{\sum_{k=1}^{n} S_k}{N_{TN}}$$

(5.3)

(5.3) formülünde SA değerini bulmak için kelimelere karşılık olarak sistemin yapmış olduğu düzeltme önerileri için bir puanlandırma yapılmaktadır [20]. Bu puanlama sistemin kelimenin düzeltilmiş hali için önerdiği kelime ilk sırada ise **1 puan**, sistem doğru kelimeyi ilk sıra dışında herhangi bir sırada öneriyorsa **0,5 puan** olarak kabul . Eğer sistem yanlış bir öneri yapıyor ise **-0,5 puan**, hiç öneri yapmıyor ise **0 puan** olarak hesaplanmaktadır. Bu şekilde sistemin toplam puanı hesaplanmaktadir. Elde edilen bu değer True Negative olarak sınıflandırılmış kelimelerin sayısına (N_{TN}) bölünerek elde edilir. Sonuç olarak elde edilen SA değeri (5.2)'deki OLP değerinin hesaplanmasında kullanılmaktadır.

Kullanılan derlemde toplamda **2.105.089** adet kelime bulunur iken, aynı derlemdeki tekil kelime sayısı **233.116** adettir. Bu kelimelerin (5.1)'deki eşitliğe göre 100 adet tivit için doğrulaması gerçekleştirilmiş ve bu doğrulama sonuçları Çizelge 5.1'de verilmiştir.

Çizelge 5.1 Metin Düzeltme Aşamasında Test Tivitleri için Oluşan Hata Matrisi

100 Tivit için Karışıklık Matrisi		**Gerçek Veri**	
		Doğru	Yanlış
Tahmin Edilen	Doğru	901	36
	Yanlış	37	374

Bu işlem yapılırken önce fm_c ve fm_i değerlerinin hesaplanması gerekmektedir. Ancak, bu değerler sonucunda fm_o ve OLP değerleri hesaplanabilmektedir (5.2).

$$fm_c = \frac{2}{\frac{1}{R_c} + \frac{1}{P_c}}$$

(5.3)

$$fm_i = \frac{2}{\frac{1}{R_i} + \frac{1}{P_i}}$$

100 tivitin doğrulanması sonucunda elde edilen veriler Çizelge 5.2'deki gibi olmuştur.

Çizelge 5.2 Metin Düzeltme Sonucunda Elde Edilen Sonuçlar

100 Tivit	Sonuç
Toplam Kelime Sayısı	1348
Hata	%30.48
$\boldsymbol{R_c}$ **(Recall Correct)**	%96.05
$\boldsymbol{R_i}$ **(Recall Incorrect)**	%91.22
$\boldsymbol{P_c}$ **(Precision Correct)**	%96.15
$\boldsymbol{P_i}$ **(Precision Incorrect)**	%91.00
SA (Suggestion Adequacy)	%87.29
$\boldsymbol{P_a}$ **(Accuracy)**	%94.58
$\boldsymbol{fm_c}$	%96.1
$\boldsymbol{fm_i}$	%91.10

Elde edilen bu sonuçlara göre metin düzeltme sonucunda elde edilen veriler gösteriyor ki metin düzeltme modülü oldukça başarılı bir sonuç vermektedir. Sonuçlara bakıldığında fm_o değeri **%94.45** olurken OLP değeri ise **%92.06** olarak ölçülmüştür. Bu da metin düzeltme performansının oldukça iyi olduğunu göstermektedir.

Ayrıca "A Cascaded Approach for Social Media Text Normalization of Turkish" isimli makalede [4] kullanılan test veri seti ile geliştirilen model test edilmiş ve **%60.02** oranında başarı elde edilmiştir. İlgili makalede aynı veri seti için elde edilen başarı oranı **%71** olarak ölçülmüştür.

5.2 Duygu Analizi

Duygu Analizi için eğitim seti olarak **5.990** adet tivit kullanılmıştır. Bu tivitler hem metin düzeltme yapılarak hem de yapılmadan 10 katlı çapraz doğrulama ile test edilmiştir. Ayrıca, bu işlem sırasında SVM, Logictic Regression, Multilayer Perceptron ve bunların birleşimi olan Ensemble yöntem kullanılılarak sonuçlar karşılaştırmalı olarak verilmiştir. Bu algoritmaların seçimine WEKA [23] üzerinde bulunan farklı algoritmalarla test yapılarak karar verilmiş ve en yüksek başarıyı veren 3 algoritma seçilerek analiz işlemleri gerçekleştirilmiştir. Ensemble yönteminin denenmesine hata matrisine bakılarak karar verilmiştir. Çünkü kullanılan bu üç algoritma sonucundaki sınıflandırmalar farklı sonuçlar vermiş ve hiçbir algoritma diğerini kapsamamıştır.

SVM algoritması için polinom kernel kullanılmıştır. Bunun seçimine de diğer kerneller ile denemeler yapıldıktan sonra karar verilmiştir. En yüksek başarı polinomial kernel ile elde edilmiştir.

Çizelge 5.3 Metin Düzeltme Yapılarak SVM algoritmasıyla Hata Matrisi

Düzeltilmiş Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (SVM)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2667	323
	Olumlu	390	2610

Çizelge 5.4 Metin Düzeltme Yapılarak LR algoritması Hata Matrisi

Düzeltilmiş Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (LR)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2641	349
	Olumlu	361	2639

Çizelge 5.5 Metin Düzeltme Yapılarak MLP algoritmasıyla Hata Matrisi

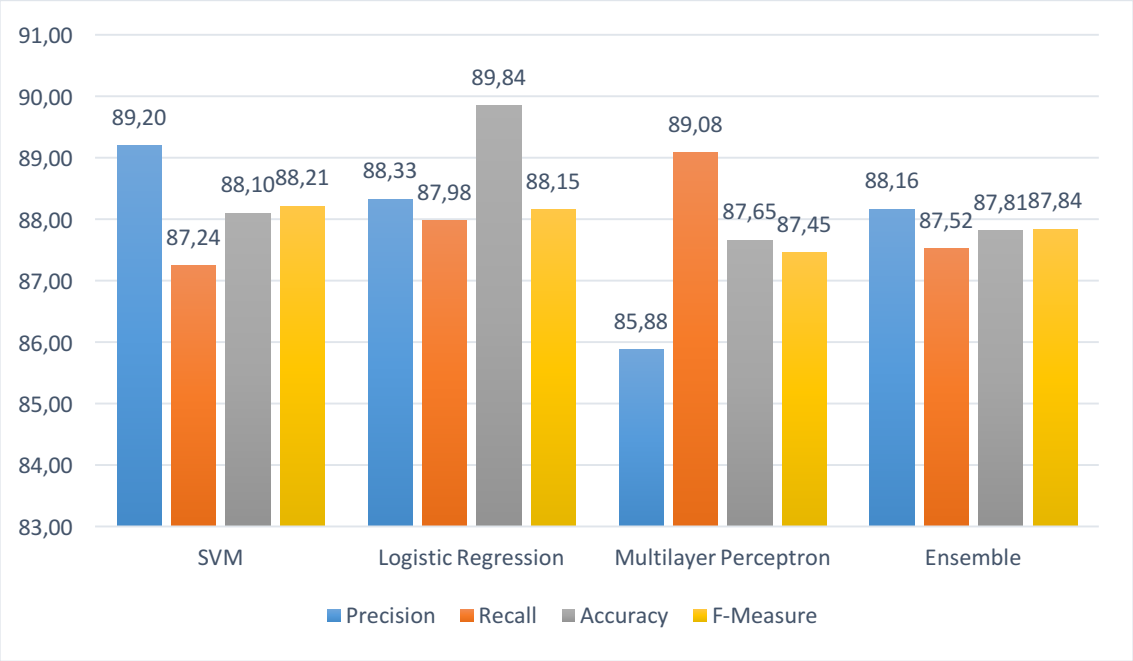
Düzeltilmiş Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (MLP)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2585	425
	Olumlu	317	2683

Çizelge 5.6 Metin Düzeltme Yapılarak Ensemble algoritması Hata Matrisi

Düzeltilmiş Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (Ensemble)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2636	354
	Olumlu	376	2624

Çizelge 5.3, Çizelge 5.4, Çizelge 5.5 ve Çizelge 5.6 incelendiğinde SVM algoritmasının en başarılı algoritma olduğu görülmektedir. Ancak, diğer algoritmalardaki sonuçlar incelendiğinde, bir kısmının olumsuz sınıf etiketlerinde başarı daha yüksek olurken,

olumlu sınıf etiketlerinde ise daha başarılı olduğu gözlemlenmiştir. Bu sebeple algoritmaların farklı yönlerini birleştirerek daha kapsamlı bir algoritma elde etmek amacıyla Ensemble yöntemi olan bu üç algoritmanın birleşimi de test edilmiştir.



