T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TERÖRİZM İÇERİKLİ ONLİNE SOSYAL YORUMLARIN DUYGU ANALİZİ VE FİKİR MADENCİLİĞİ

DOKTORA TEZİ

Ibrahim Amine FADEL

Enstitü Anabilim Dalı	**:**	**BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Prof. Dr. Cemil ÖZ**

Eylül 2020

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TERÖRİZM İÇERİKLİ ONLİNE SOSYAL YORUMLARIN DUYGU ANALİZİ VE FİKİR MADENCİLİĞİ

DOKTORA TEZİ

Ibrahim Amine FADEL

Enstitü Anabilim Dalı : BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ

Bu tez 02/09/2020 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

Prof. Dr. **Cemil ÖZ** **Jüri Başkanı**	**Prof. Dr.** **Raşit KÖKER** **Üye**	**Doç. Dr.** **Osman ÖZKUL** **Üye**
Doç. Dr. **Metin VARAN** **Üye**		**Dr.** **Ali GÜLBAĞ** **Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Ibrahim Amine FADEL

16.07.2020

TEŞEKKÜR

Doktora eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasından yazılmasına kadar tüm aşamalarında yardımlarını esirgemeyen, teşvik eden, aynı titizlikte beni yönlendiren değerli danışman hocam Prof. Dr. Cemil ÖZ'e teşekkürlerimi sunarım.

Bu araştırmaya kaynak sağladıkları için sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi çok teşekkür ederim. Aynı şekilde, Turkiye Burslari'ye yurtdışı eğitime finansal destek sağladığı için teşekkür etmek istiyorum. Ayrıca eşim, oğlum, ebeveynlerim ve kardeşlerime sevgi, destek ve dilekçeleri için teşekkür etmek istiyorum. Deneme sürelerinde benim için orada olduğun için çok minnettarım.

Son olarak, tüm arkadaşlarıma ve çalışma arkadaşlarıma destekleri ve yardımları için teşekkür ediyorum.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

DT	: Decission Tree
IŞİD	: Irak İslam Devleti ve Levant
LDA	: Latent Dirichlet Allocation
LR	: Logistic Regression
ML	: Machine Learning
NB	: Naive Bayes
NLP	: Natural Language Processing
NLTK	: Natural Language Toolkit
POS	: Part-Of-Speech
SVM	: Support Vector Machine
TF-IDF	: Term Frequency and Inverse Document Frequency

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Duygu analizi, Terrorist, Makine öğrenimi, Konu modelleme.

Duygu analizi, insanların konuştukları veya yazdıkları şeyler hakkındaki duygularını ortaya çıkarmak için kullanılmıştır. Duygu analizi, Cümleyi oluşturan pozitif ve negatif duygu terimlerinin frekansları hesaplanarak yapılan bir metin madenciliği yöntemidir. Bu araştırmada, metin birimlerinin olumsuz ve olumlu değerlerinin sıklığının, bir terör saldırısından sonra tepki duygusunu keşfetme alanı gibi bazı alanlarda içerikleri hakkında yeterli olduğunu söylemiyoruz.

Bu tezde, duygu analizi, konu modelleme ve bulanık mantık sistemi arasında bir melez olan bir model önerdik. Bu model, 2017 yılında Londra ve Barselona'da meydana gelen neredeyse aynı iki terörist saldırının tepkileri üzerine değerlendirildi. Reaksiyon veri setleri Twitter'dan toplandı.

İlk olarak, tweet içeriğinin anlambiliminin saldırının her biri ile nasıl ilişkili olduğunu anlamak için tweet içeriğinin ön analizi yapıldı. daha sonra reaksiyonların ana konuları insan kararıyla LDA konu modelleme algoritması kullanılarak çıkarılmıştır. Fikir kelimelerini puanlamak için kullanılan sözlük yaklaşımı, tweetlerdeki tüm görüş kelimelerinin, görüş sözlüğündeki kelimelerle eşleştirilerek tanımlandığı tweetlerden oluşmaktadır. Daha sonra polariteyi belirlemek ve tweet'i duygusal sınıflandırma için etiketlemek için bir bulanık mantık sistemi kullanıldı. POS ve N-gram olarak özellikler çıkarılmıştır. Son olarak, tweet'lerin duygusallığını sınıflandırmak için farklı makine öğrenme yöntemleri kullanılmaktadır.

Araştırmada elde edilen sonuçlara göre, insanların tepkileriyle ilgili konuların konuyla ilgili olduğu ve metindeki gerçek duyguları önemli ölçüde tahmin etmeye yardımcı olabileceği sonucuna varılmıştır.

OPINION MINING AND SENTIMENT ANALYSIS FOR TERRORIST REVIEWS ON SOCIAL MEDIA

SUMMARY

Keywords: Sentiment analysis, Terrorist, Machine learning, Topic modeling.

Sentiment analysis has been used to bring out the emotions of people on the things they talk or write about. It is a text mining method, which is made by calculating the frequencies of positive and negative sentiment terms that make up the sentence. In this research, we argue that the frequency of negative and positive values of text units do not say enough about their content in some domains, for example exploring the sentiment of reactions after a terrorist attack.

In this thesis, we proposed a model which is a hybrid between Sentiment Analysis, Topic modeling and Fuzzy logic system. This model has been evaluated on reactions about two almost identical terrorist attacks that occurred in London and Barcelona in 2017. A dataset of reactions to these events was collected from Twitter.

First, a preliminary analysis of the tweet content was performed to understand how the semantics of tweet content relates to each of the attacks. Then the main topics of the reactions were extracted using the LDA topic modeling algorithm with human judgment. The lexicon approach was used to score the opinion words in the tweets. All opinion words from the tweets were identified by matching them with the words in the opinion lexicon. Then, a fuzzy logic system was used to determine polarity and label the tweet for sentimental classification. Features such as POS and N-grams were extracted. Finally, different machine learning methods were used to classify the sentiment of the tweets.

According to the results obtained in this research, it was concluded that the topics from people's reactions are relevant and it can help to significantly predict the real emotions in the text.

BÖLÜM 1. GİRİŞ

Günümüzde, sosyal medya (Twitter, facebook gibi) insanların iletişim ve fikirlerini paylaşmaları için vazgeçilmez bir kanal haline gelmiştir. Bu fikirler ve duygular, ürün ve hizmetlerin değerlendirilmesinde giderek önem kazanmaktadır. sosyal medyayı izlemek ve kullanıcıların duygularını analiz etmek ve müşterilerine halkın işleri, ürünleri veya ilgilendikleri konular hakkında neler hissettiği hakkında bilgi vermek için çalışan birçok uzmanlaşmış şirket vardır.

Duygu analizi, Doğal Dil İşleme (NLP), hesaplamalı dilbilim ve metin madenciliği problemlerinin uygulanmasını ifade eder (Pang ve ark., 2012). duygu analizinin temel amacı, belirli bir kişinin bir konuya ilişkin tutumunu belirlemektir.

Duygu analizi, son yıllarda en hızlı büyüme gösteren araştırma alanlarından biridir. sadece işletmelerin müşterilerinin ürün hakkındaki düşüncelerini veya duygularını değerlendirmelerine yardımcı olmak değil. Aksine, kamu kurumlarının kamuoyu araştırması ve vatandaşların tutumlarını öğrenmesi önemli hale gelmiştir.

Terörizm olgusu hala uluslararası gündemde yer alan konulardan biridir ve sosyal medyada insanlar arasında nedenleri ve bunların önlenmesi ile ilgili tartışma ve soruları gündeme getirmektedir. Sosyal medya kullanıcılarının terör olaylarına tepkileri, güvenlik duygusu kaybetme, güçsüz hissetme, öfkeli ve korkulu hissetme ile bazı etnik veya dini gruplara karşı hoşgörüsüzlük arasında değişmektedir.

Bu nedenle, bu tepkileri incelemek ve izlemek, yetkililerin bu sorunlarla başa çıkmak için belirli yardım programlarının tanımlamasına ve sunuımasına yardımcı olabilir (Cohen-Louck ve Ben-David, 2017). Terör olaylarından sonra sosyal medya kullanımı ve tepkiler üzerine çeşitli çalışmalar vardır. Mesela (Garg ve ark., 2017)

son retweet, retweet sayısı, sık kullanılan sayısı gibi faktörleri ölçerek Twitter'da terörist operasyonlara verilen yanıtların seviyesini incelemiştir. (Simon ve ark. 2014) ise sosyal medyanın kriz yönetimine nasıl katkıda bulunduğunu değerlendirmek için Kenya'daki bir terör saldırısından sonraki 4 günlük kuşatma ile ilgili tweet'leri incelemiştir. Kenya hükümeti, durumu gerçek zamanlı olarak daha iyi anlamak, halkla iletişim kurmak, güvenliklerini artırmak ve söylentilerle mücadele etmek için terör saldırısı sırasında sosyal medyayı yaygın olarak benimseyip ve kullanmıştır. Çalışma, acil durumlarda sosyal medya izlemenin kullanılmasının, görgü tanıklarından durumları anlamada ve ayrıca doğru durumları halka açıklamada çok değerli bilgiler elde etmeye yardımcı olduğunu belirlemiştir. (Jones ve ark. 2016) ayrıca Twitter'ın kitlesel şiddet olaylarının topluluklar üzerindeki duygusal etkisini incelemek için uygun bir veri kaynağı olabileceğini bulmuştur.

Tweetin terör olaylarıyla ilgili bilgi akışı modelindeki duyguların etkisi de araştırılmıştır (Burnap ve ark., 2014) 2013 yılında Londra'da bir terör saldırısı üzerinde çalıştılar. Twitter verilerini kullanarak, kapsamı tahmin etmek için istatistiksel modeller geliştirdiler ve olayla ilgili bilgi akışının kalıcılığı. Retweet miktarının kamuoyunun ilgisini ve bilginin onaylandığını gösterdiğini belirttiler. Tweet'in duygularının, bilgi akışlarının kapsamını ve kalıcılığını önemli ölçüde yordadığını buldular. Sonuçlarına göre, olumlu duygularla ifade edilen içeriğin, olumsuz içeriğe kıyasla onaylanması ve daha uzun süre dayanması daha olasıdır. Tersine, (Garg ve ark. 2017) Uri terör saldırıları konusunu tartışan, olumsuz tweetlerin olumlu tweetlerden daha fazla kalma eğiliminde olduğunu buldu. İnsanlara çok fazla olumsuz bilgi hakim olursa, örneğin insanlar bir topluluğu hedef almaya başlayabilir ve bu da sivil huzursuzluğa yol açabilir. (Mirani ve ark. 2016), ISIS ile ilgili tweetler üzerinde duygu analizi sunar. Fikir sözlüğü kullanılarak etiketlenmiş bir veri kümesi üzerinde beş farklı algoritma eğittiler. Tweetlerin duygularının neredeyse olumsuz olduğunu gördüler.

İlişkili kişiler arasındaki duyguların benzerliklerinin sosyolojik gözlemi tarafından yapılmıştır (Chong, 2016). çalışma, #prayforparis için Twitter duyguların anlamlarını analiz ederek ve konuları çıkararak araştırıyor. Diğer çalışmalar, terörizme karşı

duyarlılığı kutupluluk açısından da araştırdı. Farklı ülkelerden bir duyarlılık sözlüğü ve coğrafi etiketli IŞİD ile ilgili tweetler kullanmak.

(Mansour, 2018) araştırmalar, batı ülkelerinden ve doğu ülkelerinden insanların IŞİD'e nasıl baktıklarına dair bir fark olup olmadığını incelemek için terörizme karşı duyarlılığı araştırdı.

Her iki taraftaki insanların IŞİD hakkında tweet atarken neredeyse aynı kelimeleri kullandıklarını keşfetti. Farklı taraftaki insanların, hangi ülkeden olursa olsun, insanların bir terörist grubu olarak bu örgüt hakkında nasıl düşündüklerini yansıtan aynı olumlu ve olumsuz kelimeleri paylaştığını buldu. Ayrıca, çoğu kullanıcı IŞİD'i nereden gelirse gelsin bir tehdit ve korku kaynağı olarak görüyor.

Terörizme sosyal medya tepkileriyle ilgili benzer bir çalışmada (Becker ve ark., 2019), Twitter kullanıcılarının Birleşik Krallık'ta meydana gelen iki terör olayı hakkındaki duygusal tepkilerini inceledi. Çalışma, bir duygu sınıflandırıcısı geliştirmek için iki derin öğrenme mimarisini test ediyor ve terörist saldırı nedeniyle duygusal bir değişim olup olmadığını ve duygusal tepkilerin olaya bağlı olup olmadığını anlamak için terörist olaylarla ilgili tweetler üzerinde bir analiz geliştiriyor ve kullanıcıların demografik özelliklerini açıklıyor.

Duygu sınıflandırması, bir metnin duygu kutuplarının (pozitif, negatif, vb.) değerlendirilmesi ve öngörülmesine odaklanır. Bu kutuplar, bazı konulardaki duyguları tanımlamak için yeterli değildir. Terör olaylarına tepkiler hakkındaki duygu ifadelerinin olumludan daha olumsuz olması muhtemeldir, çünkü kelimeler içerdiği için olumsuz duygular vardır. Bu konuda normal duygu analizi yönteminin kullanılması, duyguların doğru bir değerlendirmesini vermemektedir.

Bu nedenle, metnin içinde temel konuların aranması, kullanıcının görüşünü daha iyi tanımak için gerçekten yararlıdır. Bir incelemede sunulan konular, bu incelemenin tematik yapısını gösterir ve kullanıcının farklı yönler hakkındaki görüşünü anlamada

önemli bir rol oynar. Bu nedenle, konu modelleme ve duygu analizini birleştirmek, iyi sonuçlar elde etmek için umut verici bir fikirdir.

Konu modelleme, bir metin topluluğundaki baskın konuları ve terimleri keşfetmek için belirli bir algoritma kullanan bilgisayar tabanlı bir metin madenciliği yöntemidir. Sosyal bilimsel araştırmalarda giderek daha popüler hale gelen bir tekniktir. Metindeki gizli konuları ortaya çıkarır ve her konu için önemli terimleri listeler. Daha sonra araştırmacı konuları niteliksel olarak tanımlayabilir.

Bu tezde, terörist saldırılarla ilgili tepkilerin gözden geçirilmesi konusundaki düşünceyi analiz etmek için kutup modelinin belirlenmesi için insan mantığı ile duygu analizi ve bulanık mantık ile duygu analizinin harmanlanmasını önerilmektedir. Bu model için Twitter kullanıcılarının iki farklı ülkede meydana gelen neredeyse özdeş terörist saldırıların tepkilerinin bir koleksiyon veri seti kullanılmıştır.

Bu araştırma aşağıdaki araştırma sorularını cevaplamayı amaçlamaktadır:

- S1: Benzer terörist saldırılar aynı duygusal yanıtı tetikler mi?

- S2: Terörist saldırılara verilen tepkilerdeki genel konu türleri nelerdir?

- S3: Konu modelleme ve duygu analizinin birleştirilmesi terörist olaylara verilen tepkilerin doğasını anlamak için daha uygun mudur?

- S4: Farklı makine öğrenimi algoritmaları kullanan n-gram ve POS gibi özellik setlerinin performansı nedir?

1.1. Arıştırma Katkısı

Bu tezde, terör olaylarına verilen tepkilerin doğasını kavramak için duygu analizi ile konu modelleme önerilmiştir. Bu tezin katkıları aşağıdaki açılardan özetlenebilir.

- Terör olaylarından etkilenen toplumların ve bunların afet sonrası anında tepkilerinin incelenmesi.

- Derin bilgi elde etmek için Twitter'dan toplanan gürültülü veriler üzerinde bazı ön analizlerin nasıl kullanılacağını açıklayın.

- Sosyal medyadaki tepkileri, duygusal bulaşma sosyal teorisinin (Duygusal bulaşma, başka bir duygusal olarak kendiliğinden anlamına gelir) insanların sosyal medyadaki terör olaylarına tepkilerinde doğru olup olmadığını doğrulamak için, neredeyse aynı iki terörist saldırıyla ilgili olarak incelenmiştir.

- Terör olaylarına verilen tepkileri sınıflandırmak için duyarlılık analizi ile konu modellemesinin kullanımını açıklayın ve konuya bağlı duyarlılık analizinin çok daha iyi performans gösterdiğini ve duygu ve konu kelimeleri arasındaki yüksek etkileşimi gösterdiğini gösterin.

- Tweetlerin kutuplarını belirlemede bulanık mantık üzerine konu modelleme ve sözlük yaklaşımı çıktılarının kullanılması.

- Hangisinin daha iyi sonuçlar verdiğini belirlemek için farklı sınıflandırma algoritmalarıyla farklı özellik türlerini test edin.

- Popüler bir denetimli sınıflandırıcının performansını konu bağımlılığı bağlamında incelenmiştir.

Hangisinin daha iyi sonuçlar verdiğini belirlemek için farklı sınıflandırma algoritmaları ile farklı özellik türlerini test edin.

1.2. Arıştırma Kapsamı

Çalışmanın kapsamı aşağıdaki gibidir:

- Veri kaynağı yalnızca İngilizce dilinde yazılmış tweetlerle sınırlıdır.

- Uygulamaya ilişkin analizler metin verisi ile sınırlıdır.

1.3. Tez Organizasyonu

Tezin geri kalanı aşağıdaki gibi düzenlenmiştir:

Bölüm 2: Literatür Taraması: Bu bölüm, terörizm ve sosyal medya ile ilgili temel fikirlerle başlıyor ve duygu analizinde son teknoloji yaklaşımları değerlendirimiştir. Ayrıca sınıflandırma algoritmalarını gözden geçirirmiş ve değerlendirme ölçümleri de kapsama alınmıştır.

Bölüm 3: Veri Toplama ve Ön Analiz: Bu bölümde, veri kümesinin twitterdan nasıl toplandığı hakkında genel bir bakış ve metnin veri kümeleri ve ön işlemesi için ön analiz verilmiştir.

Bölüm 4: Konuya bağlı duygu analizi: Bu bölümde, ana konuları ayıklamak için LDA konu modelleme algoritmasını kullanıyoruz ve verileri sınıflandırmak için makine öğrenme algoritmalarında kullandığımız etiketli bir veri kümesi oluşturmak için bunları sözcüksel yaklaşımda kullanılmıştır.

Bölüm 5: Tartışma ve Sonuç: Son bölümde farklı özellik türlerini kullanarak sınıflandırıcıların performansını sunuyoruz. Ayrıca bu çalışmanın sonuçlarını da sunulmuştur. Bu bölüm ayrıca gelecekteki çalışmalar için öneriler getirecektir.

BÖLÜM 2. LİTERATÜR TARAMASI

Terör olgusu bu yeni yüzyılın ayırt edici bir özelliği haline gelmiştir. Bu yüzyılın başında, 11 Eylül 2001'de Amerika Birleşik Devletleri'nde meydana gelen terörist saldırılardan sonra. terörizm ve onunla mücadele konusu, dünyanın birçok ülkesinin politikalarında en önemli konulardan biri haline gelmiştir. 2018'de, tüm dünyada, 7.290 fail ve 15.690 kurban da dahil olmak üzere 22.980'den fazla insanı öldüren 9.600'den fazla terörist saldırı gerçekleşmiştir. Bu eylemler her yıl milyarlarca dolarlık ekonomik zarara yol açıyor. 2019 Küresel Terörizm Endeksi'ne göre. Terörün küresel ekonomi üzerindeki etkisi, IŞİD grubunun saldırılarının zirveye ulaştığı 2014 yılında 111 milyar dolara ulaşmıştır. Terörist operasyonların azalması nedeniyle zararlar 2017'de 54 milyar dolara geriledi (Institute for Economics &. Peace, 2019).

Terörizmin tanımı konusunda bir anlaşma olmamasına rağmen, bu fenomenle mücadele etmek hükümetleri birçok strateji benimsemeye itmiştir. Bu stratejilerden biri, hükümetlerin terörist grupların faaliyetleriyle mücadele etmek için yeni teknoloji kullanmasıdır. Bu, araştırmaya yönelik bilgilerin işlenmesine yardımcı olmak için büyük ölçekli veri toplama ve veri madenciliğini içerir.

İstatistiksel sosyal bilim, insanların internet davranışından elde edilen verileri kullanan insanların sosyal davranışlarını incelemek için yeni bir alandır (Cioffi-Revilla, 2010). Bu alan, insan davranış kalıpları elde etmek için önemli bir alan haline gelmiştir ve insan beyninin sosyal medyada nasıl çalıştığını araştıran araştırmacıların derinlemesine bir analizini sunmaktadır.

“Sosyal medya, topluluk temelli girdi, etkileşim, içerik paylaşımı ve işbirliğine adanmış çevrimiçi iletişim kanallarının toplanmasıdır” (Wise ve Shorter, 2014).

Sanal iletişim ve ağlarda insanların kendi aralarında içerik oluşturmasına, paylaşmasına, paylaşmasına ve yorumlamasına olanak tanır. Örneğin, Twitter ve Facebook artık insanlar tarafından düşüncelerini ve fikirlerini paylaşmak ve diğer insanlarla etkileşim kurmak için giderek daha fazla kullanılmaktadır. Bir tür gerçek dünya yansıması olarak kabul edilir (Yakushev ve Mityagin, 2014).

Terör örgütleri sosyal medya platformlarını giderek daha fazla bilgi alışverişinde bulunmak ve yeni üyeler ile destekçileri işe almak ve mesajlarını yaymak için kullanıyorlar, çünkü ucuz, erişilebilir, mesajların hızlı ve geniş bir şekilde yayılmasını kolaylaştırıyor ve ana akım haber kaynakları (Calvin Dark, 2011). Hayfa Üniversitesi Gabriel Weimann tarafından yapılan araştırmaya göre, internette örgütlü terörizmin yaklaşık% 90'ının sosyal medya aracılığıyla gerçekleştiğini buldu (CBC News, 2016). Sosyal medya izleyicisi olan Recorded Future, IŞİD'in terörist grubu tartışan toplam 700.000 hesapla hype yaratmayı başardığını buldu. Bunun için Twitter, terörist bağlantılardan şüphelenilen binlerce hesabı askıya alarak IŞİD'e karşı koymaya çalıştı (Telegraph Reporters, 2014). Bu bağlamda, hükümetler giderek terörist unsurları tespit etmek ve terörist arasındaki ilişkiyi tanımlamak için bilgisayar teknolojilerini kullanma arayışındadır. İnternette teröristlerin tespit edilmesinin daha fazla terörist saldırıyı önleyebileceğine inanılmaktadır (Kelley, 2001). Veri madenciliği bu teknolojilerden biridir.

Veri madenciliği, büyük veri toplamalarından bilgi elde etmek için uygulanan bir teknolojidir, kullanıcıların birçok farklı boyut ve açıdan verileri analiz etmelerine, kategorize etmelerine ve belirlenen ilişkileri özetlemelerine olanak tanır (Russell, 2011). Veri madenciliği tarafından sağlanan bu avantaj hükümetleri terörizmle mücadelede kullanmaya teşvik etti. son yıllarda bu kullanıcıların ilgi alanlarını anlamak için sosyal medyada terörist davranışların madenciliğinde önemli araştırma çabaları olmuştur. çünkü sosyal medyadaki kullanıcı etkinlikleri, büyük, geniş ve kullanıcı tercihleri, ilgi alanları, görüşler ve ilişkilerin göstergesi olan davranışsal veriler üretmiştir (Akcora ve ark.., 2014). Bu faaliyetler çeşitli alanlarda önem kazanmaktadır. İncelemelerimizden, terörizm ve sosyal medya ile ilgili makalelerimizden, bu araştırmaları altı alanda sonuçlandırabilir.

