TOBB EKONOMİ ve TEKNOLOJİ ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL MEDYA PROFİLLERİ ARASINDA BENZERLİK TESPİTİ VE GÖSTERİMİ

YÜKSEK LİSANS TEZİ

Ahmet Enis ERDOĞAN

Bilgisayar Mühendisliği Ana Bilim Dalı

Tez Danışmanı: Doç. Dr. Tansel ÖZYER

AĞUSTOS 2018

Fen Bilimleri Enstitüsü Onayı

……..
Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksininlerini sağladığını onaylarım.

……
Prof. Dr. Oğuz ERGİN
Anabilimdalı Başkanı

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün **151111056** numaralı Yüksek Lisans Öğrencisi **Ahmet Enis ERDOĞAN**' ın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"SOSYAL MEDYA PROFİLLERİ ARASINDA BENZERLİK TESPİTİ VE GÖSTERİMİ"** başlıklı tezi **08.08.2018** tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

Tez Danışmanı :	**Doç. Dr. Tansel ÖZYER** TOBB Ekonomi ve Teknoloji Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Ali Aydın SELÇUK (Başkan)** TOBB Ekonomi ve Teknoloji Üniversitesi	.....
	Dr. Öğr. Üyesi Hüseyin Uğur Yıldız TED Üniversitesi	.....

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Öğrenci Adı Soyadı

İMZA

ÖZET

Yüksek Lisans Tezi

SOSYAL MEDYA PROFİLLERİ ARASINDA
BENZERLİK TESPİTİ VE GÖSTERİMİ

Ahmet Enis ERDOĞAN

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Tansel ÖZYER

Tarih: Ağustos 2018

"Sosyal Medya" kullanıcıların bilgi paylaşımında bulunduğu platformlara verilen genel addır. Sosyal medya kullanımı son senelerde oldukça yaygınlaşmıştır. Kullanıcılar birden fazla Sosyal Medya Platformunda kişisel veya gündemdeki olaylar ile alakalı paylaşımlarda bulunmaktadır. Sosyal medyanın geniş kitleler tarafından kullanılması sosyal medya kullanıcıları hakkında bilgilerin çıkarılması ve kullanıcılar arasındaki benzerliklerin tespit edilmesi arayışını ortaya çıkarmıştır. Sosyal medya platformlarındaki kullanıcıların birbirlerine benzerliği tespit edildiği takdirde kullanıcıların eğilimleri, ilgi alanları, önem verdiği konular belirlenebilir. Ayrıca, reklamların hedef kitleye ulaşmasında da bu benzerlik bilgilerinden faydalanılabilir. Bununla birlikte, farklı amaçlarla gerçek sahibinin kim olduğunun doğrulanmasına ihtiyaç duyulan hesaplar da bu bilgiler ışığında tespit edilebilir. Bu çalışmada Doğal Dil İşleme(DDİ) teknikleri kullanarak kullanıcıların paylaşımları arasındaki benzerliklerin tespit edilmesi için geliştirdiğimiz teknikler sunulmuştur. Konu Modelleme ve

Adlandırılmış Varlık Tespiti teknikleri kullanılarak kullanıcıların yazılı paylaşımlarından nitelik çıkarımı yapılmıştır. Bu nitelikler Word Embedding teknikleri ve Word Mover's Distance tekniği kullanılarak kullanıcılar arasındaki benzerlikler çıkarılmıştır. DDİ tekniklerinin yanında kullanıcıların sosyal medya platformlarını kullandığı saatlerden, kullanıcı bilgilerinden ve kullanıcının arkadaşlarının isimlerinden yola çıkarak kullanıcılar arası benzerlik tespiti için teknikler önerilmiştir. Geliştirilen tekniklerden DDİ teknikleri ile oldukça etkili sonuçlar alınmıştır. Bu teknikler yardımı ile verilen bir kullanıcı kümesindeki toplulukların ve öbeklerin tespit edilebildiği gösterilmiştir. Ayrıca farklı sosyal medya platformlarındaki profiller arasında benzerlik tespiti yapılarak aynı kişiye ait sosyal medya profillerinin tespit edilebildiği gösterilmiştir. Çalışmada geliştirilen metriklerin farklı kullanıcı kümelerinde daha kolay denenmesi amacıyla bir web uygulaması geliştirilmiştir. Bu web uygulaması kullanıcılar arası benzerliklerin detaylı incelenmesini mümkün kılmıştır.