Şekil 5.1 Metin Düzeltme ile Duygu Analizi Sonuçları

Gerçekleştirilen test işlemi sonucunda elde edilen veriler, Ensemble gibi bir yöntem kullanıldığı durumlarda bazı veriler SVM'e göre düşüş gösterse de, LR'nin Accuracy yüksek çıkması, ML'in de Precision değerinin yüksek çıkması böyle bir birleşim algoritmasının kullanılması ile daha verimli sonuçların alınacağını düşündürmüştür. Nitekim Ensemble algoritmasında Precision, Recall, Accuracy ve F-Measure değerlerinin bir birine yakınlığından daha kararlı bir algoritma olduğu çıkarımı yapılabilmektedir.

Çizelge 5.7 Metin Düzeltme Olmadan SVM Algoritması Hata Matrisi

Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (SVM)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2670	320
	Olumlu	387	2613

Çizelge 5.8 Metin Düzeltme Olmadan LR Algoritması Hata Matrisi

Eğitim Seti için 10 Katlı Çapraz Doğrulama Hata Matrisi (LR)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2634	356
	Olumlu	330	2670

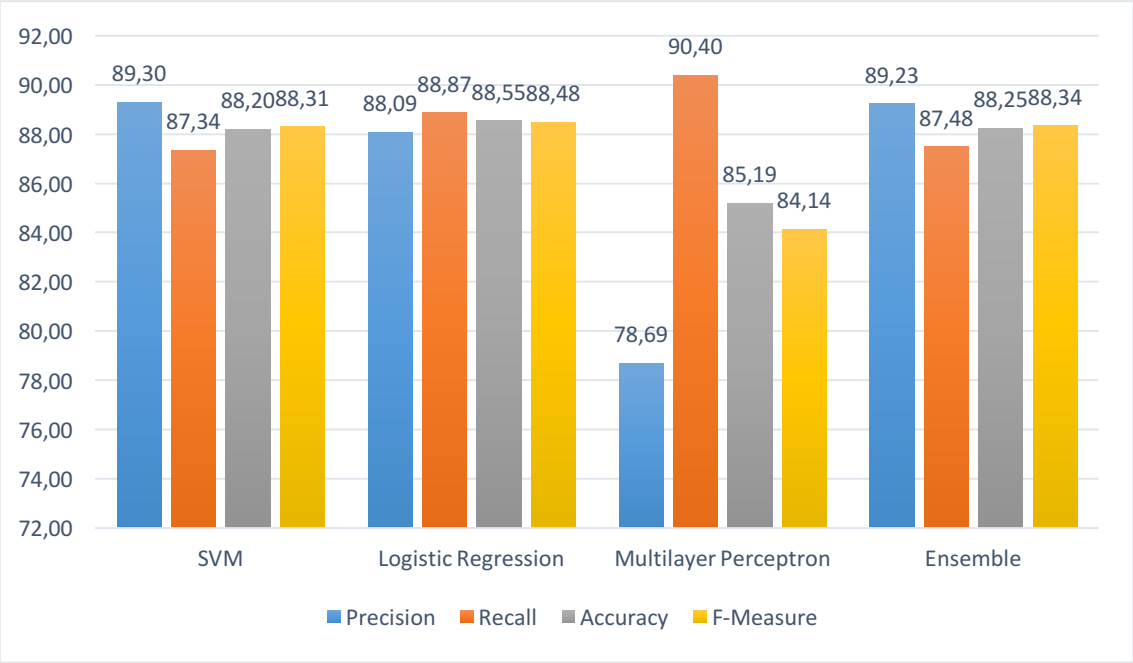
Çizelge 5.9 Metin Düzeltme Olmadan MLP Algoritması Hata Matrisi

Eğitim Seti için 10 Parçalı Çapraz Doğrulama Karışıklık Matrisi (MLP)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2353	637
	Olumlu	250	2750

Çizelge 5.10 Metin Düzeltme Olmadan Ensemble Algoritması Hata Matrisi

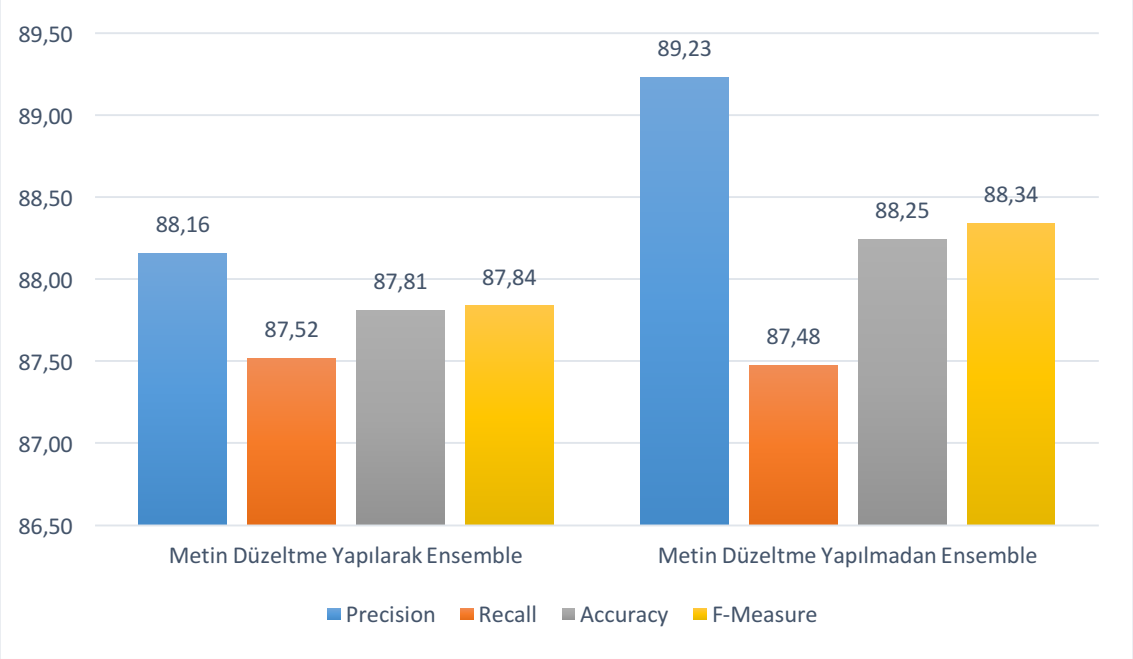
Eğitim Seti için 10 Parçalı Çapraz Doğrulama Karışıklık Matrisi (Ensemble)		**Gerçek Veri**	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	2668	322
	Olumlu	382	2618

Yine aynı şekilde metin düzeltme yapılmadan elde edilen algoritma sonuçları metin düzeltme yapılarak elde edilenler sonuçlar ile paralellik göstermektedir. Çizelge 5.7, Çizelge 5.8, Çizelge 5.9 ve Çizelge 5.10 incelendiğinde görüleceği üzere metin düzelme ile elde edilen sonuçlar ile örtüşmektedir.