- Grup Tespiti, birbirine benzeyen unsurları belirleyerek terörist grupları tespit etmek için kullanılır (Plantié ve Crampes, 2013).

- Üyelerin kendilerinden ziyade terörist üyeler arasındaki bağlantılara odaklanan Link Tahmini, anormal iletişimleri tanımlamak için de kullanılır (Rai, 2017).

- Yanıltıcı, tutarsız, çelişkili veya yanlış bilgileri keşfetmenin bir yöntemi olan Aldatmacayı Tahmin Etme. Terörist gruplar, kendilerini korumak ve tespit edilmekten kaçınmak için çoğunlukla sosyal medyada kullanırlar (Alowibdi ve ark., 2014).

- Anahtar Oyuncular, içgörü kazanmak için yapısal özelliklerine odaklanarak ve merkezi mevkileri işgal eden üyeleri (lider) belirleyerek terörist ağı analiz ederlerdi (Memon ve Larsen, 2006).

- Davranış Analizi, terörist davranışların keşfi ve sosyal medyada etkileşime odaklanmaktadır (Zafarani ve Liu, 2014) Bu, belirli bir bağlantıya veya sayfaya tıklamak, profil taraması, dostluk tepkisi vb. (Elovici ve ark., 2004).

- Terörist grubun görüşlerini sosyal medyada yayınlanan bilgilere dayanarak değerlendirmek ve anlamak için kullanılan Duygu Analizi.

2.1. Duygu Analizi

Duygu analizi, metinde ifade edilen görüşler, tutumlar ve duygular gibi öznel bilgileri saptamak için otomatik araçlar kullanmayı amaçlar ve son yıllarda doğal dil işlemeye olan ilginin hızla artmasını sağlamıştır. Çeşitli araştırma alanlarını içerir: NLP, hesaplamalı dilbilim ve metin analizi. Genellikle metin biçiminde ham verilerden öznel bilgilerin çıkarılmasını ifade eder (De Groot, 2012). Geniş bir problem alanını temsil eder. Ayrıca, duygu analizi, fikir madenciliği, fikir çıkarma,

duygu madenciliği, öznellik analizi, etki analizi, inceleme madenciliği vb gibi birçok isim ve biraz farklı görevler de vardır. Bununla birlikte, şimdi hepsi duygu analizi veya fikir madenciliği şemsiyesi altındadır (Liu, 2012a).

Bu iki ifade duygu analizi ve fikir madenciliği birbiriyle değiştirilebilir. Karşılıklı bir anlam ifade ederler. Bununla birlikte, bazı araştırmacılar, fikir madenciliği ve duygu analizinin biraz farklı kavramlara sahip olduğunu belirtmiştir (Tsytsarau ve Palpanas, 2012). Bu nedenle, duygu analizinin amacı, fikir bulmak, ifade ettikleri duyguları tanımlamak ve daha sonra kutuplarını sınıflandırmaktır. Bunlar değiştirilebilir. Karşılıklı bir anlam ifade ederler. endüstride, duygu analizi terimi daha yaygın olarak kullanılırken, akademide hem duygu analizi hem de fikir madenciliği sıklıkla kullanılmaktadır.

(Liu, 2012b) aşağıdaki (Denklem 2.1) 'de görüşün açık bir şekilde ifade edilebilmesi için bu beş parametrenin gerekli olduğunu belirtmektedir:

$$fikir = (o, f, s, h, t) \tag{2.1}$$

Burada o görüşü ifade edilen kavramı, f ise bu kavrama ait bir özelliği nitelemektedir. s hedef (o,f) için düşünülen duygu değerini, h görüşü ifade eden bireyi ve t ise ifadenin edildiği zamanı belirtmektedir.

Duygu Analizi, bir sınıflandırma süreci olarak düşünülebilir. Onun çalışmaları bir metnin polaritesini için üç farklı seviyelerde sınıflandırmaktır:

- Doküman düzeyi: Doküman tamamını olumlu veya olumsuz bir görüş veya duyguyu ifade eden bir görüş olarak sınıflandırmayı amaçlamaktadır. Yanı dokümanın tamamını, bir konu hakkında konuştuğu anlamına gelen temel bir bilgi birimi olarak görmektedir (Pang ve ark., 2002). Doküman düzeyinde sınıflandırma en iyi şekilde belge tek bir kişi tarafından yazıldığında ve tek bir varlık hakkında bir görüş / düşünce ifade ettiğinde işe yarar.

- Cümle düzeyi: Cümlenin olumlu veya olumsuz görüşlerini ifade edip etmediğini belirlemek için cümlenin öznelliğine veya nesnelliğine bağlı olarak her cümle içinde ifade edilen duyguyu sınıflandırmayı amaçlamaktadır. (Wilson ve ark., 200).

- Özellik tabanlı düzeyi: Duyguyu varlıkların belirli yönlerine göre sınıflandırmayı amaçlamaktadır. Her varlık daha sonra olumlu ya da olumsuz duyguların ya da fikirlerin varlığını belirlemek için analiz edilir (Hu ve Liu, 2004). Fikir sahipleri aynı işletmenin farklı yönleri için farklı görüşler verebilir. Örneğin, cümle "Bu telefonun ses kalitesi iyi değil, ancak pil ömrü uzun", telefonun iki yönünü, arama kalitesini ve pil ömrünü değerlendirir. Telefonun arama kalitesindeki duygu negatif, ancak batarya ömrüyle ilgili duygu olumlu. Telefonun arama kalitesi ve pil ömrü fikir hedefleridir. Bu analiz düzeyine dayanarak, yapılandırılmamış metni yapılandırılmış verilere dönüştüren ve her türlü kalitatif ve kantitatif analiz için kullanılabilen, kurumlar ve bunların yönleri hakkında görüşlerin yapılandırılmış bir özeti üretilebilir.

Duygu sınıflandırmasında ana endişe temsil ve özelliklerdir. Belgeleri modellemek için farklı metin gösterimleri kullanılmıştır. Bir sonraki bölümde, bu tezde kullanılan bu özellikler ele alınmatadır.

2.2. Duygu Sınıflandırma Özellikleri

Duygu Analizinde önemli bir özellik, herhangi bir veri örneğinden çıkarılabilen bilgilere karşılık gelir. Başka bir özellik, verilerin sahip olduğu özellikleri benzersiz bir şekilde tanımlar. Duygu analizinde kullanılan veriler, yüksek boyutlu bir özellik alanına yansıtılan özelliklerden oluşur. Bu yüksek boyutlu özellikler, veriler hakkındaki bilgileri mümkün olduğunca koruyan az sayıda düşük boyutlu değişkenle eşleştirilmelidir.

Özellik seçim yöntemleri, veri kümelerinin ilgili özelliklerini veya sınıflandırmasını yakalamak için belirli bir özellik kümesinden küçük bir özellik kümesi seçen tekniklerdir (Nicholls ve Song, 2010). Bu özellikler, orijinal verileri olabildiğince açıklamak için ayrımcı olmalıdır. Bu, farklı hedeflere ulaşmak içindir: hesaplama maliyetini azaltmak, fazla takmaktan kaçınmak ve model için sınıflandırma doğruluğunu arttırmak (Yousefpour ve ark., 2017).

Bu tezde, Konuşma Parçası etiketleme (POS etiketleme) ve N-gram'ı bir özellik olarak uygulayacaktır.

2.2.1. POS etiketleme

POS etiketleme, sözdizimsel ve morfolojik davranışlarına dayanarak belgedeki her terime dil kategorisi (genellikle POS etiketi olarak adlandırılır) atanmasını ifade eder (Das ve ark., 2015). POS etiketlemesinin basit anlaşılması, bir cümlede sıfat, zarf, fiil, isim, bağlaç vb. Gibi sözdizimsel anlamını ayırt etmek için bazı özel etiketler kullanmaktır. POS etiketleme, birçok duygu analizi sisteminin önemli bir özelliğidir (Agarwal ve ark., 2011).

Duygu Analizinde bu özellikleri seçmenin arkasındaki fikir, bir cümledeki sadece sınırlı bir kelime kümesinin, duygu-kelime olarak adlandırılan duyguyu göstermesidir. İngilizcede sözcük kategorilerine örnekler isim, fiil, zarf ve sıfattır. Bu sözcük kategorilerinden bazıları sıfatlar, fiiller ve zarflar gibi daha sık duygu kelimeleri içerir. POS etiketleyicisinin bir sorunu, birden fazla sözlük kategorisinde görünebilen bir kelime olabilir, ancak yalnızca bir sözlük kategorisindeki bir duyguyu ifade eder (Silva ve ark., 2014).

Sözcük kategorileri tüm diller için aynı değildir, her dilin kendi POS etiketleyicisini kullanması gerekir. İngilizce POS etiketlerinin listesi Tablo 2.1.'de gösterilmiştir. Birkaç makalede uygulanmış olan POS etiketleme tekniği sıklıkla kullanılır.

Table 2.1. İngilizce POS etiketleri

Etiket	Açıklama	Etiket	Açıklama
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

2.2.2. N-gram

N-gram tabanlı teknikler modern NLP ve uygulamalarında baskındır. N kelimelerini gözlemledikten sonra bir sonraki kelimeyi tahmin etmek için olasılıksal yöntemler kullanan bir kelime tahmin algoritmasıdır. Bu nedenle, bir sonraki kelimenin olasılığını hesaplamak, bir kelime dizisinin olasılığını hesaplamakla yakından ilgilidir. Bu yaklaşımın arkasındaki ana motivasyon, benzer kelimelerin yüksek oranda ortak N gramına sahip olacağıdır. N için tipik değerler 1, 2 veya 3; bunlar sırasıyla unigram, bigram veya trigram kullanımına karşılık gelir.

Unigramlarda (n = 1), her metin (tweet) bir belgedir ve kelimelere bölünür. Tüm belgelerdeki tüm kelimelerin sıklığını saymak, bir sözcük sıklığı tablosuyla sonuçlanır. Yüksek sık kullanılan kelimelerin metinlerde görünme olasılığı daha yüksektir, bu nedenle bu kelimeler veri kümesini daha iyi tanımlar.

Tüm belgeler arasındaki sıklığı ölçmenin iki yolu vardır: toplanan terim sıklığı ve belge sıklığı. Bu iki frekans ölçümü sırasıyla frekans terimine ve mevcudiyet terimine dayanmaktadır. Toplam terim sıklığı (stf) (Denklem 2.2) belirli bir terimin t tüm terim frekanslarını D hesaplanmış olarak tüm belgelerde toplar:

$$stf(t, D) = \sum_{d \in D} tf(t, d) \tag{2.2}$$

$$df(t, D) = |\{d \in D : w \in d\}|$$

döküman frekansının $df(wt, D)$ verilen terimin t , tüm belgeler arasında D

Ancak, sık kullanılan kelimeler sınıflandırma için mutlaka iyi özellikler değildir. Çok sık kullanılan bir kelimenin dağılımı sınıflar arasında eşit olarak dağıtılırsa, ayrımcı gücü düşüktür.

Daha yüksek sıralı n-gramlara kadar uzanan metinler, tek uzunluktaki öğeler olarak değil, n uzunluktaki öğeler olarak ayrılır. N = 2 (bigram) durumunda, maddeler birbirini takip eden iki kelimeden oluşur. Set, orijinal metinde birbirini takip eden iki kelimenin tüm kombinasyonlarını içerir. Aynı şey n = 3 (trigram) ve daha yüksek n değerleri ve ayrıca harf tabanlı n-gram için de geçerlidir. Sezgisel olarak, bu yüksek dereceli n-gramlar ardışık kelimelerin ilişkisini daha iyi yakalar gibi görünür, ör. olumsuzluk içinde.

2.2.3. TF-IDF

TF-IDF ölçüsü, bir kelimenin bir dizi belge içindeki önemini yansıtan bir istatistiktir. Metin sınıflandırma görevlerinde kullanılan ortak bir metriktir, ancak duygu analizinde kullanımı daha az yaygındır ve şaşırtıcı bir şekilde unigram özellik ağırlığı olarak kullanılmış gibi görünmemektedir. (Sebastiani, 2002).

TF-IDF, Terim Frekansı ve Ters Doküman Frekansı olmak üzere iki frekansdan oluşur. Terim Frekansı, belirli bir dokümanda belirli bir terimin gerçekleşme sayısını sayarak bulunur ve Ters Doküman Frekansı, toplam doküman sayısının TF-IDF'de

belirli bir kelimenin göründüğü belge sayısına bölünmesiyle bulunur. (Denklem 2.3) olarak hesaplanır:

$$tfidf(t,d,D) = tf(t,d).idf(t,D) \quad (2.3)$$

$f(t,d)$ terim frekansıdır, ve $idf(t,D)$ ters doküman frekansıdır. Hem tf hem de idf'in çeşitli varyantları vardır, ancak en basit biçimlerinde tf, bir dokümanda bir t teriminin kaç kez geçtiğini ve idf'in (Denklem 2.4) olduğunu belirtir.

$$idf(t,D) = log\frac{N}{|\{d \in D: t \in d\}|} \ , \ |\{d \in D: t \in d\}| \neq 0 \quad (2.4)$$

Burada D, tüm dokümanların bütünüdür ve N, korpustaki toplam doküman sayısıdır. Bir terimin belirdiği doküman sayısı arttıkça, logaritma içindeki oran 1'e düşecek, bu da idf yaklaşımını 0 yapacaktır.

Bu değerler birlikte çarpıldığında, birkaç dokümanda frekans görünen kelimeler için en yüksek, her dokümanda frekansa görünen terimler için düşük (ve her dokümanda seyrek görünen terimler için en düşük) bir puan alırız. bir dokümanda önemli olan terimler.

(Agarwal ve Mittal, 2012)'de İngilizce inceleme analizi için öne çıkan özellik çıkarımı üzerinde çeşitli deneysel karşılaştırmalar yapılmıştır. Unsurlar unigram, bi-gram, iki-etiketli özellik ve bağımlılık ayrıştırma ağacı tabanlı özellikler kullanılarak çıkarılmıştır. Ayrıca, özellik vektöründen gürültülü ve alakasız özellikleri ortadan kaldırmak için bilgi kazancı ve minimum yedeklilik maksimum alaka düzeyi özellik seçim yöntemleri kullanılmıştır. İnceleme belgesini pozitif veya negatif bir sınıfta sınıflandırmak için SVM ve çok terimli NB sınıflandırıcılar kullanıldı. Sonuç, multinom NB'nin ikili duygu sınıflandırması için doğruluk ve yürütme süresi açısından SVM'den daha iyi performans gösterdiğini göstermiştir.

(Pang ve ark., 2002) üç sınıflandırıcı NB, Maximum Entropy ve SVM'nin Doküman sınıflandırmasında sadece unigramlar, bigramlar, her ikisinin kombinasyonu,

unigramları ve POS'ları birleştirmek, sadece sıfatlar almak ve unigramları ve konum bilgilerini birleştirir. Sonuç, özellik varlığının özellik sıklığından daha önemli olduğunu ve özellik kümesi küçük olduğunda NB, SVM'den daha iyi performans gösterdiğini göstermiştir. Ancak özellik alanı artırıldığında SVM'ler daha iyi performans gösterir. Özellik alanı arttığında, Maksimum Entropi, Naïve Bayes'ten daha iyi performans gösterebilir, ancak aşırı uyumdan da muzdarip olabilir.

(Zheng ve ark., 2018) özellik seçiminin Çin çevrimiçi incelemelerinde duygu analizi üzerindeki etkilerini araştırdı. N-char-gram ve N-POS-gram potansiyel duygusal özellikleri seçmek için kullanılır. Özellik alt kümeleri, iyileştirilmiş bir belge sıklığı yöntemi kullanılarak seçilir ve özellik ağırlıkları Boole ağırlıklandırma yöntemi kullanılarak hesaplanır. Deneysel sonuçların önemini test etmek için ki-kare testi yapılır. Sonuç, N-char-gramları özellik olarak alırken, düşük dereceli N-char-gramların daha yüksek dereceli N-char-gramlardan daha iyi bir performans elde edebileceğini göstermektedir.

(Zhang ve ark., 2005) metin sınıflandırması için geliştirilmiş bir TF-IDF yaklaşımı önermiştir. Geleneksel TF-IDF yaklaşımı belgedeki her kelimeyi ne kadar farklı olduğuna göre ölçmek için kullanılır. Aksi takdirde, TF-IDF yöntemi kelimeler, metin belgeleri ve belirli kategoriler arasındaki alaka düzeyini yakalar. Sistem, metin sınıflandırmasının hatırlanmasını ve kesinliğini artırmak için güven, destek ve karakteristik kelimeler kullanır. Güven, belirli bir özellik sözcüğü ile belirli bir sınıfın oluşturulması için kesinlik ölçüsüdür. Belirli bir özellik sözcüğünün potansiyel yararlılığı destek ile gösterilir. Bir kelime listesi tarafından tanımlanan eşanlamlılar, bu gelişmiş TF-IDF yaklaşımında işlenir. Sonuçlar, geliştirilmiş TF-IDF yaklaşımının, geleneksel TF-IDF yaklaşımına kıyasla metin sınıflandırmasının doğruluğunu geliştirdiğini göstermektedir.

2.3. Duygu Sınıflandırma Yaklaşımları

Duygu sınıflandırması, iki veya daha fazla sınıfta verilen metnin duyarlılığının yönünün belirlenmesidir. Duygu sınıflandırması, sınıflandırma ile yakından

ilişkilidir, ancak standart metin sınıflandırmasından biraz farklıdır. Duygusal özellikleri bakış açıları, tercihler ve tutumlar gibi metinde sınıflandırmaya çalışır, oysa metin sınıflandırması temalara odaklanır (Ding ve ark., 2008).

Temel olarak iki tür Duyarga sınıflandırma yaklaşımı vardır ve bunlar Makine Öğrenme Yaklaşımı ve Sözcük tabanlı yaklaşımı içerir. Hibrit Yaklaşım her iki yaklaşımı da birleştirir ve çok yaygındır (Medhat ve ark., 2014). Duygu sınıflandırmasının çeşitli yaklaşımları Şekil 2.1.'de gösterilmektedir.

Şekil 2.1. Duygu sınıflandırma teknikleri.

2.3.1. Makine öğenmesi yaklaşımları

"Makine Öğrenmesi" terimi, bilgisayarların ileride karşılaşabileceği diğer veriler hakkında bir tahmin yapılmasına izin veren bir dizi veri vererek öğrenmesine izin vermekle ilgilidir. Makine öğrenimi, ünlü makine öğrenme algoritmalarını uygular ve dil özelliklerini kullanır. Makine öğrenimi yaklaşımını kullanan metin sınıflandırma yöntemleri kabaca denetimli, denetimsiz ve yarı-denetimli öğrenme yöntemlerine ayrılabilir (Chandrakala ve Sindhu, 2012).

2.3.1.1. Denetimli öğrenme yöntemi

Etiketli eğitim belgelerinin / ifadelerinin varlığına bağlıdır. Duygu analizinde, denetimli makine öğrenimi, bilgisayara bir dizi ifadeler (özellikler) ve kutupları sağlayarak bilgisayara görünmeyen metnin polaritesini tahmin etme yeteneği vererek gerçekleştirilir (Medhat ve ark., 2014). Olasılıksal, doğrusal, kural tabanlı ve karar

ağacı sınıflandırıcıları olan birkaç tür denetimli sınıflandırıcı vardır (Yusof ve ark., 2015).

Olasılıksal sınıflandırıcılar sınıflandırma için karışım modelleri kullanır. Karışım modeli, her sınıfın karışımın bir bileşeni olduğunu varsayar. Her karışım bileşeni, söz konusu bileşen için belirli bir terimi örnekleme olasılığını sağlayan üretken bir modeldir. Doğrusal bir sınıflandırıcı, sınıflandırmasını bir dizi ağırlık ile özellik vektörünü birleştiren doğrusal bir tahmin işlevine göre yapar.

Kural tabanlı sınıflandırıcılarda, veri alanı bir dizi kuralla modellenir. Sol taraf, özellik setinde ayrık normal formda ifade edilen bir koşulu temsil ederken, sağ taraf sınıf etiketidir. Karar ağacı sınıflandırıcıları, eğitim veri alanının hiyerarşik bir ayrışmasını sağlarken, verileri bölmek için özellik değerinde bir koşul kullanılır.

Bu tezde, denetimli sınıflandırıcılar için dört tür algoritma kullanacak ve bu algoritmaları duyarlı sınıflandırma algoritmaları bölümünde tartışacaktır.

2.3.1.2. Denetimsiz öğrenme yöntemi

Öte yandan, büyük miktarlarda serbestçe kullanılabilir veriler büyüdükçe ve denetimli bir yaklaşım kullanmak için yeterli veriyi etiketlemek artık mümkün değildir. Bu, ham verilerdeki kalıpların bilinip bilinmediğini veya beklenen çıktının ne olduğunu göstermek için makinenin istenen çıktı üretip üretmediğini değerlendirmeyi zorlaştırır. Gözetimsiz yaklaşım, bir insan için saf gürültü gibi görünen örüntüleri bulmak için kullanılır.

Gözetimsiz öğrenme yöntemlerinde, yalnızca kendileri için girdi değeri belirtilen ve çıktı hakkında doğru bilgi bulunmayan bir dizi eğitim örneği dikkate alınır. Kümelenmeye dayalı yaklaşımlar, herhangi bir insan katılımı, dilbilim bilgisi veya eğitim süresi olmadan orta derecede doğru analiz sonuçları üretebilir (Li ve Liu, 2014).

Kümeleme, henüz toplanmamış ve önceden tanımlanmış sınıfların bulunmadığı veri toplamalarında bir yapı bulmayı amaçlayan denetimsiz bir yöntem olarak kabul edilir. Diğer bir deyişle, kümelenme, her gruptaki üyelerin belirli bir bakış açısından birbirine benzediği farklı gruplara veri koymaktır. k-means, bilinen kümeleme sorununu çözen en basit denetimsiz makine öğrenme algoritmalarından biridir. Bu algoritmanın amacı, K değişkeninin temsil ettiği grup sayısı ile verilerdeki grupları bulmaktır. Algoritma, sağlanan özelliklere göre her bir veri noktasını K gruplarından birine atamak için yinelemeli olarak çalışır. Veri noktaları özellik benzerliğine göre kümelenir.

(Claypo ve Jaiyen, 2015) K-means gelen kümelenme ve MRF özellik seçim tekniğini kullanarak Tay restoran incelemesinde ilgili özellikleri seçmek için bir görüş önermektedir. Bu teknik özellik sayısını ve hesaplama sürelerini azalttı. Daha sonra K-means, Tayland restoranları incelemesini iki olumlu ve olumsuz gruba kümelemek için uyarlandı. Çalışma, K-means kümelemenin MRF özellik seçimi ile uyumlu olduğu ve kümelemede en iyi performansı sağlayabileceği sonucuna vardı.