Anahtar Kelimeler: Sosyal medya Analizi, Doğal Dil İşleme, Kelime Katıştırma, Word2Vec, Word Mover's Distance, Adlandırılmış Varlık Tespiti, Konu Modelleme, Latent Dirichlet Allocation, Conditional Random Fields, Paragraph2Vec

ABSTRACT

Master of Science

DETECTION AND VISUALIZATION OF SIMILARITIES BETWEEN SOCIAL MEDIA PROFILES

Ahmet Enis ERDOĞAN

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Computer Engineering Master of Science Programme

Supervisor: Assoc. Prof. Tansel ÖZYER

Date: August 2018

Social Media is a generic name given to the digital platforms where users share information. The use of social media has become very popular in recent years. Users share personal information or their ideas about daily events on different social media platforms. The widespread use of social media has led to an interest to extract information from social media profiles and detect similarities among users. If similarity is detected between users, this information can be used to determine the tendencies, interests of users, the things that he/she cares about. Also such an information would make advertisements reach to their target audience. In addition, this information can be used in cases where there is a need to verify the real owner of a social media account. In this study, we have developed techniques to determine the similarities among the users using Natural Language Processing (NLP) techniques. Additionally attempts have been made to identify similarities between users by users' access hours to the platform and

users' friends list. However, more effective results were obtained with NLP techniques. Topic Modeling and Named Entity Recognition techniques were used to extract features from the posts of users. Similarities between users were derived by feeding these features to Word Mover's Distance algorithm. Effective results have been obtained using the techniques developed with NLP techniques. It has been shown that communities and clusters in a user set can be detected with the help of these techniques. It has also been shown that social media profiles belonging to a person can be detected by identifying similarities among profiles in different social media platforms. In addition, a web application is developed to make it easier to experiment with the metrics developed in this work.

Keywords: Social Media Analysis, Natural Language Processing, Word Embeddings, Word2Vec, Word Mover's Distance, Named Entity Recognition, Topic Modelling, Latent Dirichlet Allocation, Conditional Random Fields, Paragraph2Vec

İÇİNDEKİLER

ŞEKİL LİSTESİ

ÇİZELGE LİSTESİ

KISALTMALAR

DDİ : Doğal Dil İşleme
SMP : Sosyal Medya Platformu
YSA : Yapay Sinir Ağı
W2V : Word2Vec
WMD : Word Mover's Distance
EMD : Earth Mover's Distance
LDA : Latent Dirichlet Allocation
AVT : Adlandırılmış Varlık Tespiti
CRF : Conditional Random Fields

RESİM LİSTESİ

1. GİRİŞ

Çevrimiçi sosyal ağ siteleri dünyada gittikçe daha çok insanı etkilemektedir. Sayıları gittikçe artan sosyal ağ sitelerinde kişiler günlük kullanımlarında gerçek kimliklerini gizleyerek kolaylıkla kendi profillerini oluşturabilmektedirler. Kullanıcılar kendi erişim ve kullanım kolaylıkları açısından bir çok farklı hesabı tek bir hesap içerisinde bulundurabilen uygulamalar da dahil olmak üzer birçok açıdan sosyal medya kullanımlarını kolaylaştırabilmektedir. Diğer taraftan bakıldığında oluşturulan bu büyük sayıdaki kullanıcı profillerinin kime ait olduğunun tespiti gittikçe zorlaşmaktadır. Bu durum çözülmesi gereken sofistike bir problem ortaya çıkarmıştır.

Bu tezde, iki popüler sosyal ağ sitesi dikkate alınmıştır. Bunlar, birçok kullanıcı tarafından yaygın olarak kullanılan Twitter ve Facebook çevrimiçi sosyal ağlarıdır.

Çevrimiçi sosyal ağ siteleri, bir topluluktaki veya organizasyondaki bir içeriğin oluşturulmasını ve paylaşılmasını kolaylaştırır. Bu içerik farklı ilgi alanlarını, bakış açılarını ve uygulamaları desteklemek için farklı kaynaklarla oluşturulabilir.