Şekil 5.2 Metin Düzeltme Yapılmadan Elde Edilen Duygu Analizi Sınıflandırma Sonuçları

Metin düzeltme olmadan yapılan 10 katlı çapraz doğrulama sonucunda elde edilen verilerde metin düzeltme yapılarak elde edilenlere benzer sonuçlar alınmıştır. Toplamda elde edilen başarı, metin düzeltme adımına göre biraz farklılık göstermektedir.



Şekil 5.3 Metin Düzeltme Yapılarak ve Yapılmadan Elde Edilen Duygu Analizi Sınıflandırma Sonuçları Karşılaştırması

Şekil 5.3 incelendiğinde görüldüğü üzere metin düzeltme yapıldığında başarı düşmektedir. Bu iki sonuç doğrultusunda, metin düzeltmenin duygu analizine çok fazla

olmasada olumsuz olarak yansıdığı görülmektedir. Bunun sebebinin argo kelimelerin ve kısaltmaların düzeltme sonucunda bozulmasıyla ortaya çıktığı düşünülmektedir. Sosyal medyadan elde edilmiş verilerden oluşturulan bir derlem ile daha iyi sonuç elde edilebileceği varsayımı yapılabilir. Metin düzeltme aşamasında yapılan işlemin sonuca çok az etki etmesinin sebebi aynı kelimelerin hem eğitim setinde hem de test verisinde aynı şekilde değiştirilmesinden kaynaklanmasını söyleyebiliriz. Giriş değeri aynı olduğu sürece çıktı da hep aynı olmaktadır. Farklılığın bozulan kelimelerden kaynaklandığı düşünülmektedir. **Metin düzeltme işleminin maliyetli bir işlem olması sebebiyle nitelik çıkarımı verisi üzerinde metin düzeltme işlemi yapılmamıştır**.

Ayrıca, kullanılan algoritmaları kendi aralarında analiz ettiğimizde SVM'in diğer algoritmalara göre daha başarılı olduğu bir gerçektir. Ancak, Logistic Regression ve Multilayer Perceptron algoritmalarıyla elde edilen sonuçlar ile SVM algoritmasının olumlu ve olumsuz olarak etiketlediği tivitler birbirinden farklılık gösterdiğinden ötürü dördüncü bir method olarak bunların birleşiminden oluşan Ensemble yöntemi ile ilerlemeye karar verilmiştir. Ensemble yönteminde başarı SVM'e göre biraz düşüş gösterse de daha fazla ortak verinin doğru etiketlenmesi adına doğru bir çıkarım olduğu düşünülmektedir.

Bu aşamada kullanılan eğitim seti ile sistem eğitildikten sonra 3 gsm operatörüne ait veriler için duygu analizi gerçekleştirilmiştir. Bu operatörlere ait veriler herhangi bir yargı oluşturmamak adına X, Y, Z olarak etiketlenerek gösterilmiştir. Bu operatörlere ait veri seti bilgisi Çizelge 5.11'de verilmiştir.

Çizelge 5.11 Operatörlere Ait Tivit Sayıları

Operatör	Tivit Sayısı	#Olumlu Tivit	#Olumsuz Tivit
X	830	210	620
Y	874	281	593
Z	805	227	578

Operatörler için elde edilen tivit verilerinin duygu analizi sonucunda elde edilen hata matrisleri Çizelge 5.12, Çizelge 5.13 ve Çizelge 5.14 'te gösterilmiştir.

Çizelge 5.12 X Operatörüne Ait Sınıflandırma Hata Matrisi

X Operatörü Hata Matrisi (Ensemble)		Gerçek Veri	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	503	116
	Olumlu	58	153

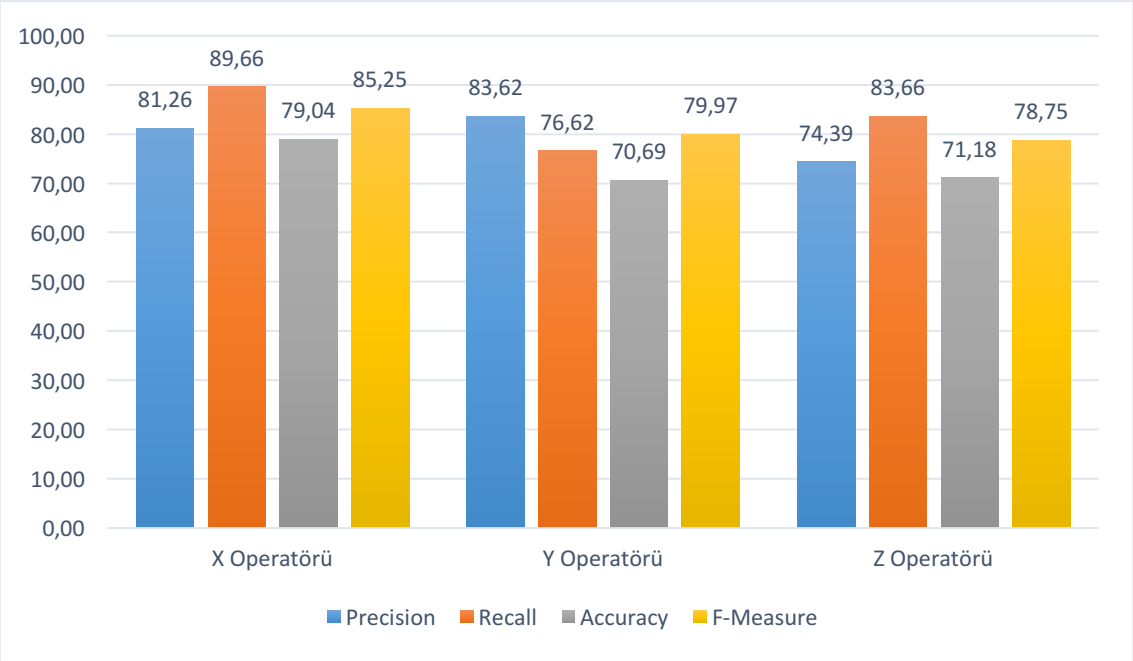
Çizelge 5.13 Y Operatörüne Ait Sınıflandırma Hata Matrisi

Y Operatörü Hata Matrisi (Ensemble)		Gerçek Veri	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	485	95
	Olumlu	148	101

Çizelge 5.14 Z Operatörüne Ait Sınıflandırma Hata Matrisi

Z Operatörü Hata Matrisi (Ensemble)		Gerçek Veri	
		Olumsuz	Olumlu
Tahmin Edilen	Olumsuz	430	148
	Olumlu	84	143

Elde edilen verilerdeki sınıflandırmanın birbirinden farklı oluşu verileri etiketleyen kişiler ile verilerin çekildiği aralıkta kullanıcıların göndermiş olduğu tivitlerin kalitesine göre değişmektedir. Bu alandaki farklılıkların sebebi veri setinden kaynaklanmaktadır.