2.3.1.3. Yarı-denetimli öğrenimi yöntemi

Yarı-denetimli makine öğrenimi, yukarıdaki iki seçeneğin en iyi kısımlarını uygular. Eğitim seansları sırasında, bir denetim otoritesi tarafından daha az miktarda veri etiketlenir ve eğitim verilerinin geri kalanı etiketlenmez. Makine, önceden etiketlenmiş verilerin yardımıyla verileri uygun bir şekilde kümelendirmelidir. Yarı-denetimli öğrenme yaklaşımları daha az insan çabası gerektirir ve daha yüksek doğruluk sağlar, bu nedenle görüş madenciliği alanında büyük ilgi görür (Zhu, 2008).

Duygu analizi için yarı-denetimli yönetimler son zamanlarda büyük ilgi gördü. İlişkili bir model kullanarak Twitter verilerindeki farklı tiplerde duygusal sinyalleri değerlendirmek için yarı-denetimli bir yaklaşım (Tang ve ark., 2015) tarafından önerilmiştir. Model, etiketli ve etiketsiz veriler üzerinde çalışan kontrollü dönüşümlü yayılma ve yerleştirme süreçlerine dayalı ikili öğrenme sunar.

(Silva ve ark., 2016) Twitter'daki görüşleri analiz etmek için yarı denetimli bir çerçeve önermiştir. Dahası, daha iyi bir tweet sınıflandırması için kendi kendine eğitim yöntemini kullanmıştır. Gerçek veri kümesindeki deney sonuçları, önerilen çerçevenin Twitter görüş analizinde doğruluk artışına neden olduğunu göstermektedir.

2.3.2. Sözcük tabanlı yaklaşım

Bir sözcük, duygu kutupları ve güç değerleri ile birlikte duygu kelimelerinin bir kelimesidir. sözcük tabanlı sınıflandırıcı yalnızca kelimeler ve anlamsal yönelimleri içeren bir polarite belgesine ihtiyaç duyar ve kullanımdan önce eğitilmesi veya başka bir şekilde işlenmesi gerekmez (Mudinas ve ark.,2012). Bu yaklaşımda iki yöntem vardır: Sözlük Tabanlı Yöneten ve Korpus Tabanlı Yöneten.

2.3.2.1. Sözlük tabanlı yaklaşım

Sözlüğe dayalı yaklaşım stratejisi, fikir çekirdeği sözcüklerini bulmaya bağlıdır ve daha sonra WordNet gibi sözlükleri eş anlamlıları ve zıtlıkları aramaya başlar (Machová ve Marhefka, 2013).

(Sharma ve ark., 2014) Cümlelerden özelliği ve görüşleri alan ve verilen cümlenin her özellik için pozitif, negatif veya nötr olup olmadığını belirleyen Unsur tabanlı bir fikir madenciliği sistemi önerdi. Cümlelerin semantik yönelimini belirlemek için denetimsiz yaklaşımın sözlük tabanlı tekniği benimsenmiştir. Fikir kelimeleri ve eş anlamlılarını ve zıt anlamlılarını belirlemek için WordNet sözlük gibi kullanılır. Deney, Amazon Hindistan'dan cep telefonlarının müşteri incelemeleri kullanılarak gerçekleştirildi. İncelemelerin yapıldığı ürünün tüm özellikleri tanımlanacak ve her özellik için cümlenin yönü belirlenecektir. Verilen cümlenin polaritesi, görüş kelimelerinin çoğunluğu temelinde belirlenir. Sonuçta sistem, kullanıcıların ürünü satın alıp almayacaklarına karar vermeleri için daha kolay okuyacak, analiz edecek ve onlara yardımcı olacak olumlu, olumsuz ve tarafsız cümlelerin akıllıca bir özetini oluşturacaktır.

2.3.2.2. Korpus tabanlı yaklaşım

Korpus tabanlı yaklaşım, sözdizimsel kalıplara veya büyük bir korpustaki diğer fikir kelimelerini bulmak için bir tohum fikir kelimesi listesi ile birlikte ortaya çıkan kalıplara bağlıdır. Bir dizi görüş kelimesi ile başlar ve daha sonra içeriğe özgü yönelimleri olan görüş kelimelerini bulmaya yardımcı olmak için büyük bir grupta başka görüş kelimeleri bulur. Bu, tohum görüş kelimelerini bulmak için istatistiksel yöntemler kullanılarak veya doğrudan duyarlılık değerleri vermek için semantik yöntemler kullanılarak ve kelimeler arasındaki benzerliği hesaplamak için farklı prensiplere dayanarak yapılabilir (Medhat ve ark., 2014). Korpus tabanlı yaklaşım, bağlama özgü yönelimlerle fikir kelimeleri bulma sorununu çözmeye yardımcı olur.

(Hu ve ark., 2012) Çevrimiçi inceleme manipülasyonunu tespit etmek ve tüketicilerin manipüle edilmiş incelemelere sahip ürünlere nasıl tepki verdiğini değerlendirmek için basit bir istatistiksel yöntem önerir. Hakemlerin yazım tarzını, derecelendirmeler, manipülasyonlar yoluyla manipülasyonun etkinliği ve okunabilirliği araştırdılar. Amazon.com'dan kitap incelemeleri üzerinde çalıştılar ve ürünlerin yaklaşık %10,3'ünün çevrimiçi inceleme manipülasyonuna tabi olduğunu keşfettiler.

Semantik yaklaşım birçok uygulamada, duygu analizinde kullanılacak fiillerin, isimlerin ve sıfatların tanımlanması için bir sözlük modeli oluşturmak için kullanılır (Maks ve Vossen, 2011) Modelleri aktörler arasındaki ayrıntılı öznellik ilişkilerini tanımladı her aktör için ayrı tutumları ifade eden bir cümlede. Bu öznellik ilişkileri hem tutum sahibinin kimliği hem de tutumun yönelimi (pozitif - negatif) ile ilgili bilgilerle etiketlenmiştir. Onların modeli, duygu analizi ile ilgili anlamsal kategoriler halinde bir sınıflandırma içeriyordu. Tutum sahibinin belirlenmesi, tutumun polaritesi ve metinde yer alan farklı aktörlerin duygu ve duygularının tanımlanması için araçlar sağlamıştır. Çalışmalarında Hollandalı WordNet'i kullandılar. Sonuçları konuşmacının öznelliğinin ve bazen aktörün öznelliğinin güvenilir bir şekilde tanımlanabileceğini gösterdi.

2.3.3. Hibrit yaklaşımı

Duygu analizine yönelik melez yaklaşım, hem makine öğrenimini hem de sözcük tabanlı yaklaşımları, her iki yaklaşımın da güçlü yanlarından faydalanacak, ancak bazı zayıf yönlerinden kaçınacak şekilde birleştirir.

Duygu analizi tekniğinde Makine öğrenimi ile sözcük tabanlı yaklaşımların kombinasyonu çok yaygındır. (Mukwazvure ve Supreethi, 2015), Guardian web sitesinin Teknoloji, Politika ve İş bölümleri hakkındaki haber yorumlarından fikir kazanmak için sözlük tabanlı tabanlı yöntem ve makine öğrenme algoritmalarını kullanan hibrit bir yaklaşım önerdi. lexicon kutupları tespit etmek için, sonra lexicon sonucu SVM ve KNN öğrenme algoritmalarını eğitmek için kullanılır. Deneysel sonuçlar SVM'nin KNN'den daha iyi performans gösterdiğini göstermektedir. KNN'nin K arttıkça üçüncü sınıfı tanımlamaması, daha az sayıda nötr eşyaya bağlanabilir.

Bulanık mantığın duygu analizi yaklaşımlarıyla hibridizasyonu da literatürde bulunmuştur. Örneğin (Appel ve ark, 2016), duyguları analiz etmek için duyarlılık sözlüğünü semantik kurallar, bulanık kümeler ve denetimsiz makine öğrenme yöntemleri ile birleştirir. Hibrit standart sınıflandırma önce yapılır ve daha sonra film incelemesi veri kümesinde dilsel sınıflandırmayı içeren gelişmiş bir yaklaşıma dönüştürülür. Hibrit yaklaşımın sonuçları NB ve maksimum entropiden daha iyidir.

2.4. Duygu Sınıflandırma Algoritmaları

Bu bölümde deneylerimizde kullandığımız sınıflandırma algoritmalarını tanıtacağız.

2.4.1. Naïve Bayes (NB)

NB en basit, güçlü ve yaygın olarak kullanılan bir makine öğrenme sınıflandırıcı modelidir. Bir örneğin her özelliğinin, bu örneğin başka herhangi bir özelliği ile ilgili olmadığı varsayımına dayanır. Örnek kategorisine c ait olma olasılığı i, (Denklem 2.5) ile ifade edilebilir (Duda ve Hart, 1974). Verilen belgeye c sınıfı atanır.

$$\hat{C} = \arg\max_{c \in C} p(c|d) \quad (2.5)$$

Temel olasılık modeli "bağımsız özellik modeli" olarak tanımlanabilir. NB sınıflandırıcısı Bayes kuralı (Denklem 2.6) kullanır.

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)} \quad (2.6)$$

Burada, $p(d)$ seçiminde hiçbir rol oynamaz C^* $p(d|c)$, Terimini tahmin etmek için, NB, denklem (2.7) 'de olduğu gibi d'nin sınıfı göz önüne alındığında, f_i'lerin şartlı olarak bağımsız olduğunu varsayarak onu ayrıştırır,

i örneğinin c kategorisine ait olma olasılığı (Denklem 2.7) ile ifade edilebilir.

$$P_{NB}(c|d) := \frac{p(c)(\prod_{i=1}^{m} p(f_i|c)^{n_i(d)})}{p(d)} \quad (2.7)$$

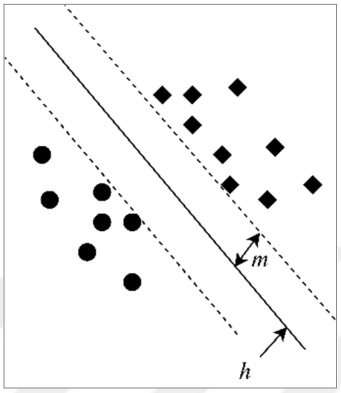
Burada, m özellik sayısı ve f_i özellik vektörüdür. $p(c)$ ve $p\ (f_i\ |c)$ bağıl frekans tahmininden oluşan bir eğitim yöntemi olarak düşünülebilir.

(Bilal ve ark, 2016) bir blogda Urduca ve İngilizce görüşlerini sınıflandırmak için üç teknik, yani Naive-Bayes, karar ağacı ve en yakın komşunun etkinliklerini karşılaştırdı. Onların sonuçları, NB'ın diğer iki teknikten daha iyi bir performansa sahip olduğunu göstermektedir. Çeşitli görüş madenciliği yöntemleri olasılıksız bir sınıflandırıcı olarak NB kullanılmıştır.

2.4.2. Destek Vektör Makineleri (SVM)

SVM, en iyi metin sınıflandırma algoritmalarından biri olarak kabul edilmektedir ve problem doğrusal olarak ayrı olduğunda sağlamdır. SVM'nin ana konsepti, farklı sınıf üyelikleri olan bir dizi nesne arasındaki doğrusal ayırıcıları belirlemektir (Moraes ve ark., 2013). (Şekil 2.2.)'de olduğu gibi, iki sınıfa (kareler ve daireler)

sahibiz ve bu sınıflar arasında birbirinden ayrılan en iyi hiperdüzlem h 'yi (doğrusal kernel) tanımlarız. Marj m, en yakın veri noktası ve hiperdüzlem arasındaki maksimum mesafedir.



Şekil 2.2. Bir SVM doğrusal olarak iki sınıfı ayırıyor.

Sınıflar için sadece ikili sınıflandırıcı olduğunu varsayalım: y = -1; 1 ve doğrusal olarak ayrılabilir eğitim seti $\{(x_i, y_i)\}$, böylece şartlar (Denklem 2.8) karşılandı.

$$w.x_i + b \geq 1 \qquad if\ y_i = 1$$
$$w.x_i + b \leq -1 \qquad if\ y_i = -1 \tag{2.8}$$

(Denklem 2.9), şartları (Denklem 2.8) bir takım eşitsizliklerle birleştirir.

$$y_i.(w_0.x + b_0) + b \geq 1 \qquad \forall_i \tag{2.9}$$

SVM, her iki sınıfı da maksimum kenar boşluğu ile ayıran optimal hiper düzlemi (Denklem 2.10) arar.

$$w_0.x + b_0 = 0 \tag{2.10}$$

Formül (Denklem 2.11), sınıflar arasındaki mesafeyi w ile verilen yönde ölçer.

$$d(w, b) = \min_{x,y=1} \frac{x.w}{|w|} - \max_{x,y=-1} \frac{x.w}{|w|} \tag{2.11}$$

(2.12) denkleminde ifade edilen optimal hiperdüzlem, d mesafesini maksimuma çıkarır $d(w,b)$ Bu nedenle, w_0 ve $b_0$parametreleri $|w_0|$. değerini maksimize ederek bulunabilir. Daha iyi anlamak için, Şekil 2.2'deki optimal ve optimal hiperplanları görümelidir.

$$d(w_0, b_0) = \frac{2}{|w_0|} \quad (2.12)$$

Sınıflandırma, nesnenin hiperdüzlemin hangi tarafında olduğuna dair basit bir karardır.

SVM sınıflandırıcıları, pek çok makalede kullanıldı , (Akaichi, 2013) Yararlı bilgiler elde etmek için SVM ve NB'ye dayalı bir yöntem önerdi ve kullanıcıların "Arap Baharı" döneminde "Facebook" mesajlarında Tunus kullanıcılarının durumlarının duygularını ve davranışlarını sınıflandırdılar. Ayrıca, çıkarılan durum güncellemelerinden gelen ifadeler, eklemeler ve kısaltmalar temelinde bir duyarlılık sözlüğü oluşturun. Dahası, iki sınıf öğrenme algoritması SVM ve NB arasında duygu sınıflandırması için bir eğitim modeli ile karşılaştırmalı deneyler yaptı.

2.4.3. Lojistik Regresyon (LR)

Maksimum Entropi olarak da bilinen LR, verileri ayrık sonuçlara sınıflandırmak için olasılıklı bir istatistiksel yöntemdir. 'Lojistik Regresyon' olarak adlandırılır, çünkü altında yatan teknik Doğrusal Regresyon ile tamamen aynıdır. Ancak en büyük fark, ne için kullanıldıklarında yatıyor. Bu model sadece regresyon için değil, sınıflandırma görevi için de kullanılır (Feng ve ark., 2014). Düşük varyans ve mükemmel verimlilik sağlayan makine öğrenme algoritmalarından biridir. lojistik regresyon formülü şu şekilde tanımlanabilir (Denklem 2.13):

$$y = f(x) = \frac{1}{1+e^{-z}} \quad (2.13)$$

Burada, y bağımlı değişkendir, x özelliğidir ve z, Fayda Fonksiyonu olarak tanımlanmakta ve aşağıdaki şekilde gösterilmektedir.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (2.14)$$

Bu formüle göre β_0 sabiti, $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$'ler ise tahmin değerleriyle çarpılan regresyon katsayılarını temsil etmektedir. n sayısı veri kümesi içerisindeki bağımsız değişkenlerin sayısını, ε ise hata terimini ifade eder.

(González-Ibáñez ve ark., 2011) 'in çalışması, Twitter verilerini kullanarak, yani olumlu veya olumsuz görüşleri doğrudan ileten alaycı tweet'leri ve alaycı olmayan tweetleri ayırt etmek için sorunu duygu analizi bağlamında inceledi. SVM ve lojistik regresyon kullanılarak denetimli öğrenme yaklaşımı benimsenmiştir. Özellikler olarak, unigram ve bazı sözlük tabanlı bilgiler kullandılar. Üç yollu sınıflandırma için deneysel sonuçlar, sorunun çok zor olduğunu gösterdi. elde edilen en iyi doğruluk sadece %57 idi.

2.4.4. Karar ağacı

Bu sınıflandırıcı, cümle yapısını bir ağaç biçiminde tarif eder; Her ağaç bir kök düğümü, dallar ve yaprak düğümleri içerir. Her bir iç düğüm bir niteliği temsil eder, her bir dal bir testin sonucunu gösterir ve her bir yaprak düğümü bir sınıf etiketine sahiptir. Çok değişkenli analiz için çok basit ve güçlü bir tekniktir. (Rokach ve Maimon, 2015) Karar ağacı, iki yinelemeli adımda problemi çözer, ilk olarak veri setindeki safsızlığı temizleyen Entropi'yi aşağıdaki formül Denk (2.15) kullanarak hesaplar.

$$E(X) = \sum_{i=1}^{c} -P_i . log_2 P_i \quad (2.15)$$

P_i'nin sınıf i 'in olasılıkları olduğu ve kümedeki sınıf i oranı olarak hesaplandığı yer.

İkinci adımda, özellik vektörlerinin belirli bir özelliklerinin ne kadar önemli olduğunu bize gösteren bilgi kazancı hesaplanır. (Denklem 2.16).

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (2.16)$$

T 'nin ana kökün toplam entropisi olduğu ve X 'in çocuk (yaprak) düğümünün entropisi olduğu durumda. Yukarıdaki iki adım, veri kümesindeki anormallikler kaldırılıncaya kadar sürekli olarak tekrarlanır. Bu sınıflandırıcı sağlamdır ve kısa zamanda büyük verilerle iyi performans gösterir. Bu tekniğin dezavantajı, ağaçta birbiriyle ilişkili çok fazla düğüm varsa, karmaşıklığın çok yükselmesi ve bu açgözlü yapı nedeniyle metni belirli bir kategoriye sınıflandırmanın çok zor olmasıdır.

(Wakade ve ark., 2012) Twitter'dan toplanan tweet'lerin duygularını sınıflandırmak için yararlı bilgileri ayıklamak için Weka veri madenciliği araçlarının kullanımını sunar. Tweet madenciliğinin sonuçları, yeni tweet'lerin duygularını değerlendirmek için kullanılabilecek karar ağaçları olarak temsil edilmektedir. Karar ağacı öğrenme süreçlerinde tweet içeren tweetlerin etkisini değerlendirerek karar ağacı öğrenimi için t-işlemek için yeni bir yöntem geliştirdi. Yöntem, iPhone ve Microsoft ile ilgili tweetlerden duygu analizi yapmak için uygulanır. Deneysel sonuçlar, karar ağacı sınıflandırıcılarının (J48) algoritmasını NB'ın daha yüksek performans göstermektedir.

2.4.5. Çoğunluk oylama

Önceden anlatılan dört sınıflandırıcının tahminlerini birleştirmek için bir topluluk tekniği kullandık. Belirli bir sınıflandırma problemi için n farklı sınıflandırma algoritması verildiğini varsayalım, bu algoritmaları tek tek algoritmalardan daha üstün bir sınıflandırıcı üretecek şekilde birleştiriyoruz.

Her sınıf için en olası olasılığı seçme olasılığı arasında oylama yapılacaktır. Oylama aşağıdaki formüle göre yapılır. (Denklem 2.17).

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (2.17)$$

$$\hat{X} = \max\{X_1, X_2, \ldots, X_m\}$$

X'in tüm algoritmalarda sınıfın olasılığının toplamını temsil ettiği durumlarda, x_i belirtilen algoritmada sınıfın olasılığıdır, n algoritma sayısıdır, $\hat{X}$ aday gösterilen adaylar arasından seçilen sınıftır ve m sınıf sayısıdır.

2.5. Değerlendirme Metrikleri

Bir metin sınıflandırıcılarının performansını değerlendirmek için birkaç farklı yöntem vardır. Bu bölümde, bu tezde kullanılan puanlama metrikleri sunulmaktadır.

Bir karışıklık matrisi, sınıflandırıcının gerçek değerlerinin bilindiği bir dizi test verisindeki performansını tanımlamak için sıklıkla kullanılan bir tablodur (Tablo 2.2.). Gerçek Pozitif (TP), pozitif sınıfa ait olarak doğru sınıflandırılan pozitif duyguların sayısıdır. Yanlış Pozitif (FP), pozitif sınıfa ait olarak sınıflandırılan negatif duygulardır. Gerçek Negatif (TN), negatif sınıfa ait olarak doğru sınıflandırılan negatif duyguların sayısıdır. Yanlış Negatif (FN) pozitif duygu, negatif sınıfa ait olarak yanlış sınıflandırılmıştır.

Table. 2.2. Karışıklık matrisi.

		Gerçek	
		Pozitif	Negatif
Tahmın	Pozitif	TP	FP
	Negatif	FN	TN

2.5.1. Doğruluk

Doğruluk, denklem kullanılarak hesaplandığı şekliyle uygun sınıflandırılmış sonuçların tüm popülasyona oranıdır. (Denklem 2.18)

$$\text{doğruluk} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{2.18}$$

2.5.2. Duyarlılık

Duyarlılık, sınıfın öngörülen büyüklüğüne göre doğru tahmin edilen uygun orandır. (Denklem 2.19) Duyarlılık aşağıdaki gibi tanımlanır.

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.19}$$

2.5.3. Hassasiyet

Hassasiyet, bir sınıflandırıcı tarafından doğru bir şekilde tanımlanmış toplam pozitif örneklerin oranını ölçer. (Denklem 2.20) Hassasiyet aşağıdaki gibi tanımlanır.

$$\text{Hassasiyet} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.20}$$

2.5.4. F1-skor

F1-Skor, tek bir önlem almak için hassasiyeti ve duyarlılığı birleştirmenin bir yoludur. (Denklem 2.21) F1-Skor aşağıdaki gibi tanımlanır.

$$F1 - score = \frac{2 * \text{Duyarlılık} * \text{Hassasiyet}}{\text{Duyarlılık+Hassasiyet}} \tag{2.21}$$

BÖLÜM 3. VERİ TOPLAMA VE İLK ANALİZ

Bu bölümde, verilerin twitterdan nasıl toplandığına dair bir genel bakış ve hashtaglere, sözlere, retweetlere, favorilere, kelime sıklığına, ilgili terimlere ve metin ön işlemlerine göre tweet'lerin genel bir tanımı verilmektedir.

3.1. Twitter'dan Veri Toplama

Twitter, kullanıcıların genellikle tweet olarak adlandırılan 140 karaktere kadar metin girişi yazmalarına izin veren popüler bir mikro blog sitesidir. Kullanıcılar günlük yaşamlarıyla ilgili çeşitli konular hakkında görüş bildirmek için tweet yazarlar. 2019'un ilk çeyreği itibarıyla Twitter'da aylık 330 milyondan fazla aktif kullanıcı ve günde yaklaşık 500 milyon tweet bulunuyor (Statista, 2018). Bu bilgi miktarı, Twitter'ı büyük bir potansiyel sunan ideal bir platform haline getirir ve herhangi bir konuya karşı duyarlılık eğilimi kazanmak için kullanılabilir.