World Wide Web (WWW), kullanıcıları bir araya getiren küresel altyapı oluşturmak için önde gelen nedenlerden biri olmuştur. Ancak, web teknolojilerine bağlı olarak pasif görüş açısıyla sınırlıydı. Web 2.0'ın ortaya çıkmasıyla, üyeleri tarafından oluşturulan topluluklar, etkileşimde bulunabilmekte ve işbirliği yapabilmektedirler.

Günümüzde birçok çevrimiçi sosyal ağ sitesi vardır; Bunlardan bazıları zamanla kullanılmadıkları için ömrünü tamamlarken bazıları ise hala günümüze kadar mevcudiyetini korumuştur. Sosyal ağ sitelerinin taksonomisi, temel özellikleri ve sosyal ağ sitelerinin evrimleşme sürecine ait değişiklikler ve faydalarına ait çalışmalar verilmiştir[1]. Buna göre, siteler, genel ve dikey (okul, meslek, hobilere, ilgi alanlarına, cinsiyete, yaşa, etnik kökene vb.) gibi amaçlarına göre kategorilere ayrılmışlardır. Çevrimiçi sosyal ağ siteleri, kullanıcıların gizlilik sınırlarını ihlal etme riskine rağmen üyelerini bu bilgileri açıklamaları konusunda teşvik etmek için tasarlanmıştır.

Bir çalışmada, bireyin kişisel gizliliğine dikkat etmesinin gerekliliği vurgulanmaktadır, kişisel bilgilerin gizliliğine dikkat edilmesi durumunda çevrimiçi sosyal paylaşım sitelerini ziyaret ederken kişisel bilgileri ifşa etmenin daha az olası olduğu ortaya koyulmuştur[2]. Bir kullanıcı sosyal ağ sitesinin sağladığı farklı özelliklerine göre bu platformu kullanıp kullanmak istemediğine karar vermektedir. Kullanıcılar gerçek kimliklerini vermekte özgürdürler. Olduklarından farklı da görünmeye çalışabilirler. Birey kendi kimliğinden ve hesaplarından haberdar olabilir ama dışarıdan aynı veya farklı sosyal ağ siteleri arasında aynı kullanıcıya ait farklı hesapların eşleştirilmesi ya da yakınlığını tespit etmek oldukça zor olmaktadır. Bu sorun, varlık çözümüne indirgenmiş, daha özel anlamda çevrimiçi sosyal ağlarda, farklı çevrimiçi profillerin çözümlenmesine indirgenmiştir. Belirli bir düzende olsun ya da olmasın varlıkların çözümlenmesi bu varlıkların çıkarılması ve eşleştirilmesi gibi bazı zorluklardan oluşmaktadır[4].

Malhotra ve diğer arkadaşları[6], birden çok sosyal medya platformunu kapsayan çalışmalarında kullanıcı profillerinin izdüşümünü incelemiştir (Twitter, YouTube ve Flickr). İzdüşüm oluşturmak için kullandıkları bilgi kaynakları, kullanıcı kimliği (kullanıcı adı), görünen ad, açıklama, konum ve bağlantı sayısı ile sınırlıdır. Dijital izdüşümlerini bir sosyal ağdan, başka bir sosyal ağa eşleştirerek kullanıcı profillerini tespit etmek için otomatik teknikler uygulamışlardır. Özellikle gözetimli bir modele sahip otomatik öğrenme araçları kullanmışlardır. Başka bir çalışmada da Twitter profili verilen bir hesaba dayanarak bu kişinin Facebook profili tespit edilmeye çalışılmaktadır. Bu çalışmada genel olarak kullandıkları yöntem kişinin profilinin aranması, kişinin içeriklerinin aranması ve bu kişinin diğer platformdaki kendi hesabındaki bir paylaşıma yer verip vermemesinden oluşmaktadır. Genel olarak bu bilgiler sosyal medya platformunun sağladığı arama özelliğinden faydalanılarak elde edilip, dönen sonuçla aranan içeriğin kosinüs benzerliği kullanılarak verilen Twitter kullanıcısına benzer Facebook hesaplarını tespit etmeye çalışmışlardır[3]. [5]'te büyük veri analizi için başlangıç aşaması olarak kural kümeleriyle kullanıcı adına dayalı eşleme önerilmiştir.