Şekil 5.4 Operatörlere Ait Tivitlerin Sınıflandırma Başarıları

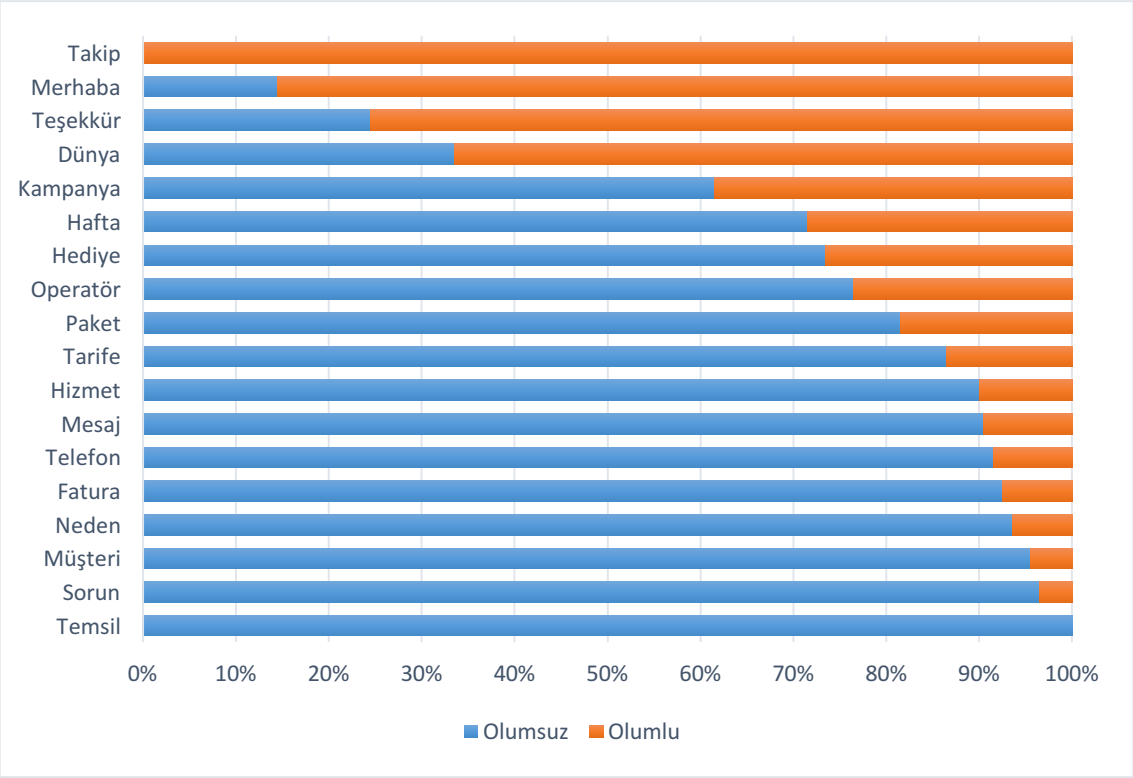
Şekil 5.4 görüldüğü üzere aynı eğitim seti kullanılmış verilerin az da olsa farklı sonuçlar vermesi verinin kalitesiyle doğru orantılı olduğundandır. Daha çok kişi tarafından etiketlenen veriler daha düzgün sınıflandırılabilmiştir.

Duygu analizinden elde edilen sonuçlar ile oluşan sınıflandırma dağılımları bu verilerin nitelik çıkarımına da direk yansımaktadır. Bu aşamada elde edilen verilerdeki sınıflandırma nitelik çıkarımı modülüne aktarılarak ilgili sınıfta hangi niteliğin o sınıfı nitelendirdiği belirlenmiştir.

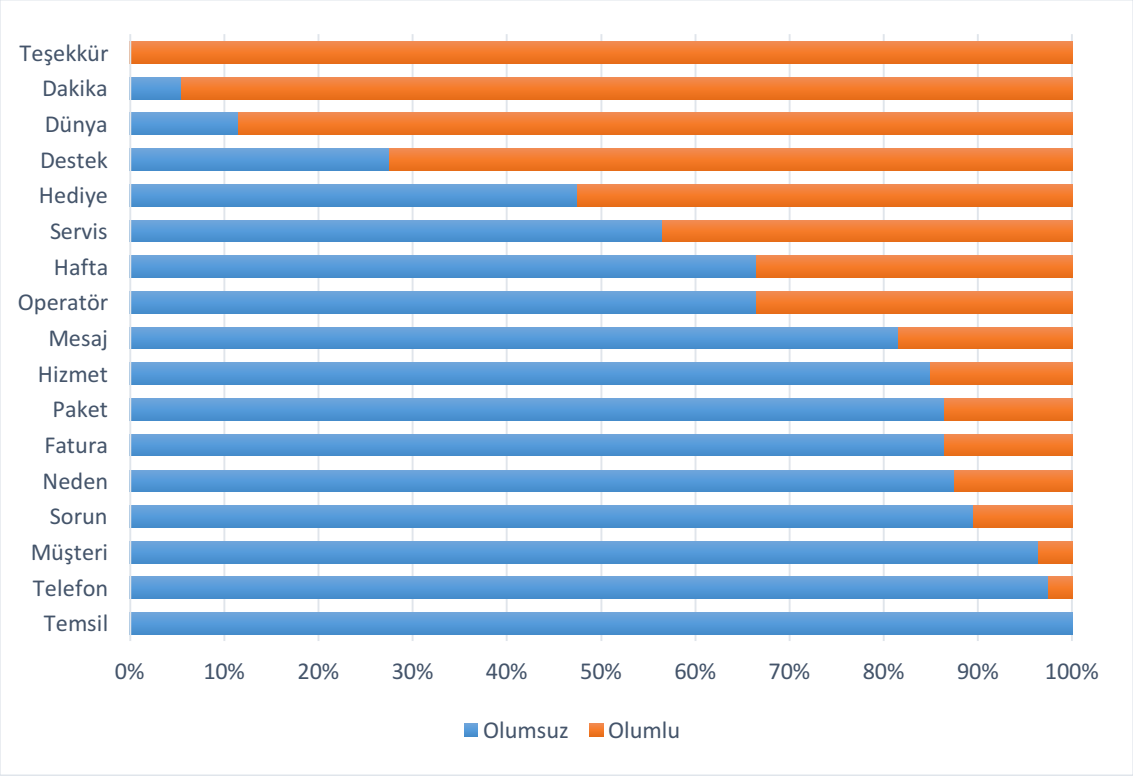
5.3 Nitelik Çıkarımı

Nitelik Çıkarımı için kullanılan veriler, Duygu Analizi modülünden elde edilen sınıflandırma doğrultusunda yapılmıştır. Bu adımda yapılan analiz ile ilgili gsm operatörünün verilerinin olumlu ve olumsuz sınıf içerisindeki frekans analizi gerçekleştirilmiştir. Bu analiz sonucunda hem olumlu hem de olumsuz sınıftaki en yüksek frekansa sahip ilk 10 kelime alınmıştır. Toplam kelime sayısı 20 olması beklenirken bazıları diğer sınıftada ilk 10 arasında yer aldığından kelime sayısı 20'ye yakın bulunmuştur. Sonuç olarak bu kelimelerin karşı sınıftaki frekanslarına da bakılarak yüzdelik olarak dağılımı elde edilmiştir.

X operatörü için elde edilen nitelik çıkarımı sonuçları Şekil 5.5'daki gibi olmuştur.



Şekil 5.5 X Operatörünün Etiketlenmiş Verilerden Elde Edilen Nitelik Çıkarım Sonuçları