Twitter tarafından oluşturulan veriler Twitter'ın API'sı aracılığıyla kullanılabilir ve değerlendirilen verilerin gerçek zamanlı bilgi akışını temsil eder. Akış API'sı, anahtar kelime, kullanıcı, coğrafi alan veya rastgele bir örneğe göre filtrelenen belirli bir veri türü için istekte bulunarak ve ardından bağlantıda hata olmadığı sürece bağlantıyı açık tutarak çalışır. Twitter'dan veri toplamak için R'yi verileri toplamak için bir platform olarak kullandık. R, derin istatistiksel analize yönelik bir programlama dilidir (Zhao, 2012). Açık kaynaklıdır ve farklı platformlarda kullanılabilir. Artık görselleştirmeler ve veri madenciliği de dahil olmak üzere çeşitli uygulamalarda kullanılıyor.

Bu tez için iki terörist saldırının verileri aşağıdaki faktörlere göre toplanmıştır: Birincisi, operasyonun bir terörist eylemi olarak sınıflandırıldığı ve bir terör örgütü

tarafından kabul edildiği; İkincisi, iki operasyonun koşulları iki farklı ülkede uygulama ve yürütme tarafı ile benzer olmalıdır; Üçüncüsü, yorumların derlenebilmesi için yeterli materyale sahip olmak için operasyonun Twitter'da oldukça etkileşimli olması gerekir.

İlk operasyon 3 Haziran 2017'de Londra'da, bir kamyonun Londra Köprüsü'ndeki yayalara kasıtlı olarak çarpmasıyla gerçekleşti. Saldırganlar daha sonra, şüphelilerin restoran ve barlarda ve çevresinde birkaç kişiyi bıçakladığı Borough Market'e doğru yoluna devam etti. Saldırıda sekiz kişi öldü, 48 kişi yaralandı. Saldırganlar polis memurları tarafından vurularak öldürüldü ve sahte patlayıcı yelek giydikleri bulundu. Saldırıyı İslam Devleti (IŞİD) üstlendi (BBC News, 2017).

İkinci operasyon, Barselona - La Rambla'da geçen bir minibüsün çalıştırıldığı 17 Ağustos 2017'de Barselona - İspanya'da iki ay sonra gerçekleşti, 13 kişi öldü ve en az 130 kişi yaralandı, Saldırgan yaya olarak kaçtı. başka bir kişi bıçakla bıçaklandı. Terörist aynı terör hücresinden beş erkek, yakındaki Cambrils'teki yayalara sürülerek bir kadını öldürdü ve altı kişiyi yaraladı. Saldırganlar polis tarafından vurularak öldürüldü. IŞİD saldırının sorumluluğunu üstlendi (the Guardian, 2017).

Bu yazıda, LondonBiridge terimi Londra saldırısına atıfta bulunurken, LaRambla terimi Barselona saldırısına atıfta bulunmaktadır. Veriler, her saldırı hakkında yorum yapmak için Twitter'da kullanılan iki trend hashtag'i temel alınarak toplanmıştır ve sadece İngilizce dilinde yazılmıştır. Tablo 3.1.'de veri toplama tarihi, toplanan tweet sayısı ve veri toplamak için kullanılan hashtag'ler açıklanmaktadır.

Tablo 3.1. Toplanan terör saldırıları.

Saldırı	Dönemi	Hashtagleri	Tweet Sayısı
LondonBiridge	03 to 06/06/2017	#Londonattacks, #Londonbridge	12,582
LaRambla	17 to 22/08/2017	#Barcelonaattack, #barcelonaterrorattack	13,767

3.2. Ön Analiz

Tweet içeriğinin anlambiliminin saldırıların her biri ile nasıl ilişkili olduğunu anlamak için tweet içerik analizi yaptık. iki terörist saldırının cesedi için altı tweet

tabanlı özellik (metin, # hashtag, @kullanıcıadı, retweet ve favoriler) içeren bir özellik kümesi belirledik. Bu analiz aşağıdakileri göstermektedir:

- Retweet dağıtımı ve favori sayımlar
- Kullanılan trend hashtag
- En çok bahsedilen kullanıcı
- İki saldırıda kullanılan ortak kelimeler.
- En sık kullanılan kelimeler ve bu kelimeler arasındaki korelasyon.

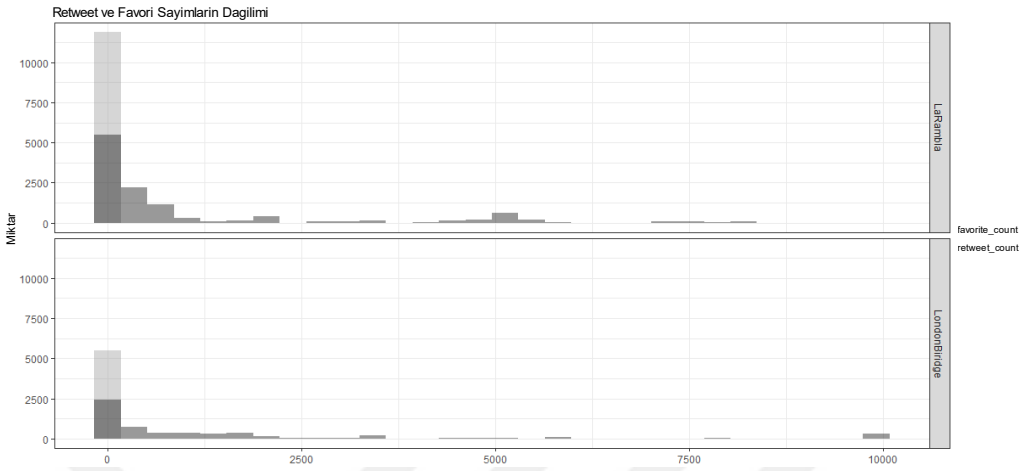
3.2.1. Retweet ve favoriler

Tekrar tweet atmak, bir tweet'in yeniden gönderilmesi ve tweet'in çevrimiçi kullanıcılar arasında sevildiğini veya popüler olduğunu gösteren göstergeler olarak tanımlandığı anlamına gelir. her iki saldırıdaki tweet'lerin çoğunun retweetlendiğini veya favorilere eklenmediğini, ancak retweetleme ve favori arasındaki karşılaştırmada, favori olmayan tweet'lerin retweet sayısının iki katı olduğunu bulduk (Şekil 3.1.).

Bu tweetleme, tweetlerinin daha geniş bir kitleye görünmesini amaçladı, yani retweetlemenin Twitter kullanıcılarının en çok kullanılan özelliği olduğu ve retweetlemenin sebebi daha geniş bir kitleye görünür olması ve bir trend haline getirilmesi anlamına geliyor.

3.2.2. Hashtag

Hashtag bir anahtar kelime veya bir konuyu veya temayı tanımlamak için kullanılan bir kelime öbeğidir. Twitter'da, dikkat çekmek, düzenlemek ve aramada daha kolay gösterilmesini sağlamak ve tanıtmak için alakalı bir anahtar kelimenin önündeki hashtag sembolünü (#) kullanın.



Şekil 3.1. Retweet ve Favori sayıların dağılımı.

Tüm hashtag'leri tüm tweet setinden çıkarırız, sonra her bir korpus için en sık hashtag'leri çıkarmak için analiz ederiz (Şekil 3.2.). Veri toplamak için kullanılan hashtag'lerin aksine, Tweetlerde bu terörist saldırılarla ilgili en yaygın kullanılan hashtag'lerin ayrılabileceğini not ediyoruz:

- Hashtag'ler, saldırının #londonbridge, #london, #spain, #rambla, #boroughmarket gibi hedef yerlerini belirtir.

- Hashtag'ler operasyonu #terrorist, #terrorism, ve #terroristattack saldırısı gibi bir terör saldırısı olarak nitelendirdi.

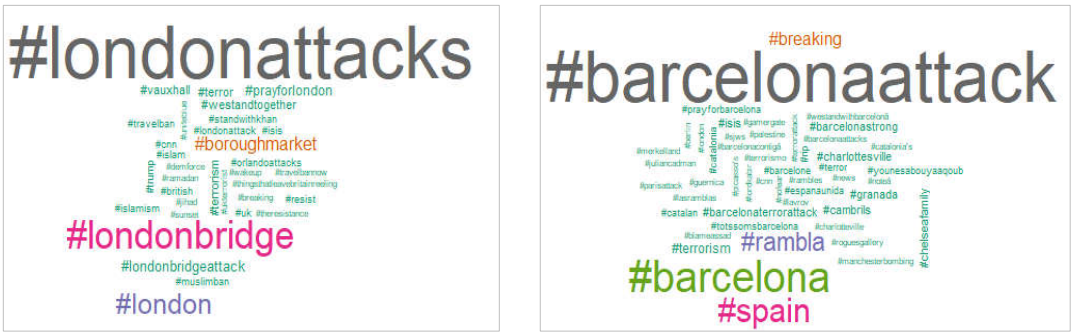
- #barcelonastrong, #standwithkhan, #westandtogether gibi destek ve dayanışma için hashtag'ler.

- Saldırı kurbanları için dua etmek için hashtag'ler, örneğin, #prayforlondon, #prayforbarcelona, #rip.

- Hashtag'ler saldırının arkasında olmakla suçlanan bir kişi veya terörist grubu ifade eder, örneğin #isis, #younesabouyaaqoub

- Hashtag'ler #islam, #ismalmism ve #muslimban gibi İslam dinini ifade eder.

- #news ve #breaking gibi medya veya haberler için hashtag işaretleri.



Şekil 3.2. En çok kullanılan hashtag'leri görüntüler.

3.2.3. Kullanıcılar adını

Bir söz, Tweet'in gövdesinin herhangi bir yerinde başka bir @kullanıcıadı içeren bir Tweet'dir, etiketlemek veya bir kullanıcının profiline başvurmak için kullanılır.

Tablo 3.2.'de veri setlerinde en çok etiketlenen hesapları ve bu hesapların sınıflandırılmasını gösterir. Çoğu medyanın, politik analistlerin ya da politikacıların hesaplarına, bazıları da polise ya da devlet dairelerinin hesaplarına atıfta bulunuyor. Ayrıca, Twitter'dan silinen hesaplar listede görünüyor.

Tablo 3.2. En çok belirtilen 15 kullanıcı adının sıklığını gösterir.

	LaRambla			LondonBiridge		
	Mention	n	Type	Mention	n	Type
1	@guardiacivil	629	police	@DineshDSouza	307	Analyst
2	@OnlineMagazin	230	media	@RealJamesWoods	241	celebs
3	@JoeBiden	181	Politician	@realDonaldTrump	209	Politician
4	@FrankRGardner	176	Analyst	@josephwillits	177	Politician
5	@XHNews	166	media	@tedlieu	169	Politician
6	@mossos	126	Police	@KTHopkins	163	journalist
7	@PDChina	84	media	@MuslimIQ	103	politician
8	@BasedMonitored	78	Suspended	@BBCBreaking	86	media
9	@ANC_USA	68	government	@BBCNews	70	media
10	@AFP	64	media	@StewartWood	68	Politician
11	@SputnikInt	62	media	@TheCartelGavin	60	Suspended
12	@tzounakis	62	journalist	@BasedMonitored	48	Suspended
13	@mendcommunity	60	Community	@MaajidNawaz	47	journalist

Tablo 3.2. (Devamı).

	LaRambla			LondonBiridge		
	Mention	n	Type	Mention	n	Type
14	@AmichaiStein1	56	journalist	@PoliticalShort	43	media
15	@BBCBreaking	55	media	@AlwaysActions	37	media

3.2.4. Tekrarlı kelimeler

Her saldırı için en sık kullanılan kelimeleri çıkarmak. Önce topluluğu (bir sonraki bölümde bu süreci açıklayacağız) hashtaglerden, linklerden, söz isimlerinden ve noktalama işaretlerinden, URL'lerden temizleriz, sonra tweetlerde bulunan kelimelerin sayısını özetleriz. Her saldırıda en sık kullanılan 20 kelime (Şekil 3.3.).

Terörist, insanlar, polis, olay, saldırı, kırma, IŞİD, öldürüldü, vb. gibi iki saldırı arasında ortaya çıkan ortak kelimeleri de çıkarıyoruz.

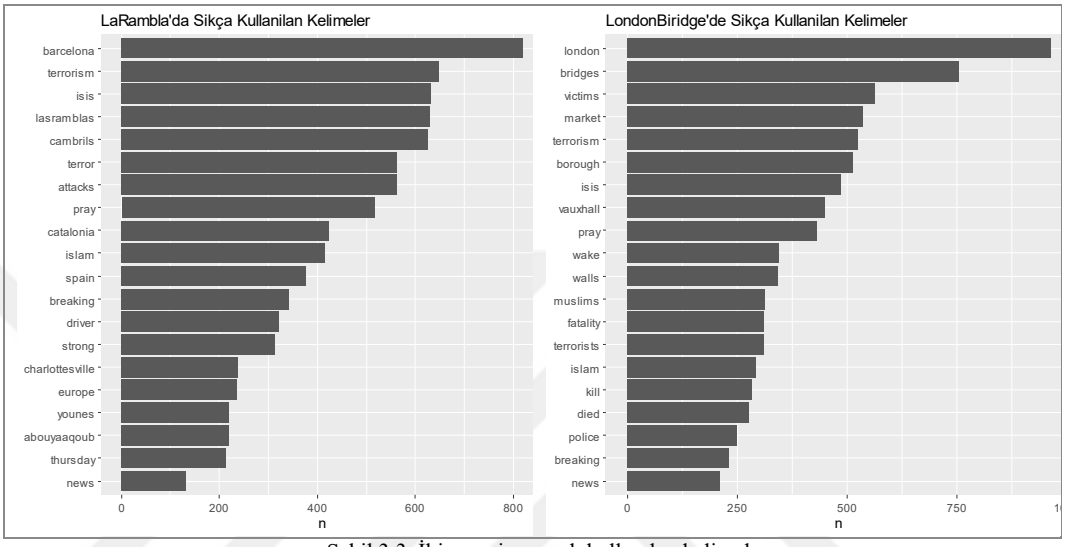
Ayrıca tweetlerdeki benzersiz kelimeleri ve kelimelerin uzunluğunu da hesaplıyoruz, sonuç tweet'lerin çoğunun 5 ila 10 kelime arasında bir orandan oluştuğunu, Ancak LaRambla söz konusu olduğunda bu oranda tweet sayısının görüldüğü gibi LondonBiridge'deki tweet sayısının neredeyse iki katı (Şekil 3.4.).

Benzersiz kelimeler söz konusu olduğunda, LaRambla'daki benzersiz kelimelerin toplam kelime sayısına oranının 3.18, Londra'da ise Biridge'nin 5.81 kelime olduğunu bulduk.

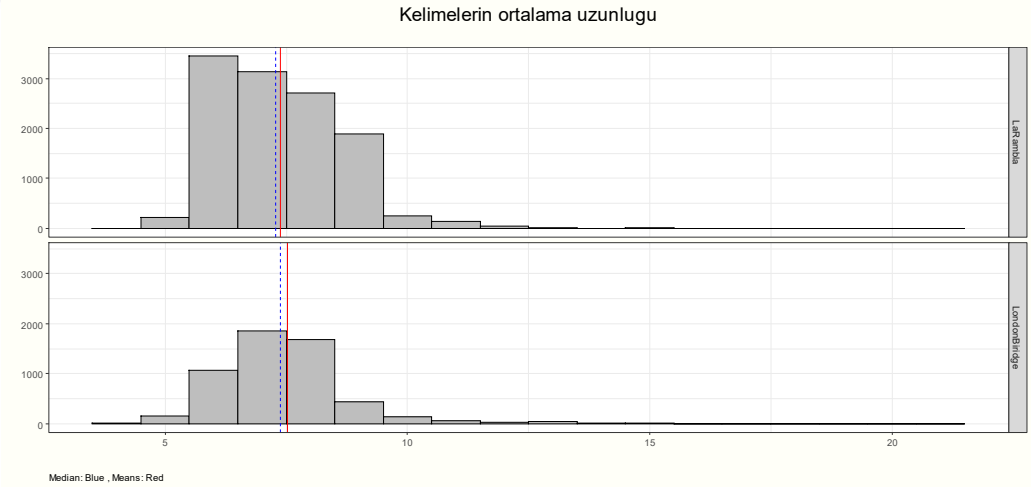
3.2.5. İlişkili terimler

Kelimelerin korpus içindeki sıklığını ve dağılımını anlamak yeterli değildir, genellikle bir korpus içindeki kelimeler arasındaki ilişkiyi anlamak isteriz. İlişkili terimler, korpustaki iki kelime arasındaki ilişki derecesini tanımlar. Bu kelimeler arasındaki ilişkiyi görselleştirmek için ağ analizini kullandık. (Şekil 3.5.) iki Corpus'ta kelime içindeki görsel korelasyonları göstermektedir. Burada korelasyonun oldukça yüksek olduğu kelime ağlarına bakıyoruz. Korpuste ilişkilendirilmiş en belirgin kelimeleri gösterir, Yani, örneğin Barcelona kelimesini seçersek, bu

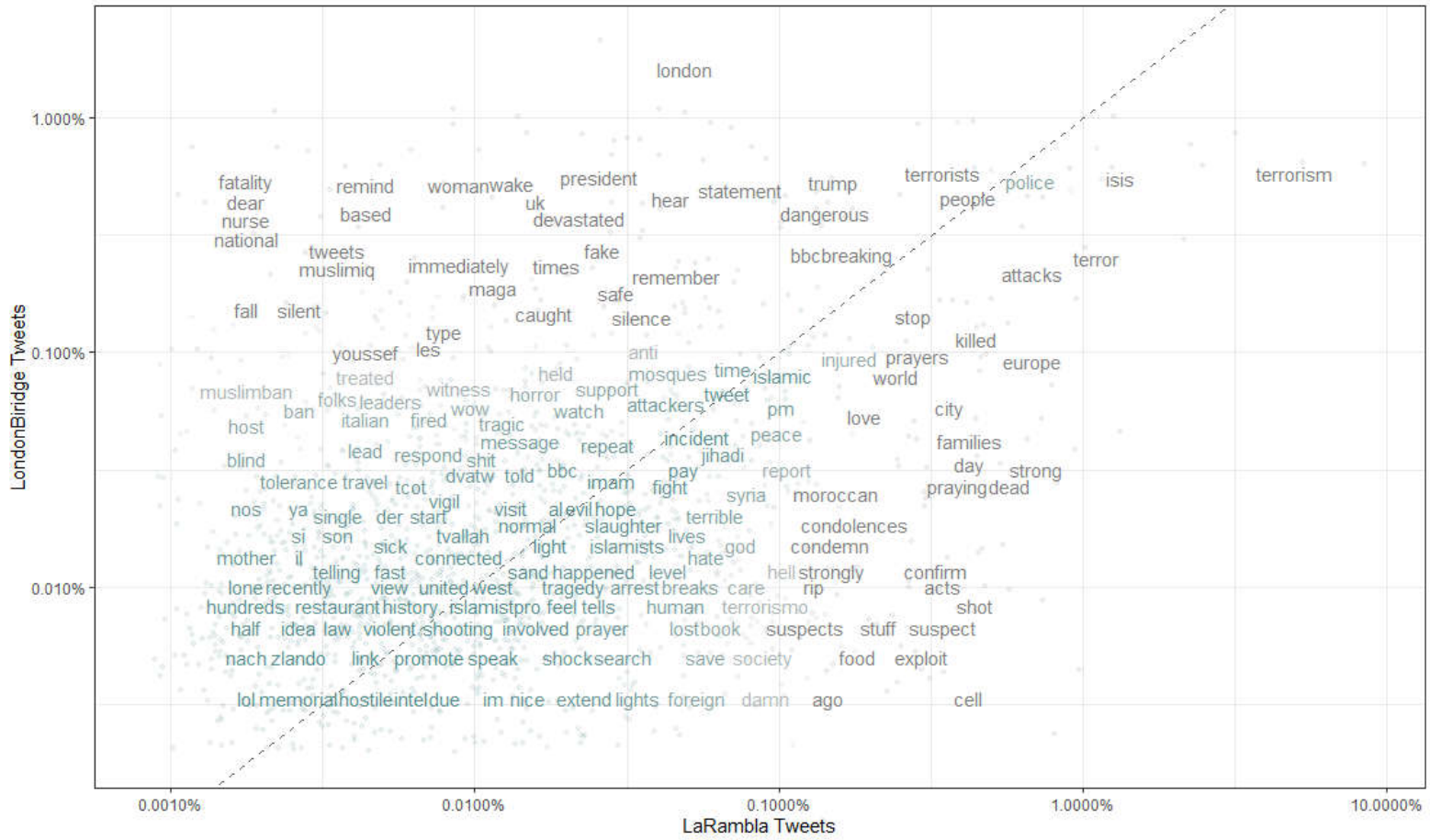
kelimenin diğer kelimelerle ilişkisini daha derin görebiliriz. rakamlar bize bunları gösteriyor. terim korelasyonu ve ağ analizi, veri kümesini özetlememize ve onu görsel olarak anlamamıza yardımcı olabilir.



Şekil 3.3. İki opersiyonu sık kullanılan kelimeler.

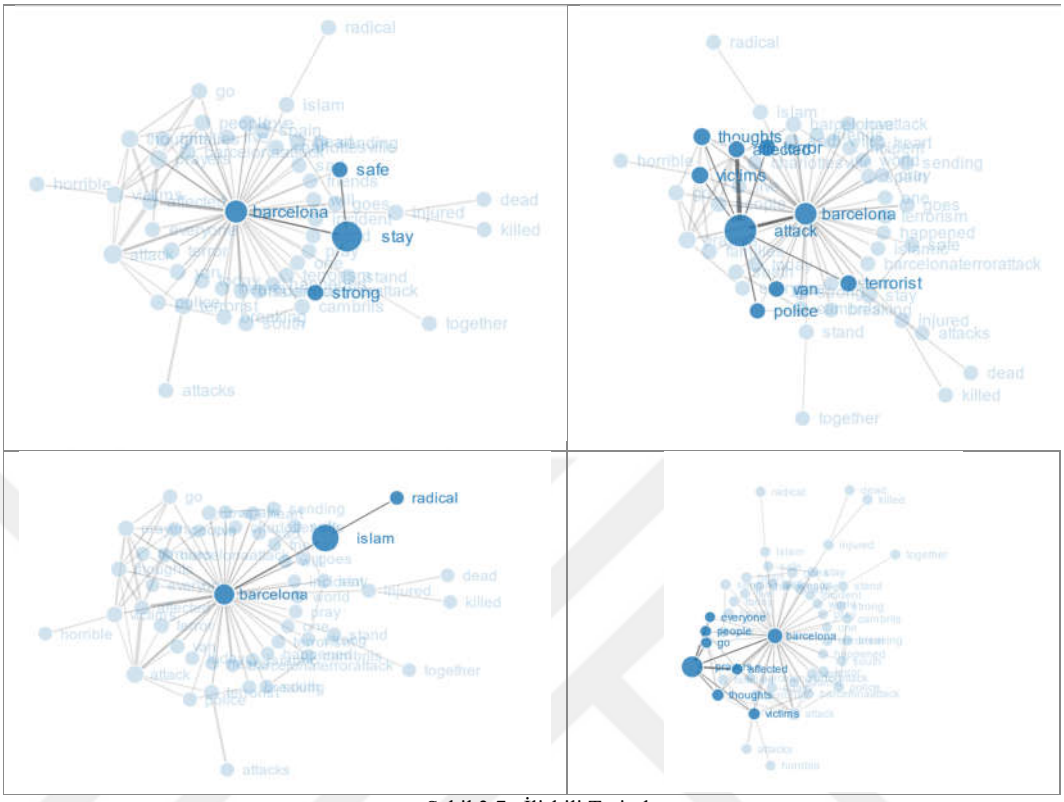


Şekil 3.4 Ortalama Kelime Uzunluğu.