Bir çalışma, stilometriden ilham alan teknikler kullanmaktadır. Kullanıcıları eşleştirmek için kullanıcıların zamansal ve dilsel stillerinden faydalanılmaktadır[7]. Tezimize

benzeyen başka bir çalışmada, kullanıcıların sosyal medya profil bilgilerini ve ağ yapısını kullanan bir yöntem önerilmektedir[8].

Bu tez çalışmasını önceki çalışmalardan ayıran özellikleri, Doğal Dil İşleme tekniklerinden; Adlandırılmış Varlık Tespiti(AVT), Konu Modellemesi ve Kelime Katıştırma(W2V) teknikleri kullanılmasıdır. AVT için Conditional Random Field, Konu Modelleme için Latent Dirichlet Allocation, kelime katıştırma için paragraf vektörleri ve kelime vektörleri algoritmaları kullanılmıştır. Kullanıcı erişim davranışları arasındaki benzerliğin çıkarılması amacıyla yeni bir metrik sunulmuştur. Ayrıca, kullanıcıların arkadaşlarının kıyaslanmasında en benzer arkadaş isimlerinin eşleştirilmesi ve benzerlik çıkarılması için bir optimizasyon probleminin çözümünden yararlanılmıştır. Bu benzerlikleri görselleştiren bir sistem geliştirilmiştir. Benzer çalışmaların aksine Türkçe dili için hazırlanmış bir çalışmadır. Ancak kullanılan teknikler ve geliştirilen metrikler başka diller için kullanılmaya uygundur. Görselleştirme ile farklı parametrik değişkenlerle farklı sonuçların alınmasına olanak tanınmıştır.

Bu tez çalışmasında geliştirilen uygulama, şu anda Facebook [26] ve Twitter [27] sosyal ağ platformları için çalışmaktadır. Farklı ağ sitelerine ait hesaplar da eklenebilir.

Bu tezde verilen katkılar aşağıdaki gibi listelenebilir:

Farklı profillerin farklı seçeneklerle birleştirilmesiyle ilgili bilgilerin keşfedilmesi için veri toplanması, benzerlik hesaplama ve görselleştirme parçaları içeren bir sistem ortaya çıkarılmıştır. Gerçekleştirilen uygulamada profiller arasındaki tüm olası benzerlik sonuçları görselleştirilmiştir. Verilen bir kullanıcı farklı bakış açılarından kapsamlı bir şekilde analiz edilebilir.

Farklı bakış açılarının sağlanması analizler esnasında farklı roller oynamaktadır. Geliştirilen uygulama ile seçilen özellikler(konu, adlandırılmış varlıklar, en sık kullanılan kelimeler, hesap aktiviteleri, arkadaş listeleri) dahil edilebilir. Benzerliklerin yorumlanması kolay bir hale getirilmiştir. İlişkiler çıkarma, konular gibi yan ürünler de dahil olmak üzere profiller arasındaki benzerlik, çizge yapısı kullanılarak basit bir gösterimi sağlanmıştır.

Tezin ana hatları şu şekildedir: İlk bölüm giriş kısmını içermektedir. Tezin ikinci bölümünde kullanılan algoritmalar ile alakalı bilgiler verilmiş, üçüncü bölümde bu algoritma ve tekniklerin nasıl kullanıldığı açıklanmıştır. Dördüncü bölümde yapılan deneyler ve alınan sonuçlar paylaşılmıştır. Dördüncü bölümdeki ilk deneyde farklı sosyal medya platformlarındaki kullanıcıların benzerliklerinin tespit edilmesi ile aynı kişiye ait sosyal medya profillerinin tespit edilmesi amaçlanmıştır. İkinci deneyde verilen bir kullanıcı kümesindeki kullanıcıların birbirleri ile benzerlikleri tespit edildikten sonra bu benzerlikler ile kullanıcılar öbeklenmiştir. Beşinci bölümde geliştirilen "Profil Benzerlik Gösterim Aracı" hakkında bilgi verilmiştir. Son olarak, sonuç ve yapılabilecek çalışmalar ile tez sona ermektedir.