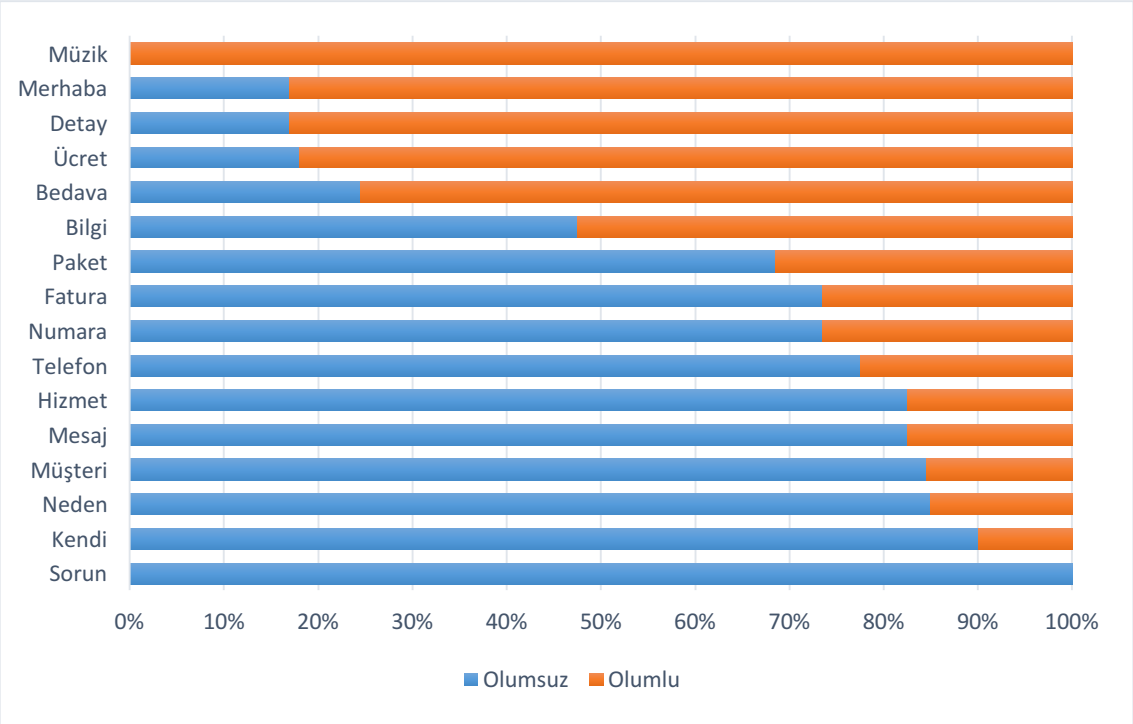


Şekil 5.6 X Operatörü Sistem Tarafından Sınıflandırma Sonucu Elde Edilen Nitelik Çıkarım Sonuçları

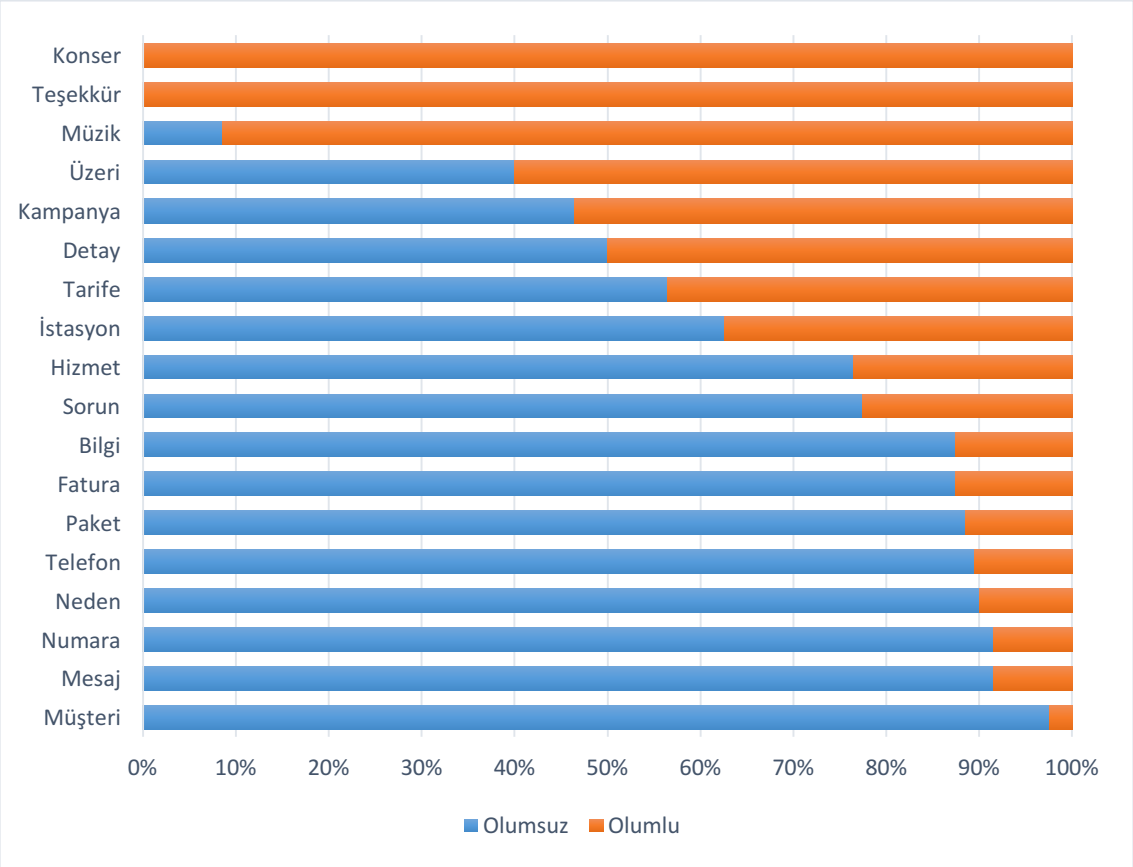
Şekil 5.5 ve Şekil 5.6 kıyaslandığında niteliklerin büyük kısmının bir biriyle aynı olduğu gözükmesine rağmen, oranlarında bazı farklılıklar göze çarpmaktadır. Böyle bir sonuç çıkmasında duygu analizi modülünde gerçekleştirilen sınıflandırmanın direk olarak etki ettiği gözükmektedir.

Ayrıca sistem tarafından yapılan sınıflandırma sonucunda Şekil 5.6 elde edilmiştir. X operatörü için elde edilen nitelik çıkarımı sonuçları gösteriyor ki; müşteri, hizmet ve temsil kelimesinde olumsuz tarafın bu kadar yüksek olması tivitlerdeki genel yargının **müşteri hizmetleri** ve **müşteri temsilcisi şikayetleri**, **müşteri kazanmak** ve **müşteri kaybetmek** ile ilgili tivitler olması dolayısıyla olumsuz yargı oluşmuştur. Olumlu yargının sebebi ise müşteri hizmetlerinin sorunu çözmesinden kaynaklı az sayıda tivitden dolayı olmaktadır. **Neden** kelimesine gelen bu kadar olumsuz yargı ise kullanıcıların şikayetleri için sordukları sorulardan dolayıdır. **Mesaj** konusundaki şikayetler, operatörün attığı mesajların zamansızlığı ve çokluğu, mesaj gönderememe gibi sorunlar sebebiyle oluşmuştur. **Servis** kelimesine gelen şikayetler, servis araçları, servislerin çekim gücü ve bazı web hizmetlerinden şikayetçi kişilerin atmış olduğu tivitlerden dolayı olmuştur.

Olumlu olan sınıfta **hediye** kelimesinin yoğunlaşması, operatörün **bedava internet** ve **mesaj paketleri** hediye etmesinden kaynaklanırken, olumsuz olan hediye tivitleri genelde hediyenin gelmemesi ya da hediye isteme gibi durumlardan dolayıdır. Operatörün destek hesabından insanların memnun kalması ve sosyal sorumluk projelerine yapılan **destek** dolayısıyla olumlu sınıfta bu yargının yüksek olmasını sağlamıştır **Dünya** kelimesine gelen olumlu tivitler ise kahve dünyası kampanyası dolayısıyla olmuştur. Hediye **dakika** dolayısıyla gelen olumlu tivitlere karşın, olumsuz olan dakika tivitleri, müşteri hizmetlerinin dakikalarca bekletmesinden kaynaklanmaktadır. Teşekkür tivitlerinin ise verilen hediyeler ve sorun çözümleri için olduğu görülmüştür.