Şekil 3.1. Yatay şekil kullanım örneği. Yazının ikinci satıra geçmesi durumunda yazı şekil numarasından sonra başlatılmalıdır.

Şekil 3.6. En İlişkili Terimler.



Şekil 3.7. İlişkili Terimler.

3.3. Metin Önişleme

Önişleme, tam teşekküllü bir analiz başlatmadan önce önemli bir adımdır. Twitter da dahil olmak üzere birincil kaynaklardan toplanan tüm veriler önemli miktarda gürüntü içerir. Örneğin, Twitter verileri semboller, URL'ler, ifadeler vb. İçerir. Bu ifadeleri normal metne dönüştürmeye ihtiyaç vardır (Tablo 3.3.).

- Retweet'ler (RT), hashtag'ler (#), Kullancı adını (@), URL'ler, ifadeler, noktalama işaretleri ve yinelenen tweetler gibi özel sembolleri otomatik olarak kaldırıyoruz.

- Hashtags “#”: Kullanıcılar konuları işaretlemek için hashtag kullanır. Twitter kullanıcıları tarafından tweetlerini daha büyük bir kitleye görünür kılmak için kullanılır.

- Retweet “RT”: bir başkasının önceki tweet'inin bir tekrarı olduğunu belirtmek için kullanılır.

- İfadeler: Noktalama işaretleri ve harfler kullanılarak yüz ifadelerinin resimsel temsilidir. İfadelerin amacı kullanıcının ruh halini ifade etmektir.

- URL: Başka bir web sayfasına ve web sitesine bağlantı.

- “@” deyin: Twitter kullanıcıları, üzerindeki diğer kullanıcılara başvurmak için “@” sembolünü kullanır.

- Noktalama işaretleri: Net bir tarzda yazmak için kullandığınız nokta, virgül veya soru işareti gibi bir işaret.

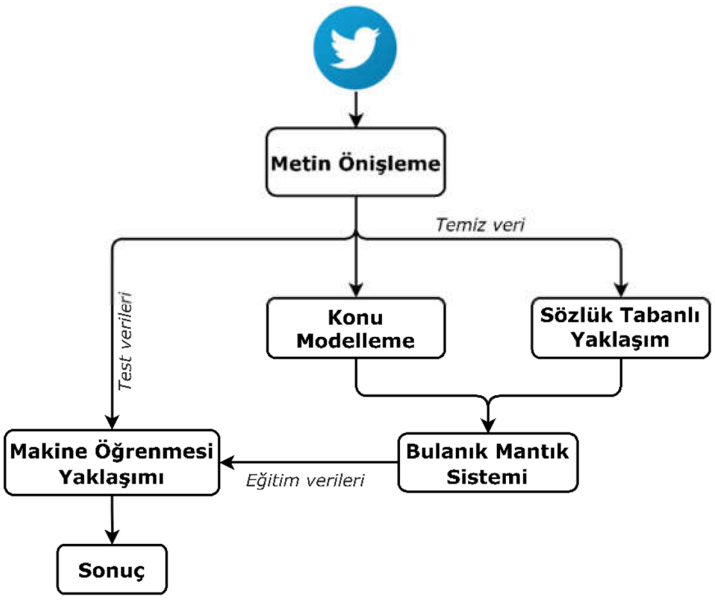
- Yinelenen tweet: başka bir kullanıcı tarafından tekrar tweetlenen bir Tweet. fazlalığı azaltmak için onu kaldırıyoruz.

Tablo 3.3. Bir tweet'in temizlenmesi işlemlerini göstermektedir.

Süreç	Metin
Orijinal Tweet	RT @JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism. ☹ #Barcelona https://t.co/NQEjjuyoev
Hashtag'leri Kaldır	RT @JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism. ☹ #Barcelona https://t.co/NQEjjuyoev
Retweet'i Kaldır	RT @JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism. ☹☹☹ https://t.co/NQEjjuyoev
Emoticons	@JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism. ☹☹☹ https://t.co/NQEjjuyoev
URL	@JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism https://t.co/NQEjjuyoev
Kullancı adı	@JunckerEU: My thoughts are with the people of Barcelona. We will never be cowed by such barbarism.
Noktalama	: My thoughts are with the people of Barcelona We will never be cowed by such barbarism.

BÖLÜM 4. KONU-BAĞIMLILIK DUYGU ANALİZİ

Bu bölümde duygu analizi, konu modelleme ve bulanık mantık sistemi arasında bir melez olan önerilen modeli tanımlayacağız. Bu modelde Twitter'dan toplanan iki benzer terörist saldırının tepkilerini kullanılıp, bu reaksiyonların ana konuları, insan kararıyla LDA konu modelleme algoritması kullanılarak çıkarılmıştır. Fikir kelimelerini puanlamak için kullanılan sözlük yaklaşımı, tweetlerdeki tüm görüş kelimelerinin, görüş sözlüğündeki kelimelerle eşleştirilerek tanımlandığı tweetlerden oluşmaktadır. Daha sonra polariteyi belirlemek ve tweet'i duygusal sınıflandırma için etiketlemek için bir bulanık mantık sistemi kullanıldı. POS ve N-gram olarak özellikler çıkarılmıştır. Son olarak, veri setindeki tweet'lerin duyarlılığını sınıflandırmak için makine öğrenme yöntemleri kullanılmaktadır. Şekil (4.1.) bu modelin ana yapısını göstermektedir.



Şekil 4.1. Önerilen Model Yapısı.

4.1. Konu Modelleme

Konu modelleme, bu tür belgelerin nümerik veriler üzerinde kümelemeye benzer şekilde denetimsiz sınıflandırılması için bir yöntemdir ve aradığımız şeyden emin olmadığımızda bile doğal öğe gruplarını bulur (Jelodar ve ark., 2019). Konu modelleme, bir metin gövdesindeki gizli semantik yapıların keşfi için sık kullanılan bir metin madenciliği aracı türüdür.

Konu modelleri, büyük miktarlarda etiketlenmemiş metni analiz etmek için basit bir yol sağlar. "Konu", sık sık birlikte ortaya çıkan bir kelime kümesinden oluşur. Bir konu modellemesi, kelimeleri benzer anlamlara bağlayabilir ve kelimelerin çoklu anlamlarla kullanımlarını ayırt edebilir. Konu modellemenin kullanılması metnin tam anlamını vermez, aksi halde elde edilemeyen konulara iyi bir genel bakış sağlar.

Bir metinden konu elde etmek için birçok yaklaşım vardır. Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation LDA) en popüler konu modelleme tekniğidir ve bu çalışmada kullanılmıştır.

Tekniğin bilinen durumunda, LDA'nın birçok makalede en çok kullanılan olduğunu buluyoruz ve araştırmacıların çoğu LDA'nın konu modelleme için tercih edilen yöntem olduğunu savunuyor. LDA, belgelerin bir konu karışımından üretildiğini varsayar (Blei, Ng and Jordan, 2003). Bu konular daha sonra olasılık dağılımlarına göre kelimeler oluşturur. Belgeler veri kümesi göz önüne alındığında, LDA bu belgeleri ilk başta hangi konuların yaratacağını bulmaya çalışır.

Diyelim ki belgelerde w sayıda kelime var. $P(w|d)$ dokümanı verilen bir kelimenin olasılığının şöyle olduğunu söyleyebilmektir:

$$\sum_{t \epsilon T} p(w|t,d)p(t|d) \qquad (4.1)$$

$P(t|d)$ belgelerindeki konuların olasılık dağılımı, $P(w|t)$ konudaki kelimelerin olasılık dağılımı, T toplam konu sayısıdır.

Konu modelleme algoritmaları, fikir modellemesinde artan bir ilgi görmektedir. (Fang ve ark., 2012) 'de, zıt fikir modellemesi için LDA temelli yeni bir denetimsiz konu modelini Çapraz Perspektif Konu (CPT) olarak önerdi, verilen bir konuya ve konudaki yeterlilik ölçütlerine göre farklılıklarına göre birden çok görüşten görüş bulmayı amaçlamaktadır. Siyasi alandaki iki veri kümesi üzerinde hem nitel hem de nicel tedbirlerle deneyler yaptılar. Böylece İki temel hat olarak corrIDA ve LDA kullanılan diğer modellerle yaklaşımlarını değerlendirmek mümkün oldu.

Sadece 140 karakter uzunluğundaki tweetlerden konu çıkarmak, önemli bir araştırma problemi haline geliyor. Her bir tweet'teki birçok kelime biraz rastgele olduğu için, tweetler bireysel belgeler olarak, konular belirgin görünmeyebilir. Ancak konu modelleme ile, tweet'lerde aktarılan temel konuları anlamak kolaydır ve oluşturulan konular bireysel tweetlerden daha tutarlı görünür.

(Alvarez-Melis ve saveki, 2016) tweetleri konuşmaya dayalı olarak daha uzun belgelere toplamak için yeni bir plan öneriyorlar. Bu çerçevede, bir belge bir tohum tweet'inden, diğer kullanıcılar tarafından gönderilen cevaplara yazılan tüm tweet'lerden ve orijinal posterin bunlara verdiği yanıtlardan oluşur. Bu tekniğin arkasındaki motivasyon, Twitter'daki kullanıcı-kullanıcıya etkileşiminin birkaç ilgili konu etrafında dönme eğilimi göstermesidir, bu nedenle tweet'leri konuşmalara göre birleştirmek daha tutarlı bir belge toplama ve dolayısıyla daha alakalı konu çıkarmaya yol açabilir.

(Hong ve Davison, 2010) tweet yazarlarının modellenmesi için LDA'ya daha basit bir uzatma önermektedir. Bir yazar tarafından yazılan tweetleri tek bir belgede toplayarak, tweet'lerin seyrek doğasından kaynaklanan dezavantajları azaltırlar. Aynı yazardan gelen mesajları birleştirme yönteminin kısa metin ortamlarında sınıflandırma ve konu modeli görevlerini iyileştirebileceğini buldular.

Konu modellemede, konu modellerinin nasıl değerlendirileceği konusunda genel bir anlaşma yoktur. Bir konu modelinin tutarlılığını veya yorumlanabilirliğini değerlendirmek söz konusu olduğunda, insan değerlendirmesi genellikle

matematiksel metriklere tercih edilir, (Gaspar ve ark., 2015) ilgili konunun genel anlayışını artıracak insan temelli bir bileşenin dahil edilmesini tavsiye eder. (Chang ve ark., 2009) insanların konu modellerini değerlendirebileceği iki görev önermektedir: kelime parazitizmi ve konu parazitizmi.

Bu tezde, LDA'ya tweet'lerde saklı olan konuları keşfetmek için insan, yorumlarını öneriyoruz. Tweetler için LDA algoritması uygulayarak ve LDA sonuçlarını manuel olarak inceledikten sonra, her iki ataktaki reaksiyonların neredeyse aynı konularda tartıştığını bulduk. Bu konular 4 ana başlık altında toplanabilir. Bu dört konuya: haber konusu, dayanışma konusu, öfke konusu ve dini konu diyoruz.

Haber konusu medya tarafından yayınlanan Tweet'leri veya etkinliğin açıklamalarını içeriyor. Dayanışma konusu, mağdurlara sempati, kınama ve dayanışma duyguları içeren ya da mağdurlara yardım ve yardım önerisi içeren tweetleri içermektedir. öfke konusu, bu konunun tweet'lerindeki, teröristler, dini gruplar, etnik, hükümeti veya politikacıları suçlamak gibi çeşitli şeylere karşı öfke, eleştiri, küfür ve kınama duygularını içerir. Dini konu bir dini suçlamaya veya bazı etnik veya göçmenlerden bahsetmeye çok odaklanan Tweet'leri içerir. Tablo 4.1. konuların en bilgilendirici kelimelerini gösterir.

Bu 4 konu LaRambla'da yaklaşık %94 ve LondonBiridge davasında %97'dir. Tartışmanın ana konusu (terörizm) ile ilgili olmayan, çoğu reklamı temsil eden veya terörist operasyonlara tepki konusu ile ilgili olmayan diğer konuları yayınlarken trend hashtag'lerini kullanan geri kalan tweetler. Bu tweetler veri setinden çıkarıldı.

Tablo 4.1. Konuların en bilgilendirici kelimeleri.

#	Haber	Dayanışma	öfke	Dini
1-	Breaking	Condemns	Terrorism	ISIS
2-	Attack	Pray	White	Terrorism
3-	London	Horrible	Horrible	Muslim
4-	Injured	Victims	Terror	Moroccans
5-	Barcelona	Hope	Radical	Mosque
6-	Victims	Peace	Terrorist	Immigrant
7-	Police	Safety	Islam	Islam

Tablo 4.2. (Devamı).

#	Haber	Dayanışma	öfke	Dini
8-	Death	Solidarity	Kill	Radical
9-	Effected	Help	ISIS	Responsible
10-	People	Heart	Hate	Peace
11	Terrorist,	Sad	Assault	Assault
12	Related	Stand	Ideology	immediately
13-	Driver	Love	Massacre	Islamic
14-	Van	Terrorist	Vigil	Prophet
15-	Arrest	strong	Us	Community

Konularıyla ilgili bir tweet örneği Tablo 4.2.'de görülebilir

Tablo 4.2. Konuların tweet örneği.

Konu	LaRambla	LondonBiridge
Öfke	A crazy kid kills someone in Charlottesville the entire right is condemned. Muslims kill 13+ in Barcelona the left cries tolerance.	Pigs are at it again London attacks London Bridge and now two other locations. When enough is enough.
Haber	BBCBreaking: Neither of two men arrested after #Barcelona attack drove the van, police say, but are linked to the incident.	Breaking: Police confirm a second incident ongoing at Borough Market in London london attacks.
Dayanışma	We strongly condemn the cowardly #BarcelonaAttack. Our thoughts are with the bereaved.	My heart love and prayers go out to all the victim’s survivors heroes and everyone else effected. #londonbridge.
Dini	muslims praying muslims slaughtering muslims blaming muslims spinning muslims hating muslims fighting.	Muslims were opening there fast whilst cowards were committing there attacks. London Attacks London Bridge

4.2. Sözlük Tabanlı Yaklaşım

Sözlük tabanlı yaklaşım denetimsiz bir yaklaşımdır. Sözlük tabanlı yaklaşım, belirli bir belgenin duygu yönünün kelimelerinin ve cümlelerinin duygu yönünden çıkarılabileceğini varsaymaktadır.

Yaklaşımda, her bir dokümanın duygu yönelimlerini belirlemek için farklı süreçlerin kullanılacağını içerir. Bu süreç, belge kelimelerini konuşma bölümlerine ayırmayı, her belge kelimesi için puan atamak için bir sözlük kullanma, olumsuzlama işleme, bazı etki alanı kelimelerinin duyarlılığını belirlerken etkisini önlemek için kullanmayı önerdiğimiz alan kelimelerini nötrleştirme içerir. tweet ve bu yaklaşımdaki son adım skor ağırlığının hesaplanmasıdır (Şekil 4.2.).

Şekil 4.2. Önerilen çerçevedeki sözlük yaklaşımı adımları.

4.2.1. POS etiketleme

Konuşma parçası (POS) etiketlemesi, NLP'nin herhangi bir dilde temel uygulamalarından biridir. Bir cümledeki her kelimeye bir etiket atama işlemidir. Sadece etiketini tanımlamamız gereken kelimeye değil, aynı zamanda komşu kelimelere de bağlıdır, çünkü aynı kelime içeriğe bağlı olarak oynamak için farklı rollere sahip olabilir. Parçalama, bağımlılık ayrıştırma, herhangi bir dilde varlık tanıma gibi görevleri yerine getirmek için bir ön görev olarak hizmet eder.

Bu çalışmada tweet'lerin tüm kelimelerini etiketlemek için python Natural Language Toolkit (NLTK) tokenize paketini uyguladık. NLTK en güçlü NLP kütüphanelerinden biridir. Bir kelimenin bir cümle içinde yer aldığı dil kategorisini otomatik olarak tanımlar.

Aşağıdaki örnek, NLTK'nın bir metni nasıl aldığını ve Token nasıl dönüştürdüğünü göstermektedir. Sonra pos_tag () yöntemini kullanarak bu belirteçler için konuşma etiketleme bölümlerini yapın. Pos_tag () belirli kelimeler için belirli etiketler çıkarır. (Algo 4.1.).

```
import nltk
from nltk.tokenize import word_tokenize
text = word_tokenize("Hello welcome to the world of to learn Categorizing
and POS Tagging with NLTK and Python")
nltk.pos_tag(text)

OUTPUT
[('Hello', 'NNP'), ('welcome', 'NN'), ('to', 'TO'), ('the', 'DT'),
('world', 'NN'), ('of', 'IN'), ('to', 'TO'), ('learn', 'VB'),
('Categorizing', 'NNP'), ('and', 'CC'), ('POS', 'NNP'), ('Tagging', 'NNP'),
('with', 'IN'), ('NLTK', 'NNP'), ('and', 'CC'), ('Python', 'NNP')]
```

4.2.2. Puanlama

İkinci bölümde açıkladığımız gibi, sözcük yaklaşımında sözlük tabanlı yaklaşım ve corpus tabanlı yaklaşım iki yöntem vardır. Bu tezde, örnekteki terimleri belirli bir sözlükle karşılaştırarak metin verilerine duygu puanları atamak için terimler ve duyarlılık puanları içeren önceden oluşturulmuş sözlükler olan sözlük tabanlı yöntemler kullanıyoruz. Bu tezde SentiWords sözlüğü kullanıldı, çünkü -1 ile 1 arasında bir duygu puanıyla ilişkili yaklaşık 155 bin İngilizce kelime içeren yüksek kapsama alan bir kaynaktır. Bu kaynaktaki kelimeler lemma # PoS biçimindedir ve sıfatlar, isimler, fiiller ve zarflar içeren WordNet listelerine hizalanır.

Bu nedenle, cümlelerde sıfatlar, isimler, fiiller veya zarflardan birine sahip olan tüm kelimelere puanları SentiWords'ten atanır. Tweetlerden örnek: “We stand with Barcelona tonight our hearts go out to the loved ones of the victims of this senseless act of terror”, POS'u kelimeleri etiketledikten sonra SentiWords sözlüğünden kelimelerin polarite puanlarını belirledik. İşlem, listedeki her kelimeyi etiketiyle, sözlükteki sözcük listesiyle karşılaştırarak yapılır. Varsa, bu sözcüğün polarite puanı değilse NA yerleştirilir (Tablo 4.3.).

Tablo 4.3. Tweet içinde etiketleme ve puan atama örneği.

#	token	lemma	upos	xpos	type	puan
1	we	we	PRON	PRP		NA
2	stand	stand	VERB	VBP	V	0.06753
3	with	with	ADP	IN		NA
4	Barcelona	Barcelona	NOUN	NN	N	0
5	tonight	tonight	NOUN	NN	N	0

Tablo 4.3. (Devamı).

#	token	lemma	upos	xpos	type	puan
6	our	we	PRON	PRP$		NA
7	hearts	heart	NOUN	NNS	N	0.48758
8	go	go	VERB	VBP	V	0.33002
9	out	out	ADV	RB		0
10	to	to	ADP	IN		NA
11	the	the	DET	DT		NA
12	loved	loved	ADJ	JJ	A	0.66256
13	ones	one	NOUN	NNS	N	0.2725
14	of	of	ADP	IN		NA
15	the	the	DET	DT		NA
16	victims	victim	NOUN	NNS	N	-0.73756
17	of	of	ADP	IN		NA
18	this	this	DET	DT		NA
19	senseless	senseless	ADJ	JJ	A	-0.42
20	act	act	NOUN	NN	N	0.16008
21	of	of	ADP	IN		NA
22	terror	terror	NOUN	NN	N	-0.56255

4.2.3. Nötrleştirme alanı kelimeleri

Her bir konudaki sık kelimeleri gözden geçirip puanlarını belirleyerek, bu konuların kelimelerinin negatiften olumluya derece olarak değiştiğini bulduk. Konu dayanışması kelimelerinin dereceleri daha olumlu dereceler taşıdığından, öfke ve din konularının sözlerine gelince olumsuz ve olumlu kelimeler içeren konu haberlerinin sözleri gelirken, sözleri daha olumsuz duygular taşır, ancak olumsuz öfke konusundaki kelimelerin derecesi dinden daha olumsuzdur.

Ayrıca konu arasında ortak kelimeler olduğunu ve alanla ilgili olduğunu tespit ettik. *Bomb, Attack, Kill, Injured, Explosion* vb. kelimeler farklı konularla tweet'lerde görünebilir. SentiWords'e göre, bu kelimelerin yüksek negatif puanları var ve tweetin son kutupluluğunu etkileyebilir, bu da tweetin duygularının negatif olduğunu gösterir. Bu etkiden kaçınmak için alanımızla ilgili bu yaygın kelimeleri seçtik ve onlara sıfır puanı verdik, yani onları etkisiz hale getirdik. Bu yaygın kelimeler Tablo 4.4.'te gösterilmiştir.

Örneğin, Tablo 4.5.'te haberler ve dayanışma konularından tweetlerimiz var. Tablo 4.6. ve Tablo 4.7. bu tweet'lerde nötrasyonu nasıl uyguladığımızı göstermektedir. Bu tweet'lerin alan kelimelerini nötralize etmeden önce, puanların kelime toplamını hesapladığımızda, bu tweet'lerin son polaritesini belirlemede bazı kelimelerin açık bir etkisi olduğunu görüyoruz. (dead, injured, arrested, victims, terror ve attack) gibi alan kelimelerini nötralize ettikten sonra, her bir tweetin toplam sonucu, her bir tweet'e göre olumsuzluktan tarafsızlığa veya pozitifliğe daha yakın hale geldi.

Tablo 4.4. Ortak etki alanı kelime ve puanları.

Kilme	PoS	Puan	Kilme	PoS	Puan
Arrest	VB	-0.67	Gun	VB	-0.34
Attack	VB	-0.75	Gunman	NN	-0.5
Attack	NN	-0.75	Incident	NN	-0.2
Attacker	NN	-0.69	Incident	JJ	-0.2
Attacking	JJ	0.16	Injured	JJ	-0.39
Blood	NN	-0.38	Kill	NN	-0.8
Bloody	RB	-0.54	Killer	NN	-0.82
Bloody	JJ	-0.54	Killing	JJ	-0.76
Bloody	VB	-0.54	Killing	NN	-0.76
Bomb	NN	-0.63	Murderer	NN	-0.77
Bomb	VB	-0.63	Shoot	VB	-0.38
Bombing	NN	-0.73	Shoot	NN	-0.38
Casualty	NN	-0.42	Shooter	NN	-0.33
Dead	JJ	-0.75	Suicidal	JJ	-0.71
Dead	RB	-0.75	Suicide	NN	-0.86
Dead	NN	-0.75	Terror	NN	-0.56
Death	NN	-0.78	Terrorism	NN	-0.85
Die	NN	-0.83	Terrorist	NN	-0.66
Die	VB	-0.83	Terrorize	VB	-0.63
Died	VB	-0.83	Victim	NN	-0.74
Explosion	NN	-0.46	Weapon	NN	-0.26
Explosive	NN	-0.22	Weaponry	NN	-0.32
Explosive	JJ	-0.22	Wound	VB	-0.44
Gun	NN	-0.34	Wounded	JJ	-0.42

Tablo 4.5. Tweet örneği.