2. KULLANILAN DOĞAL DİL İŞLEME ALGORİTMALARI

2.1. Kelime Katıştırma (Word Embedding)

DDİ görevlerinin her birinde kelimeleri algoritmalara vermeden önce kelimeleri algoritmaların anlayacağı bir formata çevirmek gerekmektedir ve çoğu zaman kelimeleri bir çeşit vektöre dönüştürmek gerekmektedir. Kelimelerin nitelik olarak kullanılması için geliştirilen teknikler genellikle kelimelerin bir dokümandaki geçme sıklığına bakarak nitelikler çıkarıyor. Ancak bu teknikler verilen iki kelime arasındaki ilişki hakkında fikir yürütülmesine imkan vermemektedir. Word Embedding yaklaşımı ise, kelimelerin bulunduğu içerikleri de göz önüne alarak kelimelerin vektör olarak belirtilmesini amaçlar ve bu vektörler kelimelerin birbirleri arasındaki benzerliklerin anlaşılmasına da imkan verir. Word2Vec[word2vec], GloVE[11], FastText[12] Word embedding tekniklerinin önemli örnekleri olarak verilebilir.

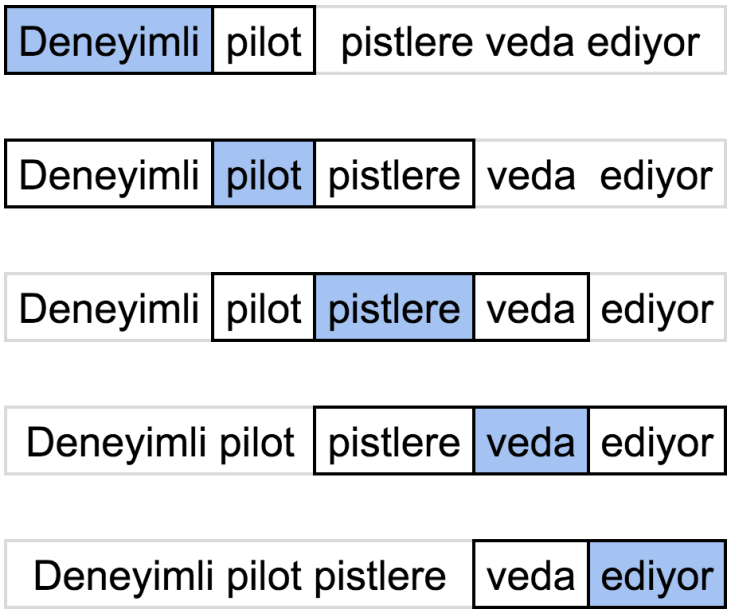
Word Embedding teknikleri ile kelimeler arasındaki ilişkileri yansıtması açısından kelimeleri nitelik olarak temsil etmede daha faydalı bilgiler kullanmamıza olanak sağlamaktadır. Örneğin “King” ve “Queen” arasındaki vektörel fark “Man” ve “Woman” arasındaki vektörel farka yakın çıkmaktadır. Bir başka etkileyici örnek “Italy” ve “Rome” arasındaki vektörel fark ile “France” ve “Paris” arasındaki vektörel farkın yakın çıkmasıdır. Yani Word Embedding teknikleri kelimeler arasındaki cinsiyet, başkent olma gibi konseptleri yakalamayı başarmaktadır. Kelimeler arasındaki bu çeşit anlamsal benzerlikleri frekans veya sayma tabanlı nitelik çıkarma teknikleri ile elde etmek mümkün olamamaktadır.

Word2Vec eğitim datanızda bir etiketleme gerektirmemesi bakımından unsupervised bir yaklaşımdır. Ancak korpusundan etiketleri kendisi çıkararak tek saklı katmana sahip bir yapay sinir ağını eğitmektedir. Eğitilen bu YSA’da önemli olan çıktılar değil; YSA’nın

eğitimi sırasında elde edilen ağırlıklardır(weight vector). Bu ağırlıklar her kelimenin vektörü çıktısı olarak belirtilmektedir.

Word2Vec kelime katıştırmaları iki farklı yaklaşım ile elde edilmektedir. Bunlar, Skip Gram ve Continuous Bag Of Words yaklaşımlarıdır.