Şekil 5.7 Y Operatörünün Etiketlenmiş Verilerden Elde Edilen Nitelik Çıkarım Sonuçları

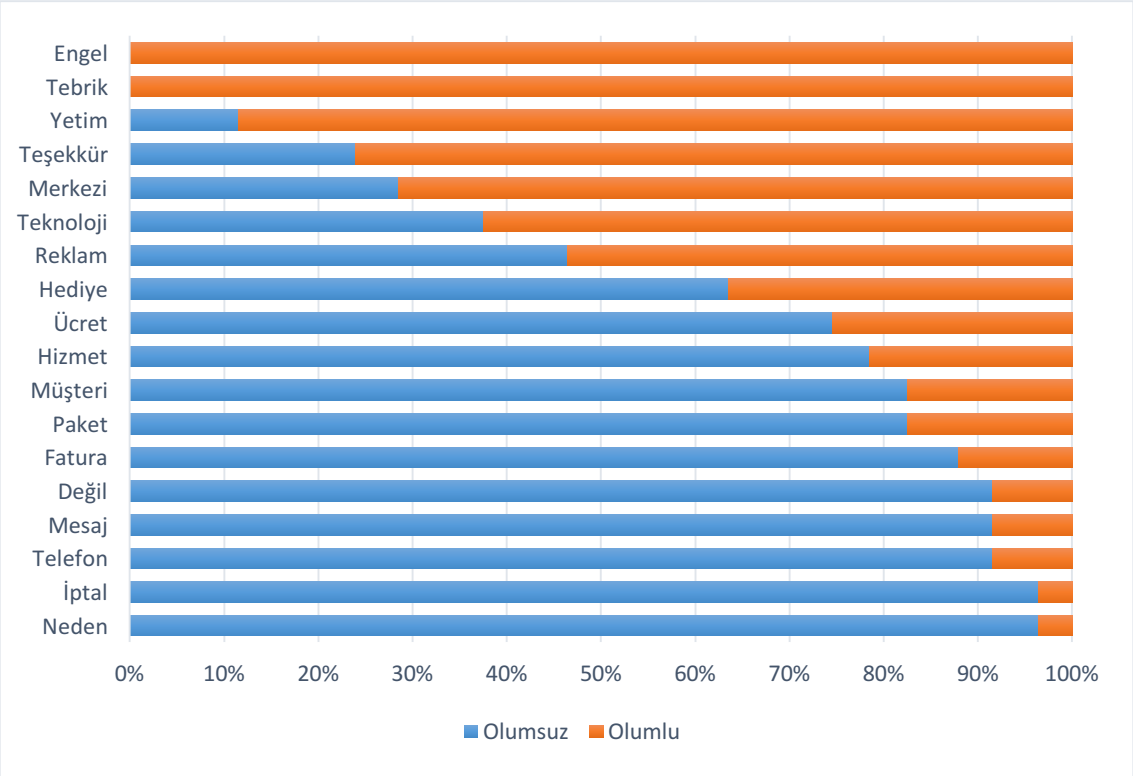


Şekil 5.8 Y Operatörü Sistem Tarafından Sınıflandırma Sonucu Elde Edilen Nitelik Çıkarım Sonuçları

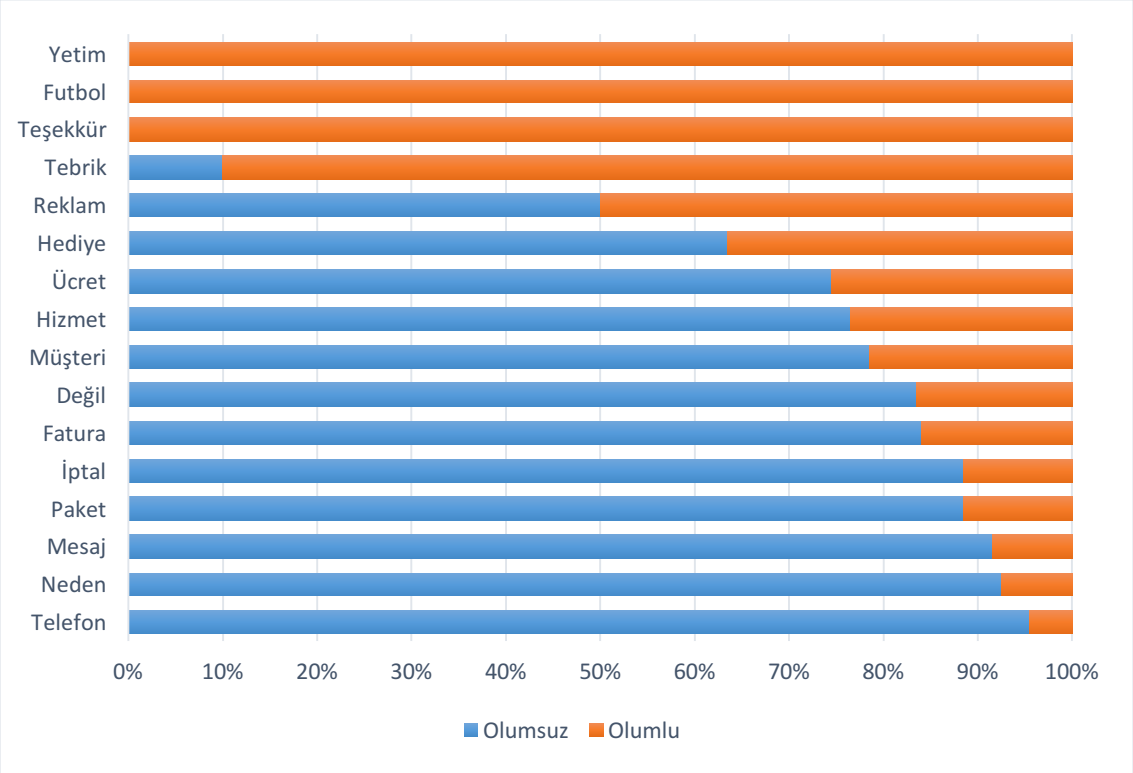
Şekil 5.7 ve 5.8 kıyaslandığında, Y operatörü için de aynı şekilde sistem tarafından sınıflandırılan veri ile etiketlenmiş veririn nitelik çıkarımı arasındaki sonucun X operatörüne göre bir birine daha çok yakınsaması bu operatör için yapılan sınıflandırmanın, X operatörüne göre daha başarılı olmasından kaynaklanmaktadır.

Ayrıca Şekil 5.8'da Y operatörü için elde edilen nitelik çıkarımı sonuçları göştermiştir ki; Y operatörü için **müşteri hizmetleri** ve **müşteri memnuniyetiyle** alakalı olumsuz görüşler dolayısıyla müşteri kelimesi bu kadar yüksek oranda olumsuzluk içermektedir. Aynı şekilde X operatöründe olduğu gibi zamansız ve yersiz **mesajlar** dolayısıyla çok yoğun şikayetler gelmektedir. **Numara** çok genel bir kelime olması dolayısıyla çok çeşitli konularda olumsuzluk içermektedir. Örneğin; **numara taşımadaki** sıkıntılar, **müşteri hizmetleri numarasının cevap vermemesi** gibi durumlar yüzünden olumsuz sınıfta yoğun bir şekilde bulunmaktadır. **İstasyon** kelimesi ile ilgili şikayetler **baz istasyonları** ve **çekim gücü** ile bağlantılı olarak ortaya çıkmıştır. Tarife kelimesine gelen olumlu tivitler ilgili opearatörde tarifelerin **ucuz** olması, olumsuz tivitlerde ise tarife değişiklikleriyle alakalı sorunlar yaşanması dolayısıyla olmuştur.

Olumlu yargı bildiren **müzik** kelimesi ise operatörün düzenlediği **müzik konserleri** ve **bilet hediyeleri** dolayısıyladır. Aynı şekilde **konser** kelimesinin yoğun çıkma sebebi de bu sebeple olmuştur.