Metin	Konu
At least 13 people killed and more than 80 injured in Barcelona van attack, Catalonia official says.	Haber
Breaking, Neither of two men arrested after Barcelona attack drove the van, police say, but are linked to the incident.	Haber
My heart goes out to those injured and the families of those who lost their lives today in Barcelona.	Dayanışma
Praying for the victims and families of those attacked in Barcelona.	Dayanışma

Tablo 4.6. Haber konuları, nötrleştirme tweetleri.

Tweet 1			Tweet 2		
Kilme	Önceki puan	Sonraki puan	Kilme	Önceki puan	Sonraki puan
At	0	0	Breaking	0	0
least	0	0	Neither	0	0
13	0	0	of	0	0
people	0.18	0.18	two	0	0
dead	**-0.75**	**0**	men	0	0
and	0	0	**arrested**	**-0.67**	**0**
more	0	0	after	0	0
than	0	0	Barcelona	0	0
80	0	0	**attack**	**-0.7**	**0**
injured	**-0.39**	**0**	drove	0.13	0.13
in	0	0	the	0	0
Barcelona	0	0	van,	0	0
van	0	0	police	-0.10	-0.10
attack	**-0.7**	**0**	say	0.23	0.23
Catalonia	0	0	but	0	0
official	0.19	0.19	are	0	0
says	0.23	0.23	linked	0	0
			to	0	0
			the	0	0
			incident	-0.2	-0.2
Toplam	**-1.24**	**0.6**		**-1.31**	**0.06**

Tablo 4.7. Dayanışma konuları, nötrleştirme tweetleri.

Tweet3			Tweet4		
Kilme	Önceki puan	Sonraki puan	Kilme	Önceki puan	Sonraki puan
my	0	0	Praying	0.18	0.18
condolences	-0.09	-0.09	for	0	0
and	0	0	the	0	0
prayers	0.18	0.18	**victims**	**-0.74**	**0**
for	0	0	and	0	0
you	0	0	families	0.56	0.56
and	0	0	of	0	0
for	0	0	those	0	0
all	0.22	0.22	**attacked**	**-0.7**	**0**
the	0	0	in	0	0
victims	**-0.74**	**0**	Barcelona.	0	0
in	0	0			
Barcelona	0	0			
Attack	-0.7	0			
peace!	0.69	0.69			
Toplam	**-0.44**	**1**		**-0.7**	**0.74**

4.2.4. Olumsuz muamele

Dilbilimdeki olumsuzlama, bir olumlamayı tersine ya da tam tersine yapma sürecidir. verilen bir duyguyu analiz ederken önemli bir konudur; bu, cümlelerin (not, don't, should vb.) bir olumsuzlama kelimesi içermesine bağlanabilir. Bu kelimeler cümlenin kutupluluğunu değiştirir. Örneğin, (I don't like the movie). ("Like") kelimesi olumlu bir anlam taşır, ancak ("don't") cümle anlamını tersine çevirir. Olumsuz muamele, olumsuzlamanın kapsamını belirlemenin ve bir olumsuzlamadan gerçekten etkilenen görüşlü kelimelerin kutuplarını tersine çevirmenin otomatik bir yoludur (Jia,ve ark. 2009).

Cümlenin olumsuzlamanın etkilediği kısmına olumsuzlamanın çevresi veya kapsamı denir. Bir olumsuzlama teriminin varlığı, cümledeki kutupları taşıyan tüm kelimelerin ters çevrileceği anlamına gelmez.

Bununla birlikte, en son teknolojide dilbilimsel fenomenle ilgilenen birçok yöntem vardır. Statik pencere ve noktalama işaretlerine dayalı yöntem gibi en popüler yöntemler. Bir olumsuzlama ifadesinin pencere boyutu, cümledeki hangi kelime dizisinin olumsuzlama sözcüklerinden etkilendiğini belirler; yani, pencere boyutu bire eşitse, bu negatif parçacıkların etkisinin sadece yanlarındaki kelimeye sunulduğu anlamına gelir (Hogenboom ve ark., 2011).

Olumsuzluk kelimesini takip eden kelimenin polaritesini tersine çevirmek için yaklaşma çağrı penceresi boyutlarını uyguladık. Tez kullanım büyüklüğü = 1 çünkü olumsuzlama kelimesinin etkisinin çoğunlukla ondan önceki kelimenin anlamını etkilediğini düşünüyoruz.

Algoritma yürütüldükten sonra, negatif bir parçacıktan etkilenen kelimelerin puanı -1 ile çarpılarak doğrulamayı tersine çevirir. Algoritma algoritma 2'de gösterilir. Alg. 4.2. Olumsuzluk kapsamını belirleme algoritması.

```
Algorithm Determining the Negation Scope
INPUT: Document
for all word ∈ Document do
  If word ∈ negative_particle then
    For i = 0 until i = windows_size do
      If (next word ∉ negative_particle) and (next word ∉ punctuation_mark) then
   word_score = * −1
Else
Break and continue with next word in the document
end if
   end for
  end if
end for
```

4.2.5. Ağırlık hesaplama

Ağırlık hesaplamasında belgedeki kelimelerin negatif ve pozitif puanlarını hesaplıyoruz. N terimini içeren bir belge D verildiğinde, belgenin toplam Puanı, Denklem (4.2) belgesindeki (t) terimlerinin pozitif (pos) ve negatif (neg) puanlarının toplamı hesaplanarak hesaplanır. Ayrıca, bir Denklem (4.3) konusuna ait terimlerin sıklığını (F) hesaplıyoruz.

$$Score(D_x) = \sum_{n=1}^{N} score(t_n) \quad (4.2)$$

$$F(D_x) = \frac{1}{N}(\sum_{t \in topic} f(t) \quad (4.3)$$

Belge puanı ve sıklığının hesaplandığı düşünüldüğünde, bir sonraki adım belgenin polaritesini belirlemektir.

Duygu analizinde polarite belirleme, bir cümlenin özneye karşı pozitif, negatif veya nötr bir duygu ifade edip etmediğini belirlemekle ilgilidir. Dolayısıyla Duygu analizine polarite tayini de denir.

Bu tezde, kutupluluk konuya göre belirlenir ve bu konuların polaritesi, öfke durumunda akut olumsuzluk ile dini durumda orta olumsuzluk arasında, haber durumunda az negatif veya pozitif, dayanışma durumunda pozitif arasında değişir. Polariteyi belirlemek için bulanık mantık kavramı kullanılmıştır.

4.3. Bulanık Mantık Sistemi

Bulanık Mantık 1965 yılında profesör Lotfi A. Zadeh tarafından başlatılmıştır (Gupta, 2011). Bulanık mantık fikri insanlara daha yakın görünmektedir. EVET ve HAYIR dijital değerleri arasındaki tüm ara olasılıkları içeren insanlarda karar verme şekli. Ancak bulanık mantık değişkenlerinde 0 ile 1 arasında değişen bir doğruluk değeri olabilir. Kontrol teorisinden yapay zekaya kadar birçok alana uygulanmıştır.

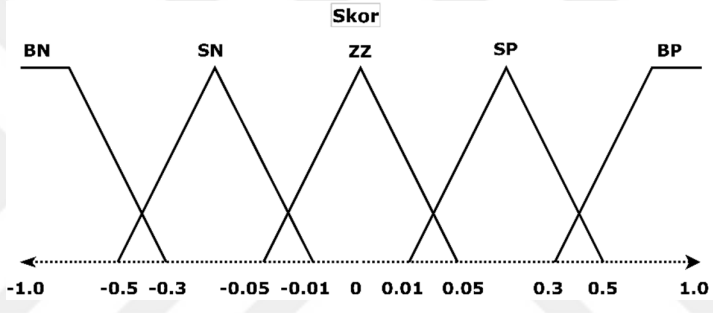
Bu çalışmada, duygu sınıflandırması yapmak için bulanık küme teorisinden yararlanılmıştır. Bulanık küme teorisini seçiyoruz, çünkü duygu kutupluluğundaki içsel bulanıklığı temsil etmek için daha basit bir yol sağlamaktadır. Üçgen ve Trapez üyelik fonksiyonları, ilgili bulanık skor ve polarite terimleri için kullanılır.

İlk adım, net bir giriş değerini, bilgi tabanındaki bilgilerin kullanılmasıyla gerçekleştirilen bulanık bir değere dönüştürme işlemi olan bulanıklaştırmadır. Burada bulanık değerleri elde etmek için kullanılması gereken bulanık kümeler ve

bulanık üyelik işlevleri. skoru ve konuyu girdi değişkenleri olarak alıyoruz, skorun üyelik fonksiyonu Büyük Pozitif (BP), Küçük Pozitif (SP), Sıfır (ZZ), Küçük Negatif (SN) ve Büyük Negatif (BN), konu için ise Haberler, Dayanışma, Din ve Öfke. Bu ilişki kümesini aşağıdaki ilişki ile tanımlarız:

$$A = \sum_{i=1}^{n} \mu_i K(x_i) \tag{4.4}$$

Bu denklem Denklemi (4.4), Söylem Evreninin ve üye işlevinin uygulandığı bulanıklaştırma sürecinde kullanılır. (Şekil 4.3.).



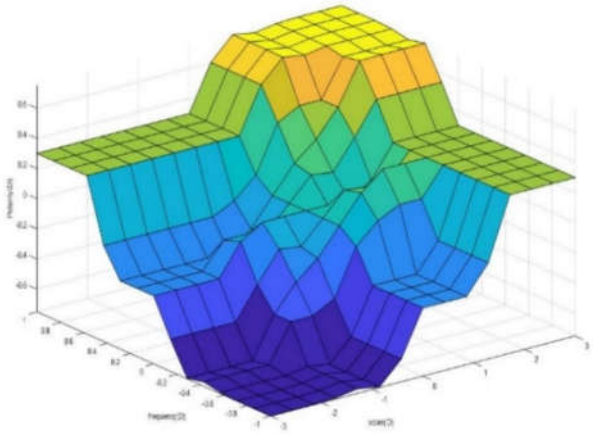
Şekil 4.3. Fuzzification.

Burada araştırmalarımızda Mamdani çıkarsama sistemini kullanıyoruz, araştırma yöntemlerinde kullanılan en genel ve son derece faydalı yaklaşımdır. Bulanık küme teorisi kullanılarak yapılan ilk kontrol sistemidir.

Tablo 4.8.'de gösterilen kuralları kullanarak kontrollü çıktıyı kolayca tanımlayabiliriz böylece daha kolay sonuç alınır. Bunlar basit IF-ELSE kurallarıdır. Bu kuralların kontrol yüzeyi Şekil 4.4. 'de gösterilmiştir.

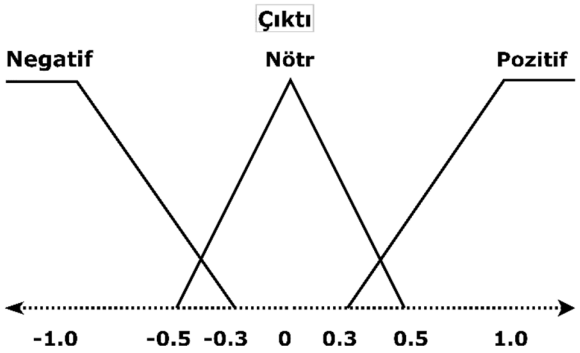
Tablo 4.8. Bulanık kural.

	Dayanışma	Haber	Dini	Öfke
Big Postive	Positive	Positive	Neutral	Neutral
Small Postive	Positive	Neutral	Neutral	Neutral
Zero	Positive	Neutral	Neutral	Negative
Small Negative	Neutral	Neutral	Negative	Negative
Big Negative	Neutral	Negative	Negative	Negative



Şekil 4.4 Kontrol yüzeyi.

Bulanıklaştırma (Defuzzification), bulanık sonuçları net hale dönüştürmek için eşlemenin yapıldığı ters bulanıklaştırma sürecidir. Bir dizi değişkeni bulanık bir sonuca dönüştüren bir dizi kural kullanır (Şekil 4.5.). Bulanık kümeler ve üyelik işlevleri içeren hesaplanabilir sonuçlar üretir. Bulanık kümeleri net değerlere eşleme çıktısı verir. Burada kesin sonuçları tanımlayan bir üçgen almaktadır.

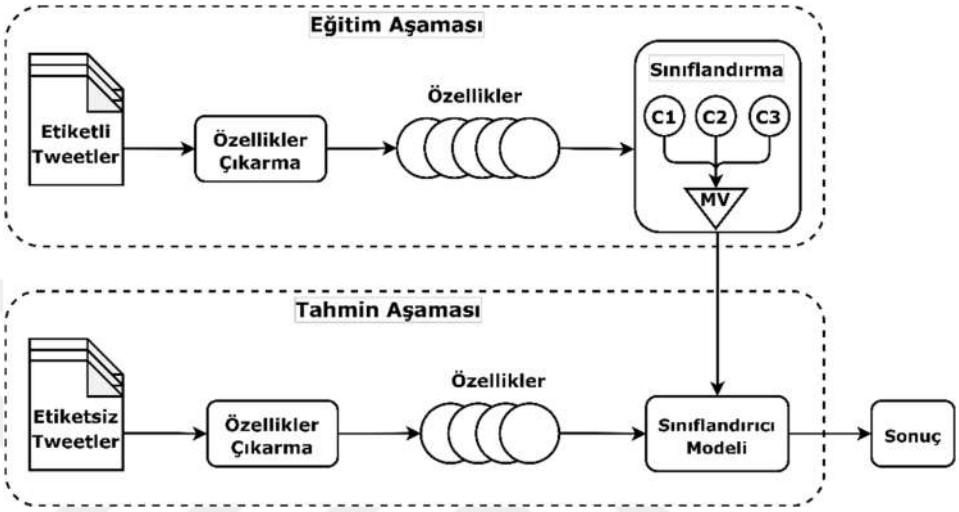


Şekil 4.5. Defuzzification.

4.4. Makine Öğrenimi Yaklaşımı

Makine öğrenmesi-yaklaşımı, duygu analizi için yaygın olarak kullanılır, metinlerdeki duyguları incelemeyi ve değerlendirmeyi amaçlayan birçok algoritma içerir. Makine öğrenimi yöntemleri kullanıldığında, özellikler yakalanır ve analiz edilir. Daha sonra çıkarılan özellikler kullanılarak bazı verilerden (eğitim verileri) bir

model oluşturulur (Şekil 4.6.). Genel olarak, makine öğrenmesinin temel amacı karmaşık yapıları tanımlamak ve otomatik olarak akıllı kararlar vermektir.



Şekil 4.6. Önerilen çerçevedeki Makine öğrenmesi-yaklaşımı adımları.

4.4.1. Özellik çıkarma

Makine öğrenimi yaklaşımında özellik çıkarma, duyguların analiz edilmesi ve sınıflandırılmasının önemli bir parçasıdır. Duyguları analiz etmek için kullanılan birçok özellik çıkarma tekniği vardır. Bu tezde POS Etiketleri ve N-gram özellik çıkarma tekniklerinin performansı üzerine bir çalışma gerçekleştiriyoruz. Hedefimiz, veri kümesinin ilgili özelliklerini yakalayan en iyi özellik türünü bulmaktır.

4.4.1.1. POS özellikleri

POS etiketleme, cümleleri kelimelere bölme ve POS etiketleme kurallarına göre her kelimeye isim, fiil, sıfat ve zarf gibi uygun bir etiket ekleme yöntemidir. Çoğu araştırmada, sıfatlar, fiiller ve zarflar gibi özellikler genellikle duyguları sınıflandırmak için kullanılır. Bazı araştırmacılar sadece bir özellik kullanırken, diğerleri sınıflandırma için farklı kombinasyonlar kullandı.

(Dray ve ark., 2009) sıfatları film incelemelerinden çıkardı ve kesinlik / hatırlamayı hesapladı. Sıfatların bir incelemeyi olumlu ve olumsuz olarak sınıflandırmak için

kullanılabileceği sonucuna vardılar. pozitif sınıf için 0.71 ve negatif sınıf için 0.62 f-skoru kaydetmişlerdir. (Pang ve ark., 2002) yalnızca film inceleme alanlarındaki sıfatları kullanarak film inceleme alan adlarında yaklaşık %82,8 doğrulukla elde etmiştir.

(Neviarouskay ve ark., 2009)'de fiilin, bireysel cümle düzeyindeki görüşleri sınıflandırma yaklaşımlarındaki rolünü tartışmışlardır. Deneme yapmak için açıklamalı 1947 fiilleri çıkardılar. Araştırmacı, cümlenin bir fiil içermesi gerektiğini, çünkü eylemin bağlı olduğu konuşma kısmı olduğunu belirtmektedir. Bu nedenle, fiil kısmı olmayan cümle, görüşleri olumlu veya olumsuz olarak sınıflandıramayabilir. %71 civarında f-skoru elde edilmişler.

(Dragut ve Fellbaum, 2014) ise yoğunlaşan sekiz zarfın, içinde bulundukları cümlelerin duygu puanları üzerindeki etkisini incelemişlerdir. Çalışma, bazı yaygın olarak kullanılan duygu sözlüklerinde temsil edilmelerinin aksine, bu zarfların içsel bir duygu polaritesi taşımadığını, ancak bağlamlarındaki kelimelerle aktarılan duyguyu farklı derecelerde güçlendirdiğini göstermiştir.

Bu çalışmada sıfat, fiil ve zarf özellikleri ayıklamamız gerekiyor. üçünü dikkate alacak ve duyguların sınıflandırılmasında hangisinin en etkili olduğunu test edilecektir.

Bu yöntemin genel işlemleri aşağıdaki gibidir. Twitter cümleleri için öncelikle POS etiketleme yapılır. Bazı cümleler kriterlere göre eğitim veri seti olarak seçilir ve daha sonra kutuplarına göre pozitif ve negatif olmak üzere iki sınıfa ayrılır. Sınıflardaki kelimeler POS etiketleri kullanılarak kümelenir. Daha sonra, sözcüğün ait olduğu kümelenmiş özellik kümesinin bağımlılığına ve sözcük ayrımcılığına bağlı olarak her sözcüğe bir ağırlık değeri atanır. Eğitim aşaması sona erdiğinde, bir istatistiksel veri tablosu elde edilir. Test belgesindeki twitter cümlelerinin duyguları istatistik tablosuna göre değerlendirilir.

Tablo 4.9. Bazı bilgilendirici özellikleri göstermektedir.

	Sıfat			Fiil			Zarf		
	Kilme	Ağ.	Sınıf	Kilme	Ağ.	Sınıf	Kilme	Ağ.	Sınıf
LaRambla	Spanish	48	neu	Praying	203	pos	Strongly	102.6	pos
	Heinous	51	neg	Affected	74	pos	Ago	83.2	neu
	Dead	67	neu	Stand	84.7	pos	Never	75.2	neg
	Devastating	66	neg	Killed	81.4	neu	Still	74.1	neg
	İllegal	60	neg	Confirm	74.2	neu	Exactly	69.7	pos
	Dangerous	54	neg	Wanted	67.7	neu	Really	67.3	pos
	Armed	48	neu	Exploit	66.6	neg	Together	65.8	pos
	Horrible	48	pos	Surround	65.8	pos	ideologically	48.1	neg
	Islamic	46	neg	Confirmed	61.2	neu	Quickly	46.1	neu
	Strong	45	Pos	Alleged	50	neu	Anymore	46.1	neg
LondonBiridge	Muslim	106	neg	Attacks	125	neu	Nearly	430	neu
	Silent	91	pos	Remind	124	neg	Honestly	324	pos
	Safe	67	pos	Died	107	neu	immediately	311	neg
	Critical	44	neu	Devastated	102	pos	Officially	308	neu
	Heartbreaking	43	pos	Fall	100	neu	Absolutely	227	pos
	Wounded	43	neu	Acknowle	94	pos	Again	187	neg
	Conservative	27	neg	Remember	88	pos	More	145	neu
	Tragic	26	neg	Stand	72	pos	Exactly	141	neg
	Sad	20	pod	Killed	71	neg	Never	127	neg
	Jihadi	20	neg	Help	70	pos	Really	92	pos

4.4.1.2. N-gram özellikleri

N-gram, belirli bir metin veya konuşma dizisinden n sürekli kelimeyi veya sesi kontrol etmek için bir stratejidir. Bu model, sıradaki bir sonraki öğeyi tahmin etmeye yardımcı olur. bilgi alma ve metin madenciliğinde yaygın olarak kullanılmaktadır. Duygu analizinde n-gram modeli, dokümanın veya metnin duygusunu tahmin etmeye yardımcı olur. n bitişik jetondan oluşan daha uzun bir özelliğe sahip bir dilim anlamına gelir.

N-gram, kelimelerin doğal bir uzantısıdır. Kelimelerin torbasında, inceleme metnindeki kelimeleri, birgram anlamına gelen tek tek kelime sayısı istatistiklerine ayırıyoruz. n-gram kelimelerin çantasından daha bilgilendirici olabilir çünkü her

kelimenin etrafında daha fazla bağlam yakalarlar. Bununla birlikte, n-gramlar kelime torbasından çok daha büyük ve daha seyrek bir özellik kümesi üretebileceğinden, bunun bir maliyeti vardır.

En yaygın n-gramlar: unigram bir büyüklükte n-gram anlamına gelir, bigram iki büyüklükte n-gram anlamına gelir, tri-gram üç büyüklükte n-gram anlamına gelir, hatta daha yüksek gram, dört-gram, beş-gram vb.

Sonuç olarak, aşağıdaki inceleme metni verildiğinde “Thinking of all those affected by today's attack”. bunu parçalara ayırabiliriz::

- unigram: “‘Thinking’, ‘of’, ‘all’, ‘those’, ‘affected’, ‘by’, ‘today's’, ‘attack’” burada tek bir kelime dikkate alınır.

- bigram: “‘Thinking of’, ‘of all’, ‘all those’, ‘those affected’, ‘affected by’, ‘by today's’, ‘today's attack’” burada bir çift kelime dikkate alınır.

- trigram: “‘Thinking of all’, ‘of all those’, ‘all those affected’, ‘those affected by’, ‘affected by today's’, ‘by today's attack”’ burada üçe eşit olan bir dizi kelime dikkate alınır.

Tablo 4.10., versisetindeki tweetlerdeki en yüksek 10 özellik vektörü (frekans) unigram, bigram ve trigram özellik vektörlerini göstermektedir.

Tablo 4.10. Unigram ve Bigram ve Trigram özellikleri.