Skip Gram yaklaşımında YSA'nın görevi bir cümle içinde geçen bir kelimeye bakarak bu kelimenin komşuluğunda olabilecek kelimelerin olasılıklarını tahmin etmektir. Yani Skip Gram yaklaşımında YSA girdi olarak bir kelime alır ve bu kelimenin etrafındaki kelimelerin geçme olasılığı çıktı olarak verilir. Yakındaki kelimelerin sayısı bir pencereye göre belirlenir. Her kelime için yakınındaki kelimelerin geçme olasılıkları verilen bir korpustan çıkarılır. YSA'daki eğitim verisinin nasıl oluşturulduğu aşağıdaki şekilde verilmektedir.



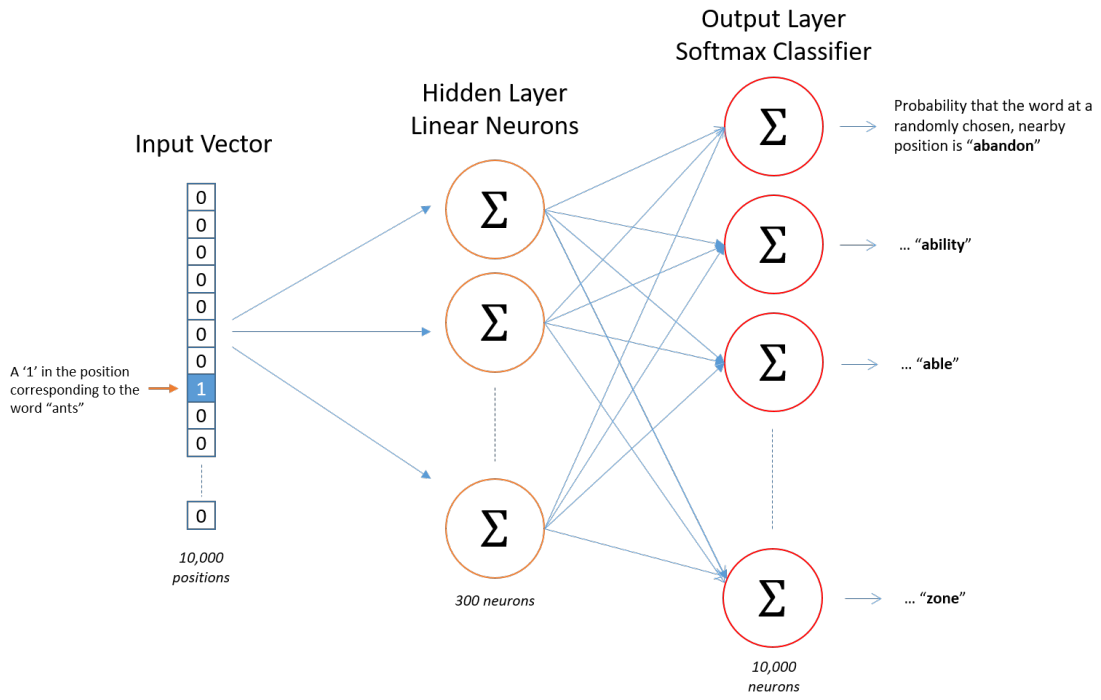
Şekil 2.1: Skip Gram Word2Vec yaklaşımında cümleden eğitim verisi oluşturma

Burada window boyutu bir olarak belirlendiği için kelimenin hemen bir sonraki komşuluğundaki kelimeler ele alınmıştır. Bu cümleden çıkarılan eğitim verisi numuneleri aşağıdaki gibidir.

[(Deneyimli, pilot), (pilot, Deneyimli), (pilot, pistlere), (pistlere, pilot), (pistlere, veda), (veda, pistlere), (veda, ediyor), (ediyor, veda)]

Buradan her ikilinin geçmesine göre verilen bir kelimenin etrafındaki kelimelerin geçme olasılıkları çıkarılarak yapay sinir ağı eğitilmektedir.

Skip Gram YSA'sının mimarisi aşağıda verilmiştir.