Şekil 5.9 Z Operatörünün Etiketlenmiş Verilerden Elde Edilen Nitelik Çıkarım Sonuçları



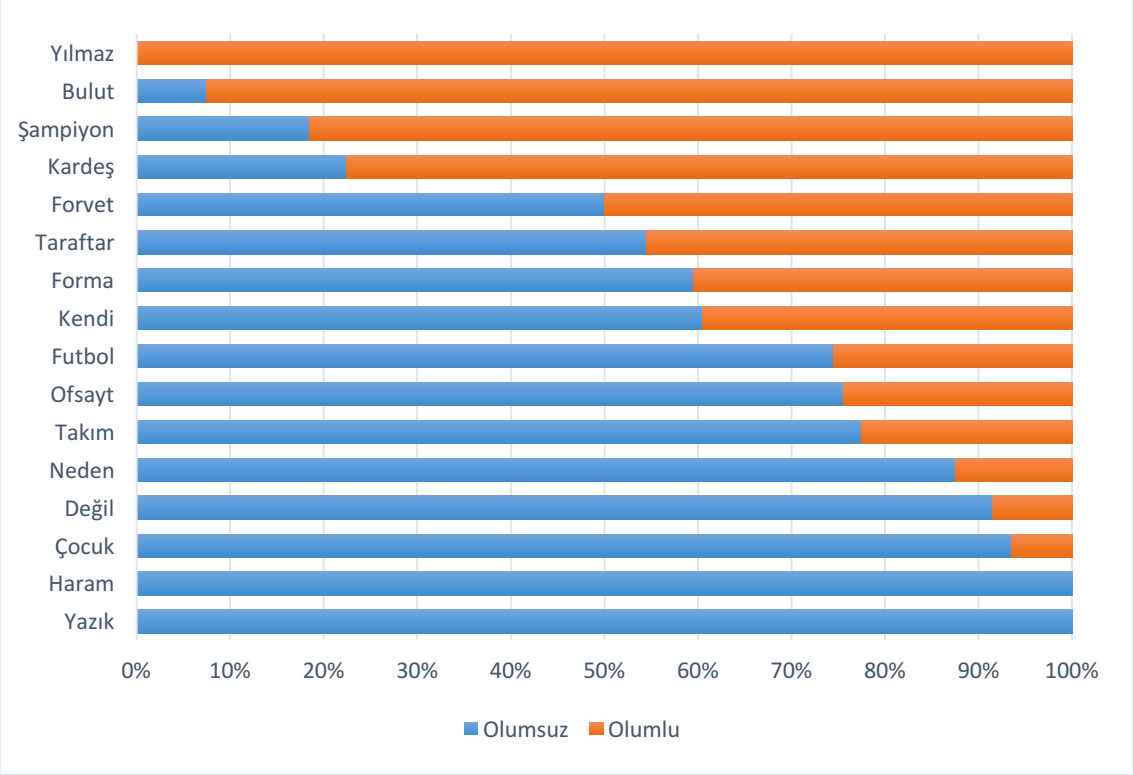
Şekil 5.10 Z Operatörü Sistem Tarafından Sınıflandırma Sonucu Elde Edilen Nitelik Çıkarım Sonuçları

Şekil 5.9 ve 5.10 incelendiğinde, diğer operatörlerin etiketlenmiş ve sistem tarafından sınıflandırılan verilerin nitelik çıkarım sonuçlarındaki farkların bu operatör de oldukça az olması, bu operatör için elde edilen duygu analizinin f-ölçüsünün %85 civarında olmasından kaynaklanmaktadır. Sonuç olarak bu operatörün duygu analizinde yapılan sınıflandırma diğerlerinden daha başarılı olmuştur. Bu operatör için sağlanan verilerin hem etiketlenmesi diğer operatörlere göre daha başarılı hem de sınıflandırması, etiketlenen verinin daha temiz olmasınından dolayı yüksek çıkmıştır. Bu da direk olarak nitelik çıkarım sonuçlarının orijinal veriden elde eden ile çok benzer olmasını sağlamıştır.

Ayrıca Şekil 5.10'da görüldüğü gibi Z operatörü için elde edilen nitelik çıkarımı sonuçları bize en yoğun kelime olan **telefon** kelimesi, cihaz kampanyaları ve telefon numarası gibi konularda olumsuzluk gösterirken, **paket iptali** veya **hat iptali** gibi durumlar için yaşanan sorunlar iptal kelimesinin bu kadar yoğun olmasına sebebiyet vermektedir. **Ücret** kelimesinin bu kadar yoğun olması ise operatörün beklenmedik ücretler yansıtması ve ücretlerin yüksekliği ile ilgili olumsuz tivitler dolayısıyladır. Olumlu ve olumsuz yoğunluğu eşit olan **reklam** kelimesi ise reklamların gerçeği yansıtmadığını söyleyen tivitler dolayısıyla olumsuz olurken, engelliler ve futbol için yaptığı reklamlar dolayısıyla olumlu tivitler almaktadır.

Futbol kelimesinin olumlu tivitlerde bu kadar yoğun olması, opertatörün **futbol** ne güzel şey isimli reklam kampanyası dolayısıyla gerçekleşmiştir. **Yetim** kelimesinin yine bu kadar olumlu tivit almasının sebebi, yetimlere yönelik oluşturulan sosyal sorumluluk projesinden kaynaklandığı görülmüştür.

K futbolcusuna ait nitelik çıkarım sonuçları Şekil 5.11'deki gibi olmuştur.



Şekil 5.11 K Futbolcusu için Sistem Tarafından Otomatik Sınıfılandırılmış Nitelik Çıkarım Sonuçları

K futbolcusu için elde edilen nitelik çıkarımı sonuçları ise söyledir. **Yazık** kelimesinin bu kadar yoğun olumsuzluk içermesi "**yazıklar olsun**" söz kalıbındaki isim olan kelimenin yoğun çıkması sebebiyle olmuştur. İlgili futbolcuya yoğun bir şekilde bu söz öbeğinin tivitlenmesi dolayısıyla böyle bir sonuç elde edilmiştir. Aynı şekilde **haram** kelimesini kullanılması alınan ücrete gönderme yapılarak futbolcunun yüksek ücret aldığını göstermektedir. **Çocuk** kelimesinin bu kadar olumsuz tivitte geçmesi küfür olacak şekilde çocuk kelimesinin kullanılmasından kaynaklanmaktadır. **Ofsayt** ve **forvet** kelimelerinin olumsuz alanda yoğunlaşma sebebi ilgili futbolcunun **çok fazla ofsayta düşen bir forvet oyuncusu** olmasından kaynaklanmaktadır.

Kardeş kelimesinin yoğun olduğu tivitlerde ise futbolcuyu motive etmek amacıyla sarf edilen sözlerdeki hitaptan kaynaklanmaktadır. Aynı şekilde **bulut** ve **yılmaz** kelimelerinin beraber oynadıkları arkadaşına gönderme yapılarak motivasyon için atılan olumlu tivitlerde geçmesi dolayısıyla böyle bir sonuç elde edilmiştir.

Nitelik çıkarımından elde edilen sonuçlar gösteriyorki, Twitter üzerinden bir kişi veya kurum hakkında bir yargıya varmak mümkündür. Bu şekilde kullanıcı yorumları incelenerek kişinin yaptığı işi veya bir kurumun verdiği hizmetin nitelikleri belirlenebilir.

BÖLÜM 6

SONUÇ

Yapılan çalışma duygu analizi ve nitelik çıkarımı için elde edilen sonuçların başarılı olduğunu göstermiştir. Bununla beraber duygu analizinde, eğitim setinin kalitesi ve kategorik dağılımının analize doğrudan etki ettiği görülmüştür. Bu sebeple duygu analizi için kullanılacak eğitim verisinin sektör bazlı olması, elde edilecek duygu analizi ve nitelik çıkarımı sonuçlarını oldukça ileriye taşıyacağı düşünülmektedir. Nitelik çıkarımı için elde edilen verilerin daha düzgün sonuçlar verebilmesi için duygu analizinin de bir o kadar başarılı olması gerektiğidir.