	Unigram		Bigram		Trigram	
	Kilme	Sınıf	Kilme	Sınıf	Kilme	Sınıf
LaRambla	People	pos	Barcelona terrot	neg	Barcelona terror attack	neg
	Crazy	neg	Terror attack	neg	Barcelona van attack	neu
	Jihadi	neg	My heart	pos	Prayers go out	pos
	Killed	neu	Jihadi idriss	neu	Barcelona stay safe	pos
	Victims	pos	Effected people	neu	Stay strong barcelona	pos
	Strong	pos	Lying media	neg	Attack jihadi terrorism	neg
	Pray	pos	Stay strong	pos	Possible terrrorist attack	neu
	Radical	neg	Van attack	neu	Denounce radical islam	neg

Tablo 4.10. (Devamı).

	Unigram		Bigram		Trigram	
	Kilme	Sınıf	Kilme	Sınıf	Kilme	Sınıf
	Stay	neg	Sending love	pos	Spanish prime minister	neu
	Van	neu	Isis group	neu	Barcelona so sad	pos
LondonBiridge	Condemning	pos	London attacks	neg	London terror attack	neg
	Skynews	neu	London bridge	neg	Attack london bridge	neg
	Muslim	neg	Stay safe	pos	London attacks pray	pos
	Pray	pos	London mayor	neg	London bridge borough	neu
	Shoot	neu	Attacks pray	pos	London attacks breaking	neu
	Borough	neu	Isis terror	neu	Islam london attacks	neg
	Manchester	neg	Borough market	neu	Attack multiple casualities	neu
	Victims	pos	Breaking attack	neu	Attack stay safe	pos
	Evil	neg	My heart	pos	People Injured numeber	neu
	Wake	neg	Vauxhall london	neu	Radical islamic spreads	neg

4.4.2. Sınıflandırma algoritması

Sınıflandırma düzeyinde, her bir kaydın bir sınıfa etiketlendiği modeli eğitmek için makine öğrenme algoritmalarıyla ortaya çıkan özellikleri kullanıyoruz. Sınıflandırma modeli, sınıf etiketlerinden birine ait temel kayıttaki özelliklerle ilgilidir. Daha sonra model, metnin sınıfını tahmin etmek için test verilerine uygulanır, metnin duyarlılığını sınıflandırmak anlamına gelir.

Sınıflandırma aşağıdaki gibi tanımlanabilir. Belgenin bir açıklaması $d \in X$ verildiğinde, burada X belge alanı, sabit bir $C = \{c_1, c_2, \ldots, c_n\}$ sınıfı seti ve $\langle d, c \rangle$ etiketli eğitim $\langle d, c \rangle \in X \times C$ bir makine öğrenme algoritması Γ kullanarak, belgeleri sınıflarla eşleyen bir sınıflandırıcı veya sınıflandırıcı işlevi $\Gamma(D) = \gamma$ learn öğrenmemiz gerekir: $\gamma: X \to C$. Duygu analizi için, C kümesi üç $C = \{positive, negative, neutral\}$ sınıfından oluştur.

Farklı türlerde makine öğrenme algoritmaları vardır, bu tezde dört sınıflandırma algoritması kullanıyoruz: Naive Bayes (NB) sınıflandırması olasılıklara dayalı, SVM ve istatistiklere dayalı Lojistik Regresyon (LR) sınıflandırması, Karar Ağacı (DT)

sınıflandırması, grafikler ve tüm bu algoritmalar bir oylama algoritmasında birleştirilmiştir.

NB basit ama etkili ve yaygın olarak kullanılan bir makine öğrenimi sınıflandırıcısıdır. Bayesci bir ortamda Maksimum A Posteriori karar kuralını kullanarak sınıflandırma yapan olasılıklı bir sınıflandırıcıdır. Aynı zamanda çok basit bir Bayes ağı kullanılarak da temsil edilebilir. Bir Naive Bayesian modeli oluşturmak çok kolaydır, bu da onu çok büyük veri kümeleri için özellikle yararlı kılan karmaşık yinelemeli parametre tahmini içermez. Basitliğine rağmen, Naive Bayesian sınıflandırıcısı genellikle şaşırtıcı derecede iyi bir performans sergiliyor ve yaygın olarak kullanılıyor çünkü çoğu zaman daha karmaşık sınıflandırma yöntemlerinden daha iyi performans göstermektir.

SVM yaygın olarak kullanılan makine öğrenme algoritmasıdır. Aynı zamanda hem sınıflandırma hem de regresyon amaçları için kullanılabilir. SVM'nin ana fikri, eğitim setindeki herhangi bir sınıf veri noktası arasında maksimum marjlı köprü oluşturmaktır, bu yeni verilerin doğru bir şekilde sınıflandırılma şansını artırabilir.

LR, verileri ayrık sonuçlara ayırmak için olasılıklı bir istatistiksel yöntemdir. 'Lojistik Regresyon' olarak adlandırılır, çünkü altında yatan teknik Doğrusal Regresyon ile tamamen aynıdır. Ancak en büyük fark, ne için kullanıldıklarında yatmaktadır (Feng ve ark., 2014). Değerleri tahmin etmek / tahmin etmek için doğrusal regresyon algoritmaları kullanılır, ancak sınıflandırma görevleri için lojistik regresyon kullanılır.

DT, endüktif çıkarım için en yaygın kullanılan ve pratik yöntemlerden biridir. Gürültülü verilere dayanıklıdır ve ayrık ifadeleri öğrenebilir. Karar ağaçları hem sınıflandırma hem de regresyon problemleri için kullanılır. ID3, C4.5 gibi karar ağacı algoritmaları çok popüler endüktif çıkarım algoritmalarıdır ve birçok eğilimli göreve başarıyla uygulanırlar.

Tüm bu algoritmaların öngörülen sınıfları bir oylama algoritmasında (MV) birleştirildi. N sayıda sınıfımız olduğunu varsayalım. Oylama aşağıdaki formüle göre yapılır (Denklem 4.5).

$$C = \{c_1, c_2, \ldots, c_n\} \tag{4.5}$$

ve m sınıflandırma algoritması Γ sayısı (Denklem 4.6).

$$\Gamma_i,\ i = \{1, \ldots, m\} \tag{4.6}$$

Oylama algoritması MV 'de c_i sınıfının p olasılığı, sınıflandırma algoritmalarında bu sınıfların olasılıklarının ortalamasıdır. (Denklem 4.7).

$$p(MV, c_i) = \frac{1}{m}\sum_{i=1}^{m} p(\Gamma_j, c_i) \tag{4.7}$$

Oylama algoritmasında oylanan "MV_C" sınıfı maksimum olasılık sınıfıdır. (Denklem 4.8).

$$MV_C = \max\{p(MV, c_1), p(MV, c_2), \ldots, p(MV, c_n)\} \tag{4.8}$$

BÖLÜM 5. TARTIŞMA VE SONUÇ

Bu bölümde, önerilen modelimizin deneylerinden elde ettiğimiz sonuçlar hakkında tartışacağız. POS özellikli sınıflandırma algoritmaları kullanıldığında elde edilen sonuçla başlayarak. Bundan sonra, N-gram özelliklerinin uygulanmasından elde edilen sonuçları tartışacağız. Ayrıca bu çalışmanın sonuçlarını sunarız ve gelecekteki çalışmalar için önerilerde bulunuruz.

5.1. Tartışma

5.1.1. POS özelliklerinin uygulanması

Tablolar 5.1. ve 5.2., farklı algoritmalar kullanarak Sıfat, Fiil ve Zarflar gibi üç tip POS özelliği kullanarak LaRambla ve LondonBiridge saldırıları hakkındaki reaksiyonun duygurlı sınıflandırma sonuçlarını göstermektedir. Burada, Sınıflandırma algoritmalarının uygulanmasından elde edilen sonuçlar arasında bir karşılaştırma yapıyoruz.

Sıfat özelliğinin kullanılması durumunda, NB'in LaRambla ve LondonBiridge veri kümelerinde sırasıyla %94.8 ve %92.5 doğrulukla en iyi sonuçları verdiği görülmektedir. MV algoritması, LaRambla'da %92,7 ve LondraBiridge'de ikinci en iyi doğruluğu %93,3 olarak gerçekleştirdi. LR gelince, iki veri setinde %93 doğrulukla aynı sonucu verir. SVM ve DT, LaRambla'da sırasıyla %92,5 ve %92,4 doğruluk ve LondonBiridge ile %90,8 ve %90,7 doğruluk ile hemen hemen aynı sonuçları verir.

Fiil özelliğinin kullanılması durumunda, DT algrithm kullanımı daha iyi doğruluk sağlar, LaRambla ve %93.1 LondonBiridge ile %94.3, MV algoithm, LR ve SVM iki

veri setinde neredeyse aynı sonuçları elde etti. NB %91.5 ve %91.3 ile biraz daha düşük doğruluk elde etti.

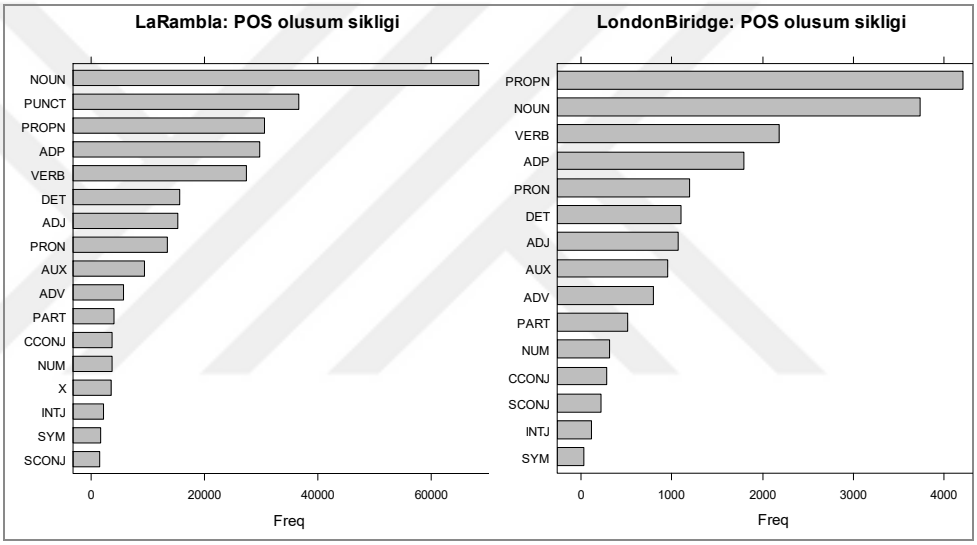
Tablo 5.1. LaRambla veri seti ile POS özelliklerinin sınıflandırma sonuçları.

Sınıflandırıcı	Özelliği	Doğruluk	Duyarlılık	Hassasiyet	F1-Skor
NB	Sıfat	94.8	93.8	98.7	96.19
	Fiil	91.5	90.57	94	92.25
	Zarf	86.8	87.7	88.9	88.3
LR	Sıfat	93.12	92.31	97.90	95.02
	Fiil	92.5	91.8	97.6	94.61
	Zarf	84.62	82.89	83.98	83.43
SVM	Sıfat	92.5	92.9	96.1	94.47
	Fiil	92.4	92.4	96.5	94.41
	Zarf	85.44	85.67	85.53	85.6
DT	Sıfat	92.4	92.4	96.5	94.41
	Fiil	94.3	93.2	98.7	95.87
	Zarf	80.48	78.75	79.84	79.29
MV	Sıfat	93.3	93.2	98.7	95.87
	Fiil	92.7	92.8	97.8	95.23
	Zarf	84.4	86.15	85.13	85.64

Tablo 5.2. LondonBiridge veri seti ile POS özelliklerinin sınıflandırma sonuçları.

Sınıflandırıcı	Özelliği	Doğruluk	Duyarlılık	Hassasiyet	F1-Skor
NB	Sıfat	92.5	92.9	96.1	94.47
	Fiil	91.3	91.3	96.3	93.73
	Zarf	85.06	83.29	85.35	84.31
LR	Sıfat	93	92.2	97.8	94.92
	Fiil	92.14	91.44	97.24	94.25
	Zarf	82.28	80.73	76.77	78.7
SVM	Sıfat	90.8	90.1	96.9	93.38
	Fiil	92.19	92.19	96.29	94.2
	Zarf	84.46	82.91	78.95	80.88
DT	Sıfat	90.7	90.8	95.8	93.23
	Fiil	93.1	92.3	97.9	95.02
	Zarf	80.58	71.29	79.89	75.35
MV	Sıfat	92.3	91.9	97.0	94.38
	Fiil	91.57	91.67	96.67	94.1
	Zarf	83.57	83.32	82.3	82.81

Zarf özelliği kullanılması durumunda, tüm algoritmaların sıfat ve fiil özelliğini kullanmaya kıyasla zarf özelliği ile kötü sonuçlar verdiğini not edebiliriz. NB tarafından LaRambla veri seti ile elde edilen %86.8'den fazla doğruluk yapmazken, en kötü doğruluk, LondonBiridge veri setiyle DT tarafından %80.48 elde edildi. Zarf kelimelerinin sayısı, Şekil (5.1.) 'de gösterildiği gibi fiil ve sıfatların sayılarına kıyasla çok daha az olduğu için, bu daha kötü sonuç gerekçelendirilebilir. Ayrıca, incelememiz yoluyla, birçok araştırmacı ayrıca zarfın tek başına (Dragut ve Fellbaum, 2014)'da belirtildiği gibi doğal duyugu polaritesi taşımadığını belirtmiştir.



Şekil 5.1. Veri setlerindeki POS sayıları arasında karşılaştırma.

Üç POS özelliğinin sonuçları arasında karşılaştırma yaparak, sıfatın kullanılmasının en iyi genel performansı verdiği sonucuna varabilirken, zarfları kullanmanın daha kötü bir performans verdiği sonucuna varabiliriz. Her ne kadar konuşmanın tüm parçaları önemli olsa da, ancak insanlar duyguların çoğunu tasvir etmek için en yaygın olarak sıfatlar kullandı. Sıfat ve cümle özneliğinin varlığı arasında yüksek bir korelasyon vardır. Literatürden, sıfatların konuşmanın tüm bölümleri arasında en çok özellik olarak kullanıldığını ve sadece özellik üretimi için sadece sıfatlara odaklanan tüm çalışmalar tarafından yüksek bir doğruluk bildirildiğini bulduk.

Genel olarak, sıfat özelliğini kullanırken NB ve LR'nin en iyi sonuçları aldığını görüyoruz. DT ise fiil özelliği ile en iyi performansı elde etti. SVM ve NB algoritmaları en iyi zarf ile performans gösterir. Burada MV en iyi şekilde çalıştığını,

farklı özelliklerle oldukça iyi doğruluk sonuçları elde ettiğini not edebiliriz. Bu, MV kullanımının her zaman bir denge ve iyi sonuçlar elde ettiğini doğrularken, diğer algoritmaların sonuçlarının farklı özelliklerle dalgalandığını doğrular.

İki veri setinin sonuçları ile karşılaştırıldığında, LaRambla'nın sonuçları LondonBiridge'den biraz daha iyidir. LaRambla'ya tepki olarak kullanılan POS özelliklerinin sayısının LondonBiridge'de kullanılandan çok daha fazla olduğunu belirtiyoruz (Şekil 5.1.). Ayrıca, LaRambla ve LondonBiridge'de kullanılan ortalama kelimeleri sayarken, LaRambla durumunda 5-10 kelime oranında tweet sayısı Londra'daki tweet sayısının neredeyse iki katıdır (Şekil 3.4.) iyi sonuçlar, diğer algoritmaların sonuçları farklı özelliklerle dalgalanır.

5.1.2. N-gram özelliklerinin uygulanması

Bu bölümde unigram, bigram ve trigram gibi N-gram özelliklerinin sınıflandırma performansını karşılaştırıyoruz. N-gram özellikli sınıflandırma algoritmaları kullanıldığında elde edilen deneysel sonuçlardan. Tablolar (5.3.) ve (5.4.) 'de gösterilen sonuçlar, unigram kullanımının en iyi performansı verirken, trigramları kullanırken daha kötü bir performans sağlar.

Trigramın başarısızlığı muhtemelen trigram özelliklerinin unigram ve bigramdan daha düşük frekans sayımlarına sahip olmasından kaynaklanmaktadır. Yeterli trigrm frekansı eksikliği nedeniyle, reaksiyon veri setlerinin toplandığı tweetlerdeki metinlerin sınırlandırılması (140 karakteri geçmeyen). önişlemdeki metinden gürültülü verileri hariç tutarak, kalan kelimeler iyi sonuçların elde edilmesine yardımcı olabilecek trigram özelliklerinin frekans sayımının üretilmesinde yeterli olmayabilir.

Unigram özelliğinin kullanılması durumunda, SVM'nin LaRamla veri seti ile %88,36 ile en iyi sonuçları verdiği görülmektedir, Ancak, LondonBiridge veri setini kullanırken en kötü sonucu almıştır. aksine, LR ve DT'nin LondonBiridge'de sırasıyla %84.5 ve %84.4 ile en yüksek sonuçları aldığını, LaRambla veri setleriyle

en kötü perforamansı gerçekleştirdiklerini görüyoruz. Bu tutarsızlığı NB sonuçlarında da buluyoruz. MV ise iki veri kümesinde üçüncü sırada yer aldmıştı.

Tablo 5.3. LaRambla veri seti ile Ngram özelliklerinin sınıflandırma sonuçları.

Sınıflandırıcı	Özelliği	Doğruluk	Duyarlılık	Hassasiyet	F1-Skor
NB	Unigram	86.93	88.23	84.62	86.39
	Bigram	83.95	86.14	80.99	83.49
	Trigram	75.58	74.16	78.68	76.35
LR	Unigram	86.22	84.4	86.73	85.55
	Bigram	84.93	88.94	84.85	86.85
	Trigram	69.9	71.15	66.45	68.72
SVM	Unigram	88.36	89.06	84.76	86.86
	Bigram	87.34	88.94	84.85	86.85
	Trigram	74.88	74.16	78.68	76.35
DT	Unigram	85.34	84.2	84.54	84.37
	Bigram	84.06	77.54	84.29	80.77
	Trigram	72.53	71.5	69.44	70.45
MV	Unigram	86.26	84.38	83.88	84.13
	Bigram	84.49	80.83	83.98	82.37
	Trigram	73.8	72.45	76.18	74.27

Tablo 5.4. LondonBiridge veri seti ile Ngram özelliklerinin sınıflandırma sonuçları.

Sınıflandırıcı	Özelliği	Doğruluk	Duyarlılık	Hassasiyet	F1-Skor
NB	Unigram	83.33	84.16	81.12	82.61
	Bigram	80.83	83.98	82.38	83.17
	Trigram	71.72	75.32	70.36	72.76
LR	Unigram	84.5	86.15	85.13	85.64
	Bigram	83.95	85.77	81.21	83.43
	Trigram	60.76	64.42	61.58	62.97
SVM	Unigram	83.07	85.77	79.79	82.67
	Bigram	82.89	83.98	83.43	83.7
	Trigram	73.19	73.59	76.33	74.93
DT	Unigram	84.4	86.73	85.55	86.14
	Bigram	80.57	84.05	82.27	83.15
	Trigram	71.29	79.89	75.35	77.55
MV	Unigram	83.67	80.57	84.05	82.27
	Bigram	81.77	78.92	81.77	80.32
	Trigram	72.43	89.06	84.76	86.86

Bigram özelliğinin kullanılması durumunda, LaRamla veri seti ile SVM sonuçları ve LondonBiridge veri seti ile LR arasında anlamlı bir fark yoktu, unigramdaki gibi en yüksek doğruluğa ulaşırlar, ayrıca MV algoritması unigramdakiyle aynı yere geldi, Ancak, unigram performansına kıyasla diğer algoritmaların geri kalan sonuçlarında net bir fark. DT'nin LaRamla veri seti durumunda NB'den biraz daha iyi performans gösterdiğini ve LondraBiridge veri seti durumunda en kötü performansı elde ettiğini gördük. aksine, NB ağının LondraBiridge veri seti ile DT'den biraz daha iyi perfotmance ve LaRamla veri seti ile en kötü performans buluyoruz.

Tigram özelliğinin kullanılması durumunda SVM, %73.19 ile LaRamla veri seti ve LondonBiridge veri seti ile ikinci yükseklik performansı ile en iyi sonuçları verdi. Ancak LR, iki veri kümesinde en kötü performansa ulaştı. MV algoritması ise iki veri seti ile üçüncü sırada yer aldı. Burada diğer algoritmaların sonuçlarında tutarsızlık olduğunu not edebiliriz. DT, LaRamla veri setiyle ikinci en yüksek doğruluğu ve LondonBiridge veri setiyle dördüncü sırada yer alırken. NB, LondonBiridge veri setiyle ilk yüksek doğruluğu elde etti ve LaRamla veri setiyle dördüncü sırada yer aldı.

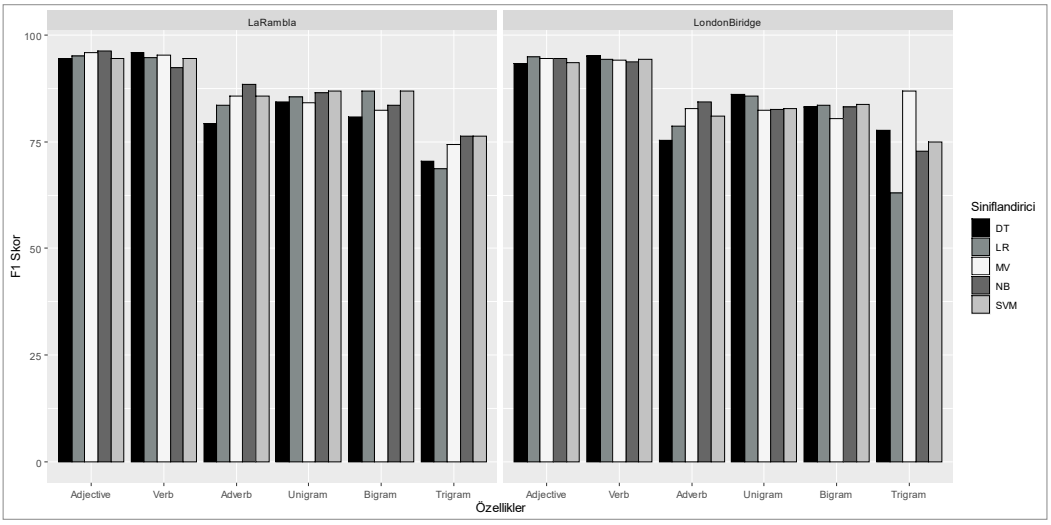
Önceki sonuçlardan, MV algoritmasının sonuçlarında net bir istikrar bulurken, diğer algoritmaların sonuçlarında özellik ve veri setisine göre dalgalanma görüyoruz. Bununla birlikte, hepsi, unigram özelliklerinin bigram ve trigramlardan çok daha yüksek bir frekans sayısına sahip olduğundan daha önce bahsettiğimiz için, bigram ve trigram üzerinde unigramların üstünlüğüne katılırlar. Ayrıca kullanılan algoritmaların ve çalışma yöntemlerinin de ungramın bigram ve trigram üzerindeki üstün sonuçlarında rol oynadığını görüyoruz.