Şekil 2.2: Skip Gram Modelinin YSA mimarisi [10]

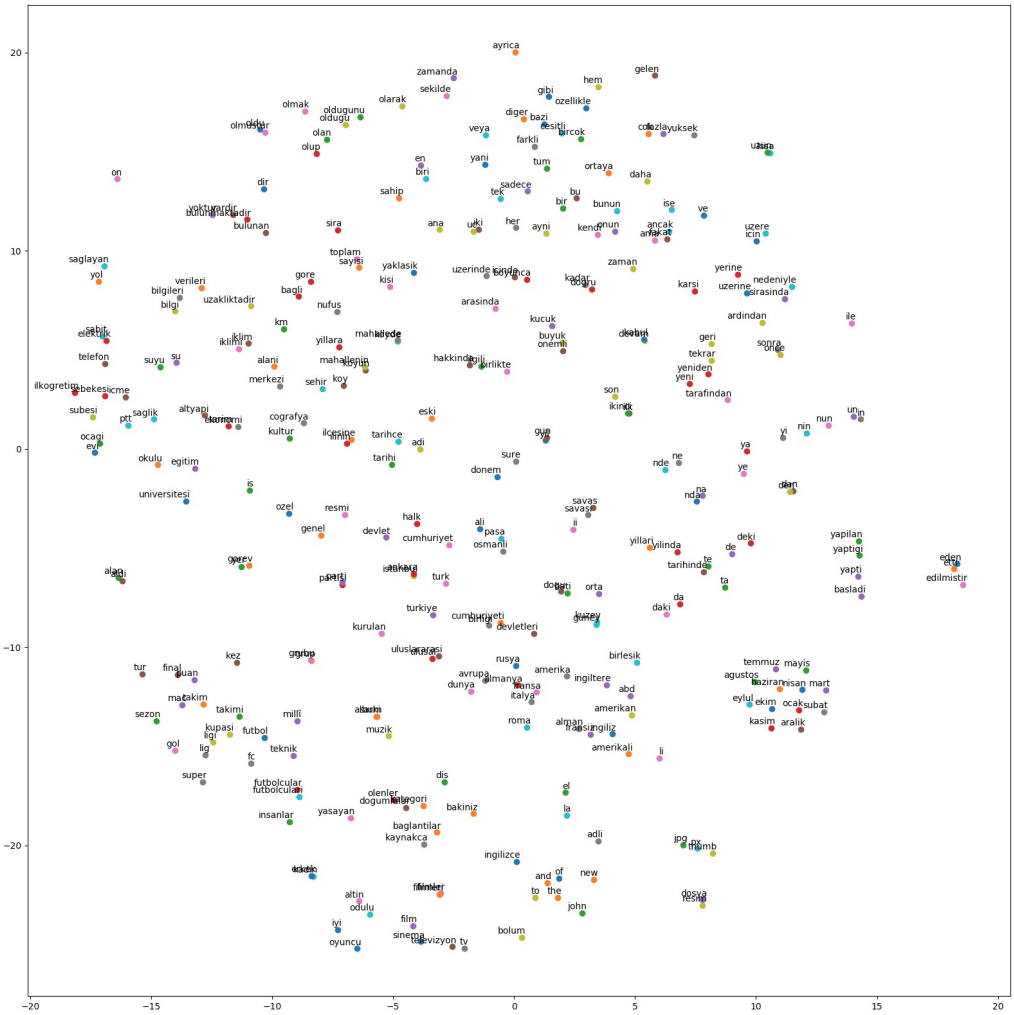
Kelimeler one-hot encoding (tek elemanı 1 geri kalan elemanları 0 olan bir vektör) ile Yapay Sinir Ağına girmektedir. Hedef olarak etrafındaki kelimelerin geçme olasılıkları verilmektedir. Çıktı katmanı için bir Softmax sınıflandırıcı kullanılmaktadır. YSA Backpropagation ile saklı katman ağırlıkları yakınsadıktan sonra saklı katmandaki her kelime için elde edilen ağırlıklar o kelimenin "kelime vektörü" olarak kullanılmaktadır.

Continuous Bag of Words yaklaşımında kullanılan YSA Skip Gram'daki YSA'nın ters çevrilmiş hali olarak düşünülebilir. Burada farklı olarak etrafındaki kelimeler girdi katmanında verilirken hedef olarak merkezdeki kelime verilmektedir. Ayrıca gizli katmanın çıktılarının ortalaması alınarak YSA eğitilir. Bu şekilde eğitilen YSA'nın gizli katmandaki ağırlıkları kelime vektörü olarak kullanılır.