Sonuç olarak yapılan çalışmanın en önemli adımı olan nitelik çıkarımı; kişi, kurum ve ürün hakkında sosyal medyadan elde edilen verilere duygu analizi yapıldıktan sonra yapılan nitelik çıkarımı, ilgili şey hakkında bir yargı elde edilebileceğini göstermiştir. Bu sonuçla ilgili kurum veya kişinin gelecekteki stratejisini kullanıcı yorumlarına bakarak yönlendirebilmesine olanak sağlayacağı düşünülmektedir.

Ayrıca, elde edilen veriler bize ilgili kaynak hakkında müşteri yorumlarını otomatize edebilecek şekilde analiz edilebildiğini ve bu sayede insan gücünden kazanç sağlanarak analiz maliyetlerinin düşürülebileceğini göstermiştir. Otomatik olarak ilgi kurum, kişi veya ürünün niteliklerinin çıkarılabileceğinin ispatı niteliğindedir. Bu sayede müşteri memnuniyetinin ve kurum, marka stratejisinin daha hızlı ve daha az maliyet ile belirlenebilmesine olanak sağlanabileceği düşünülmektedir.

Bu analiz sonucunda elde edilen bir diğer çıkarımın, markanın günlük yapılan kampanyalar veya reklamlar doğrultusunda hızlı geri bildirim alınabilmesinin önünü açmasıdır. Sosyal medyadan gelen yorumların analizi sonucu, ilgili kampanya veya

reklamın çok hızlı sürede nasıl bir etki bıraktığının görülebilmesine de olanak sağlayacağı düşünülmektedir.

Ayrıca, çıkarılan bu sonuçlarla şirket politikalarının ve verilen hizmetin kalitesinin iyileştirmeye ihtiyacı olup olmadığını oldukça hızlı yapılabilecek olmasıdır. Bunun sonucu olarak yapılan iyileşmenin de geri dönüşümü büyük bir hızla alınabildiğinden, ilgili iyileştirmenin işe yarayıp yaramadığı daha hızlı görülebileceği düşünülmektedir. Bu da gereksiz yatırım ve maliyetin önüne geçmektedir. Aynı zamanda çalışmanın, kurumlar için oldukça önemli bir CRM aracı olabileceği düşünülmektedir.

KAYNAKLAR

[1] Twitter Statistics, Statistic Brain, http://www.statisticbrain.com/twitter-statistics/ , 25 Eylül 2015.

[2] Jurafsky, D. ve Martin, J.H., (2008). "Speech and Language Processing 2nd Edition", Prentice Hall.

[3] Huyssteen, G. B. V., Eiselen, R. ve Puttkammer, M., (2004). "Evaluating Evaluation Metrics for Spelling Checker Evaluations", *International Proofing Tools and Language Technologies Workshop,*. 91-99

[4] Torunoğlu, D ve Eryiğit, G, (2014) "A Cascaded Approach for Social Media Text Normalization of Turkish" *In 5th Workshop on Language Analysis for Social Media (LASM) at EACL*

[5] Vinodhini, G. ve Chandrasekaran, R., (2012). "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, 2:6

[6] Domingos, P. ve Pazzani, M., (1997). "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", Machine Learning, 29:103-103.

[7] Niu, Z., Yin, Z. ve Kong, X., (2012). "Sentiment classification for microblog by machine learning", Computational and Information Sciences (ICCIS) 2012 Fourth International Conference on, IEEE , 286–289.

[8] Barbosa, L. ve Feng, J., (2010). "Robust sentiment detection on twitter from biased and noisy data", 23 Uluslararası Konferans, ,Computational Linguistics, 36–44.

[9] Celikyilmaz, A., Hakkani-Tur, D., ve Feng, J., (2010). "Probabilistic model-based sentiment analysis of twitter messages", Spoken Language Technology Workshop, 79–84

[10] Balahur, A., Hermida, J. M. ve Montoyo A., (2012). "Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model", Affective Computing, :3 88–101

[11] Xia, R., Zong, C., & Li, S., (2011). “Ensemble of feature sets and classification algorithms for sentiment classification”, Information Sciences, 1138-1152.

[12] Peddinti , V. M. K. ve Chintalapoodi, P., (2011). “Domain adaptation in sentiment analysis of twitter”, Analyzing Microtext Workshop, AAA.

[13] Neethu, M. S., ve R. Rajasree., (2013). “Sentiment analysis in twitter using machine learning techniques”, Communications and Networking Technologies (ICCCNT), 2013 4. Uluslararası Konferansı, IEEE, 1-5.

[14] Mejanova, Y., (2009). “Sentiment analysis: An overview”, Comprehensive Exam paper, http://www.academia.edu/291678/Sentiment_Analysis_An_Overview .

[15] Stemler, S., (2001). “An Overview of Content Analysis”, Practical Assessment, Research & Evaluation, http://pareonline.net/getvn.asp?v=7&n=17

[16] Kutlu, M., Cığır, C. ve Çiçekli İ., (2010). “Generic Text Summarization for Turkish”, ISCIS 2009. 24. Uluslararası Sempozyum, Computer and Information Sciences, 224-229

[17] Şimşek, M. U., & Özdemir, S., (2012). “Analysis of the relation between Turkish twitter messages and stock market index”, Application of Information and Communication Technologies (AICT), 2012 6th International Conference, IEEE, 1-4

[18] Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., ve By, T., (2012). “Sentiment analysis on social media”, 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE, 919-926.

[19] A. Griffin, Technology News, http://www.independent.co.uk/life-style/gadgets-and-tech/news/twitter-to-expand-tweet-s-character-limit-could-increase-it-to-10000-a6797906.html, 5 Ocak 2016.

[20] Hamilton, J, (2008). “Facebook Releases Cassandra as Open Source”, http://perspectives.mvdirona.com/2008/07/facebook-releases-cassandra-as-open-source/ , 12 Haziran 2008

[21] Datasets, Yıldız Teknik Üniversitesi, http://www.kemik.yildiz.edu.tr/?id=28 .

[22] Zemberek NLP Kütüphanesi, https://github.com/ahmetaa/zemberek-nlp , 20 Ocak 2013

[23] Weka: Data Mining Software, http://www.cs.waikato.ac.nz/ml/weka/ , 3 Mart 2014

[24] Twitter Search API, https://dev.twitter.com/rest/reference/get/search/tweets , 24 Ekim 2013

[25] Apache Cassandra, Database, http://cassandra.apache.org/ , 16 Haziran 2015

[26] Apache POI, the Java API for Microsoft Documents, https://poi.apache.org/ , 29 Eylül 2015

[27] Apache Lucene, Search Engine, https://lucene.apache.org/core/ , 24 Ağustos 2015

[28] What is Object/Relational Mapping, Hibernate Overview. JBOSS Hibernate, http://hibernate.org/orm/what-is-an-orm/ , 2011

[29] Java Cassandra Driver, DataSlax, https://github.com/datastax/java-driver, 10 Kasım 2015

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı :Tugay ÖZGİRGİN

Doğum Tarihi ve Yeri :03-01-1986 Hamburg / Almanya

Yabancı Dili : İngilizce / Rusça / Kırgızca

E-posta : girginsoft@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2016
Lisans	Bilgisayar Mühendisliği	Kırgız-Türk Manas Üniversitesi	2009
Lise	Fen Bilimleri	Nevşehir Anadolu Lisesi	2004

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2015-Halen	İncir A.Ş	Yazılım Mimarı / ERP Sorumlusu
2009-2015	Kartaca Bilişim	Yazılım Uzmanı / Yazılım Bölüm Sorumlusu

YAYINLARI

Bildiri

1. “Sosyal Medya Üzerinden Duygu Analizi ve Arka Plan Sebebi Tespiti”, Tugay Özgirgin, Banu Diri, 2016. *ASYU 2016 (*Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*)*, 29 Eylük-1 Ekim 2016, Düzce