NB, özelliklerin birbirinden bağımsız olduğu olasılıklı bir yöntemdir. Bu nedenle, analiz unigram ve bigram kullanılarak gerçekleştirildiğinde, elde edilen doğruluk değeri, trigram kullanılarak elde edilenlerden nispeten daha iyidir. Ayrıca, LR algoritması durumunda, koşullu dağıtım ve kelime sınıfı çiftine dayanan LR algoritması incelemeyi sınıflandırmaya yardımcı olduğundan, analiz için tek

kelimeyi dikkate alan unigram özelliği, diğer özelliklerle karşılaştırıldığında en iyi sonucu sağlar. SVM yöntemi olasılık dışı bir doğrusal sınıflandırıcı olduğundan ve veri setini ayırmak için hiper düzlemi bulmak için model çalıştırdığından, analiz için tek kelimeleri analiz eden unigram modeli daha iyi sonuç verir. DT sınıflandırıcısı, verileri bölmek için öznitelik değeri üzerindeki bir koşulun kullanıldığı veri alanının hiyerarşik bir ayrışmasını sağlar. Bu nedenle, analiz, unigram veya bigram özelliği kullanılarak gerçekleştirildiğinde, elde edilen doğruluk değeri, trigram kullanılarak elde edilenlerden nispeten daha iyidir.

İki veri setinin (LaRambla ve LondonBiridge) sonuçları arasında karşılaştırma yaparken, LaRambla'nın sonuçlarının LondraBiridge'den daha iyi olduğunu görebiliriz. Çünkü LaRambla veri kümesindeki kelime sayısı Londra'dakinden çok daha fazladır.

Duyarlılık, Hassasiyet ve F1-Skor de farklı sınıflandırıcıların çeşitli özellikleri için hesaplanır. Elde edilen değerlerden, sıfat ve fiiller özelliklerinin diğer tüm POS ve Ngram özelliklerinden iyi performans gösterdiği bağlanabilir (Şekit 5.2.).



Şekil 5.2. Özelliklerin F1 skor sonuçları arasında karşılaştırma.

5.2. Sonuç

Duygu analizi son yıllarda aktif bir araştırma alanıdır ve farklı alanlarda birçok yararlı uygulama gördü. Duygu analizi, farklı alanlardaki duygu ifadelerinin oldukça farklı olabileceği alan adına bağlıdır. Aynı alanda bile, sınıflandırma doğruluğunu geliştirmek için hakkında derin bilgiye ihtiyacımız olabilir.

Terörist saldırılara yönelik reaksiyonların duygu analizi gibi bir etki alanı olarak, buradaki reaksiyonların güvenlik hissi kaybetme, güçsüz hissetme, öfkeli ve korkulu hissetme ve bazı etnik veya dini gruplara karşı hoşgörüsüzlük arasında değiştiğini görebiliriz.

Bu alanda özne genellikle olumsuz bir duyguya eğilimlidir, bu nedenle bu alanların duygularını sadece geleneksel duygu analizi yöntemlerini kullanarak tanımlamak yetersiz ve yanıltıcı olabilir ve bu yöntemlerin etkinliğini gerçek duyguları çıkarımdan saptamak zordur. Bu nedenle, metin içinde altta yatan konuları araştırmak ve bu reaksiyonların gerçek duygularını çıkarmada onlardan faydalanmak gerekir.

Konu modelleri, büyük miktarlarda etiketlenmemiş metni analiz etmek için basit bir yol sağlar. Bir konu, sık sık birlikte ortaya çıkan bir kelime kümesinden oluşur. aynı belgede bir arada olma eğiliminde olan konu kelimeleri, bu konunun ne hakkında olduğunu kolayca yorumlayabilir, çünkü bu birlikte ortaya çıkan kelimeler birbirleriyle sıkı bir anlambilimsel ilişkiye sahip olma eğilimindedir. Belge koleksiyonlarından çıkarılan bu tür konu kümeleri, bu belgelerin içeriğini belirleme, sorgu verilen en alakalı belgeleri alma, zaman içinde içerikteki değişiklikleri izleme ve belgeler arasındaki benzerlikleri ölçme gibi birçok görevde faydalıdır.

Bu çalışmada, iki farklı ülkede en benzer iki terör saldırısı için sosyal medyadaki tepkilerin veri setini kullandık, Öncelikle, iki terörist saldırının tepkileri arasında olmuş, daha sonra bu terörist saldırılar gerçekleştikten sonra Twitter'da hangi

konuların tartışıldığını ortaya çıkarmak için LDA konu modelleme algoritmasını uyguladık.

Terörist operasyonlara verilen tepkilerden kaynaklanan bu konular arasında dayanışma konusu da var. Bu konunun tweetleri arasında dayanışma, sempati, kurbanlar için dua etme, onlara yardım ve yardım sunma, hatta saldırıyı kınama duyguları yer alıyor. Bu konudaki tepkilerin çoğu olumlu duygular taşıyor.

Haber konusu, bu konu haberleri bildirmek ve olaylar hakkında bilgi vermekle ilgilidir. Bu konu, medya tarafından Twitter'da ve hatta insanlar tarafından yayınlanan saldırı hakkında tweet'ler içeriyor. Bu olayla ilgili bilgilerin yanı sıra, yetkililere, politikacılara ve hatta sıradan insanlara cevap veren tweetler de içeriyor. Bu tepkiler kişisel duygular taşır.

Öfke konusu, Bu konu terörist saldırıların vatandaşlara verdiği olumsuz duyguların bir yönünü ortaya koymaktadır. terörist, dini gruplar, hükümeti ya da politikacıları suçlamak ya da öldürme ve intikam hakkında konuşmak gibi öfkelerini ya da küfürlerini ifade eden insanların tepkilerini içerir. bu konudaki tepkiler daha olumsuz duygular taşır.

Yirmi birinci yüzyılda terörist grupları ayırt eden güdüler, dini olarak motive olmuş gruplardır. Terörist saldırılara karşı birçok tepki dinler hakkında yorumlar içermektedir. Örneğin, tepkilerini analiz ettiğimiz iki terörist saldırıyı benimseyen IŞİD. Bu saldırılar İslam diniyle bağlantılıdır çünkü IŞİD İslam adına kullanan bir örgüttür ve ayrıca çoğu İslam ülkesine ait olan mülteci ve göçmenlere de atıfta bulunmaktadır. Batı ülkelerinde her zaman göçmenlere ve Müslümanlara karşı iddianame bulunan birçok ırkçı ve aşırılık yanlısı grubun aktif olduğu ve bu terör olaylarını Müslümanlara ve göçmenlere karşı propagandalarını onlara karşı nefret duygularını körüklemek için kullandıkları belirtilmelidir. Dini konu da birçok olumsuz duygu içerir.

Bu dört konu kategorisini bulanık mantık sisteminde sözcüksel yaklaşımda kelimelere atadığımız kutuplarla uyguladık. Bu bulanık sistem, metin için bulanık hissi elde etmek ve kesin duyguların değerlerine daha yakın bir karar belirlemek için konu ile kutupluluk arasındaki temel bulanıklığı ayırmanın doğrudan yolunu sağlamıştır.

NB, LR, SVM, DT ve MV algoritması gibi farklı makine öğrenimi sınıflandırıcıları kullanılarak çeşitli Özellik tipi POS ve N-gramları eğitilir. Amacımız, veri kümesinin ilgili özelliklerini yakalayan en iyi özellik türünü bulmak ve popüler bir denetimli sınıflandırıcının performansını konu bağımlılığı bağlamında incelemektir. Sonuçlar, POS özelliklerinin kullanımının N-gram özelliklerine kıyasla daha iyi sonuçlar verdiğini gösterdi.

POS özelliklerinde, sıfat özelliğinin akranlarıyla karşılaştırıldığında yüksek doğruluk elde ettiğini bulduk. Her ne kadar konuşmanın tüm bölümleri önemli olsa da, insanlar duyguların çoğunu tasvir etmek için en yaygın olarak sıfatlar kullanmaktadır. Sıfat ve cümle özneliğinin varlığı arasında yüksek bir korelasyon vardır. Literatürden, sıfatların konuşmanın tüm bölümleri arasında en çok özellik olarak kullanıldığını ve sadece özellik üretimi için sadece sıfatlara odaklanan tüm çalışmalar tarafından yüksek bir doğruluk bildirildiğini bulduk.

Ngramda, unigramın en iyi sonuçları, ardından bigramı gerçekleştirdiğini ve trigramın en kötü doğruluklara ulaştığını bulduk. Trigramın başarısızlığı muhtemelen trigram özelliklerinin unigram ve bigramdan daha düşük frekans sayımlarına sahip olmasından kaynaklanmaktadır. Yeterli trigrm sıklığının olmaması nedeniyle, tweetlerdeki metinlerin sınırlandırılması (140 karakteri geçmeyen), reaksiyon toplandı. önişlemdeki metinden gürültülü verileri hariç tutarak, kalan kelimeler iyi sonuçların elde edilmesine yardımcı olan trigramların sıklık sayısını üretmede yeterli olmayabilir.

Gelecekteki çalışmalarda, mevcut araştırma çalışmaları çeşitli şekillerde genişletilebilir. Örneğin, deneylerimizde bazı özelliklerin POS'ta sıfat ve N-gram

özelliğinde unigram gibi daha iyi duyarlılık doğruluğu elde ettiği fark edildi, bu nedenle duygu sınıflandırmasında modellerin performansını artırmak için bu özelliklerden yararlanılması ve bazı farklı özellikleri birleştirmeyi öneririz.

Bu tezde sunulan modeller, konu sayısının bilindiği ve sabit olduğu varsayılan LDA'ya dayanılarak geliştirilmiştir. Gelecekte başka bir yön, hiyerarşik bir Dirichlet süreç çerçevesi kullanmak için genişletmek mümkün olacaktır, bu da konu sayısının verilerden otomatik olarak çıkarılmasını sağlar. Ayrıca, Latent Semantic Analysis (LSA) ve robabilistic Latent Semantic Analysis (PLSA) gibi başka bir konu modelleme algoritması kullanmayı öneriyoruz.

KAYNAKLAR

[1] Pang, B., Lee, L., Vaithyanathan, S., Thumbs up?: sentiment classification using machine learning techniques. The Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, USA, 10, 79–86, 2002

[2] Cohen-Louck, K., Ben-David, S., Coping with terrorism: Coping types and effectiveness. International Journal of Stress Management, 24(1), 1–17, 2017

[3] Garg, P., Garg, H., Ranga, V., Sentiment Analysis of the Uri Terror Attack Using Twitter. International Conference on Computing, Communication and Automation (ICCCA2017), Greater Noida, India, 17–20, 2017

[4] Burnap, P., Williams, L., M., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., Voss, A., Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Social Network Analysis and Mining, 4(1), 1–14, 2014

[5] Chong, M., Sentiment analysis and topic extraction of the twitter network of #prayforparis. Proceedings of the Association for Information Science and Technology, 53(1), 2016

[6] Becker, K., Harb, J. G., Ebeling, R., Exploring deep learning for the analysis of emotional reactions to terrorist events on Twitter. Journal of Information and Data Management, 10(2), 97–115, 2019

[7] Institute for Economics. Peace., Global Terrorism Index 2019: Measuring the Impact of Terrorism,. Sydney, 2019

[8] Cioffi-Revilla, C., Computational social science. WIREs Computational Statistics, 2(3), 259–271, 2010

[9] Wise, E. K., Shorter, J. D., Social Networking and the Exchange of Information. Information Systems, 15(11), 103–109, 2014

[10] Yakushev, A., Mityagin, S., Social networks mining for analysis and modeling drugs usage. Procedia Computer Science, 29, 2462–2471, 2014

[11] Calvin Dark., Social Media and Social Menacing. Foreign Policy Blogs, 2011

[12] CBC News., Terrorist groups recruiting through social media - Technology & Science - CBC News. Retrieved from http://www.cbc.ca/news/technology/terrorist-groups-recruiting-through-social-media-1.1131053, 2016

[13] Telegraph Reporters., How terrorists are using social media. The Telegraph. Retrieved from http://www.telegraph.co.uk/news/worldnews/islamic-state/11207681/How-terrorists-are-using-social-media.html, 2014

[14] Kelley, J., Terror groups hide behind Web encryption. USA TODAY. Retrieved from http://usatoday30.usatoday.com/tech/news/2001-02-05-binladen.htm, 2001

[15] Russell, M. A., (1st ed). Mining the Social Web. In O'Reilly, Vol. 54, Issue 2, Beijing: O'REILLY, 2011

[16] Akcora, C. G., Carminati, B., Ferrari, E., Kantarcioglu, M., Detecting Anomalies in Social Network Data Consumption. Social Network Analysis and Mining, 4(231), 1–14, 2014

[17] Plantié, M., Crampes, M., Survey on Social Community Detection. In Social Media Retrieval. London, England: Springer Publishers, 2013

[18] Rai, A. K., A Survey on Link Prediction Problem in Social Networks. International Journal for Research in Applied Science and Engineering Technology, 5(4), 2017

[19] Alowibdi, J. S., Buy, U. A., Yu, P. S., Stenneth, L., Detecting Deception in Online Social Networks. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Beijing, China, 383, 2014

[20] Memon, N., Larsen, H. L., Principles Practical approaches for analysis, visualization and destabilizing terrorist networks. First International Conference on Availability, Reliability and Security (ARES'06), Vienna, Austria, 8 pp. – 913, 2006

[21] Zafarani, R., Liu, H., Behavior Analysis in Social Media. IEEE Intelligent Systems, 29(1–4), 4–7, 2014

[22] Elovici, Y., KANDEL, A., Last, M., Shapira, B., ZAAFRANY, O., Using Data Mining Techniques for Detecting Terror-Related Activities on the Web. Journal of Information Warfare, 3(1), 17–29, 2004

[23] De Groot, R., Data Mining for Tweet Sentiment Classification. Utrecht University, 2012

[24] Liu, B., Sentiment analysis and opinion mining. In Synthesis lectures on human language technologies. Morgan & Claypool Publishers, 2012

[25] Tsytsarau, M., Palpanas, T., Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478–514, 2012

[26] Wilson, T., Wiebe, J., Hoffmann, P., Recognizing contextual polarity in phrase level sentiment analysis. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada 347–354, 2005

[27] Hu, M., Liu, B., Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIKDD International Conference on Knowledge Discovery and Data Mining KDD 04, Seattle, WA USA, 04, 168–177, 2004

[28] Nicholls, C., Song, F., Comparison of Feature Selection Methods for Sentiment Analysis. Canadian Conference on Artificial Intelligence AI 2010: Advances in Artificial Intelligence, 286–289, 2010

[29] Yousefpour, A., Ibrahim, R., Hamed, H. N. A., Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. Expert Systems with Applications, 75, 80–93, 2017

[30] Das, B. R., Sahoo, S., Panda, C. S., Patnaik, S., Part of speech tagging in odia using support vector machine. Procedia Computer Science, 48(C), 507–512, 2015

[31] Wakade, S., Shekar, C., Liszka, K. J., Chan, C., Text Mining for Sentiment Analysis of Twitter Data. International Conference on Information and Knowledge Engineering IKE 12, 109–114, 2012

[32] Silva, A. P., Silva, A., Rodrigues, I., An Approach to the POS Tagging Problem Using Genetic Algorithms. Studies in Computational Intelligence, 577, 3–17, 2014

[33] Sebastiani, F., Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47, 2002

[34] Agarwal, B., Mittal, N., Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification. The 2nd

Workshop on Sentiment Analysis Where AI Meets Psychology, Mumbai, India, 17–26, 2012

[35] Zheng, L., Wang, H., Gao, S., Sentimental feature selection for sentiment analysis of Chinese online reviews. International Journal of Machine Learning and Cybernetics, 9(1), 75–84, 2018

[36] Zhang, Y., Gong, L., Wang, Y., An improved TF-IDF approach for text classification. Journal of Zhejiang University SCIENCE, 6(1), 49–55, 2005

[37] Ding, X., Liu, B., Yu, P. S., A holistic lexicon-based approach to opinion mining. Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08, California, USA, 231, 2008

[38] Medhat, W., Hassan, A., Korashy, H., Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113, 2014

[39] Chandrakala, S., Sindhu, C., Opinion Mining and sentiment classification a survey. ICTACT Journal on Soft Computing, 3(1), 420–427, 2012

[40] Yusof, N. N., Mohamed, A., Abdul-Rahman, S., Reviewing Classification Approaches in Sentiment Analysis. International Conference on Soft Computing in Data Science, 43–53, 2015

[41] Li, G., Liu, F., Sentiment analysis based on clustering: A framework in improving accuracy and recognizing neutral opinions. Applied Intelligence, 40(3), 441–452, 2014

[42] Claypo, N., Jaiyen, S., Opinion mining for thai restaurant reviews using K-Means clustering and MRF feature selection. Proceedings of the 2015-7th International Conference on Knowledge and Smart Technology, KST 2015, Chonburi, Thailand, 105–108, 2015

[43] Zhu, X., Semi-Supervised Learning Literature Survey Contents. In SciencesNew York, Madison, 10(1530), 2008

[44] Tang, J., Nobata, C., Dong, A., Chang, Y., Liu, H., Propagation-based Sentiment Analysis for Microblogging Data. Proceedings of the 2015 SIAM International Conference on Data Mining, Arizona, USA, 577–585, 2015

[45] Silva, N. F. F. da, Coletta, L. F. S., Hruschka, E. R., Jr., E. R. H., Using unsupervised information to improve semi-supervised tweet sentiment classification. Information Sciences, 355–356, 348–365, 2016

[46] Mudinas, A., Zhang, D., Levene, M., Combining lexicon and learning based approaches for concept-level sentiment analysis. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM ’12, Beijing, China, 1–8, 2012

[47] Machová, K., Marhefka, L., Opinion mining in conversational content within web discussions and commentaries. International Conference on Availability, Reliability, and Security, 8127 LNCS, Regensburg, Germany, 149–161, 2013

[48] Sharma, R., Nigam, S., Jain, R., Mining of Product Reviews at Aspect Level. International Journal in Foundations of Computer Science Technology, 4(3), 87–95, 2014

[49] Hu, N., Bose, I., Koh, N. S., Liu, L., Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision Support Systems, 52(3), 674–684, 2012

[50] Maks, I., Vossen, P., A verb lexicon model for deep sentiment analysis and opinion mining applications. The 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis WASSA 2011, Portland, Oregon, USA, 10–18, 2011

[51] Mukwazvure, A., Supreethi, K. P., A hybrid approach to sentiment analysis of news comments. 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 1–6, 2015

[52] Appel, O., Chiclana, F., Carter, J., Fujita, H., A hybrid approach to sentiment analysis. 2016 IEEE Congress on Evolutionary Computation (CEC), Cci, 4950–4957, 2016

[53] Duda, R. O., Hart, P. E. (). Pattern Classification and Scene Analysis. In John Wiley & Sons, Wiley Interscience Publication, 7(4), 1974

[54] Bilal, M., Israr, H., Shahid, M., Khan, A., Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University - Computer and Information Sciences, 28(3), 330–344, 2016

[55] Moraes, R., Valiati, J. F., Gavião Neto, W. P., Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621–633, 2013

[56] Akaichi, J., Social networks' Facebook' statutes updates mining for sentiment classification. 2013 International Conference on Social Computing, 886–891, 2013

[57] Feng, J., Xu, H., Mannor, S., Yan, S., Robust Logistic Regression and Classification. Neural Information Processing Systems (NIPS), 1–9, 2014

[58] González-Ibáñez, R., Muresan, S., Wacholder, N., Identifying Sarcasm in Twitter: A Closer Look. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 581–586, 2011

[59] Rokach, L., Maimon, O., (2nd ed.) Data Mining With Decision Trees - Theory and Applications, world scientific, 2015

[60] Wakade, S., Shekar, C., Liszka, K. J., Chan, C., Text Mining for Sentiment Analysis of Twitter Data. International Conference on Information and Knowledge Engineering IKE 12, Las Vegas, USA, 109–114, 2012

[61] Statista., Twitter: number of active users 2010-2018 | Statista. Statista. Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/, 2018

[62] Zhao, Y., R and Data Mining: Examples and Case Studies. Academic Press, Elsevier, 2012

[63] BBC News., London Bridge attack: Timeline of British terror attacks. BBC. Retrieved from https://www.bbc.com/news/uk-40013040, 2017

[64] The Guardian., Spain attacks – a visual guide. The Guardian. Retrieved from https://www.theguardian.com/world/2017/aug/17/what-happened-in-barcelona-las-ramblas-attack, 2017

[65] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 2019

[66] Blei, D. M., Ng, A. Y., Jordan, M. I., Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022, 2003

[67] Fang, Y., Si, L., Somasundaram, N., Yu, Z., Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model. The Fifth ACM International Conference on Web Search and Data Mining. ACM, 2012

[68] Alvarez-Melis, D., Saveski, M., Topic Modeling in Twitter: Aggregating Tweets by Conversations. The Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), Cologne, Germany, 2016

[69] Hong, L., Davison, B. D., Empirical Study of Topic Modeling in Twitter. The First Workshop on Social Media Analytics, 80–88, 2010

[70] Gaspar, R., Barnett, J., Seibt, B., Crisis as seen by the individual: the Norm Deviation Approach / La crisis vista por el individuo: el Enfoque de la Desviación de la Norma. Bilingual Journal of Environmental Psychology, 6(1), 2015

[71] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., Blei, D. M., Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems 22 (NIPS 2009), 288–296, 2009

[72] Jia, L., Yu, C., Meng, W., The effect of negation on sentiment analysis and retrieval effectiveness. He 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, 2009

[73] Hogenboom, A., Van Iterson, P., Heerschop, B., Frasincar, F., Kaymak, U., Determining negation scope and strength in sentiment analysis. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, 2589–2594, 2011

[74] Gupta, M. M., Forty-five years of fuzzy sets and fuzzy logic-A tribute to professor Lotfi A. Zadeh (the father of fuzzy logic). Scientia Iranica, 18(3 D), 685–690, 2011

[75] Dray, G., Plantié, M., Harb, A., Poncelet, P., Roche, M., Trousset, F., Opinion mining from blogs. International Journal of Computer Information Systems and Industrial Management Applications, 1(2150–7988), 205–213, 2009

[76] Neviarouskaya, A., Prendinger, H., Ishizuka, M., Semantically distinct verb classes involved in sentiment analysis. The IADIS International Conference Applied Computing, Rome, Italy, 2, 19-21, 2009

[77] Dragut, E. C., Fellbaum, C., The Role of Adverbs in Sentiment Analysis. Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore, Maryland, USA, 38–41, 2014

ÖZGEÇMİŞ

Ibrahim Amine FADEL, 13.03.1981'da Encemine'de doğdu. İlk, orta ve lise eğitimini Encemine'de tamamladı. 2000 yılında başladığı Afrika Uluslararası Üniversitesi Bilgisayar Mühendisliği Bölümü'nü 2004 yılında bitirdi. 2005 yılında Afrika Uluslararası Üniversitesi Üniversitesi'nde araştırma görevlisi olarak çalışmaya başladı. 2009 yılında Sudan Bilim ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü'nde yüksek lisans bitirdi. 2014 yılından beri Doktora eğitimine Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümü'nde devam etti.