Sonuç olarak, Word2Vec'te YSA'nın sadece eğitim sonrası oluşan ağırlıkları dikkate alınır.

Bu çalışmada eğitim verisi olarak Türkçe Wikipedia makaleleri kullanılmıştır. Kelimeler küçük harflere ve Türkçe karakterlerin ascii karşılıklarına çevirildikten sonra noktalama işaretleri kaldırılmıştır.

Ayrıca 100 kereden az geçen kelimeler korpusa dahil edilmemiştir. YSA'nın gizli katmanındaki düğüm sayısı 200 olarak belirlenerek Continuous Bag-of-Words yaklaşımı ile eğitilmiştir. Bu şekilde bir Türkçe kelime katıştırması(word embedding) elde edilmiştir. Aşağıdaki şekilde elde edilen kelime katıştırmanın t-SNE [15] kullanılarak iki boyuta indirilmiş hali verilmiştir. Şekilde en sık geçen 200 kelimeler gösterilmektedir.



Şekil 2.3: Elde edilen Türkçe kelime vektörlerinin t-SNE ile iki boyutta gösterimi

2.2. Word Mover's Distance (WMD)

WMD Kusner et al. tarafından, iki doküman arasındaki uzaklığı hesaplamak için önerilmiş bir tekniktir. WMD Word2Vec veya başka bir word embedding üzerinde çalışan bir tekniktir. Dokümanlardaki kelimelerin kelime katıştırmaları birbirleri ile yakın olacak şekilde eşleştirildikten sonra her eşleşmeden oluşan fark çıkarılarak iki doküman arasındaki uzaklık bu farkların toplamı olarak hesaplanmaktadır. WMD kelime katıştırmalarını kullanarak aynı kelimeler geçmese bile kelimeler arasındaki anlamsal yakınlıkları kullanarak iki doküman arasındaki benzerliği çıkarmaktadır[16].

WMD bir ulaştırma problemi olan EMD'nin çözümünü kullanarak sonuca ulaşmaktadır.

$\mathbf{T} \in \mathbb{R}^{n \times n}$ bir akış matrisi olarak kabul edelim ve d dokümanındaki i kelimesinden d' dokümanındaki j kelimesine giden akis $\mathbf{T}_{ij} \geq 0$ olsun. d'yi tamamen d` dokümanına akıtmak icin d'nin kelimelerinden çıkan akışların toplamının d` dokümanının kelimelerine gelen akış toplamına eşit olması gerekmektedir. Sonuç olarak iki doküman arasındaki uzaklık d den d` ye bütün kelimeleri minimum şekilde taşıyacak akışın ağırlıklı olarak toplamı ile elde edilir.

Bu açıklamanın lineer programlama olarak ifade edilmiş hali aşağıda verilmiştir.

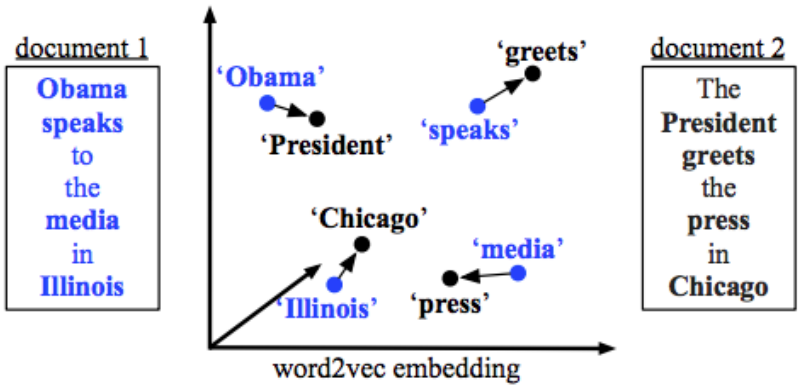
$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j)$$

$$\text{subject to: } \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}.$$

[16]

İki dökümandaki kelime sayıları birbiri ile aynı olduğu durumda içgüdüsel olarak birbirine en yakın kelimeler arasındaki eşleşmelerden oluşan farkların toplamı dokümanlar arasındaki fark olmuş oluyor bu durumu aşağıdaki figürde verilmiştir.



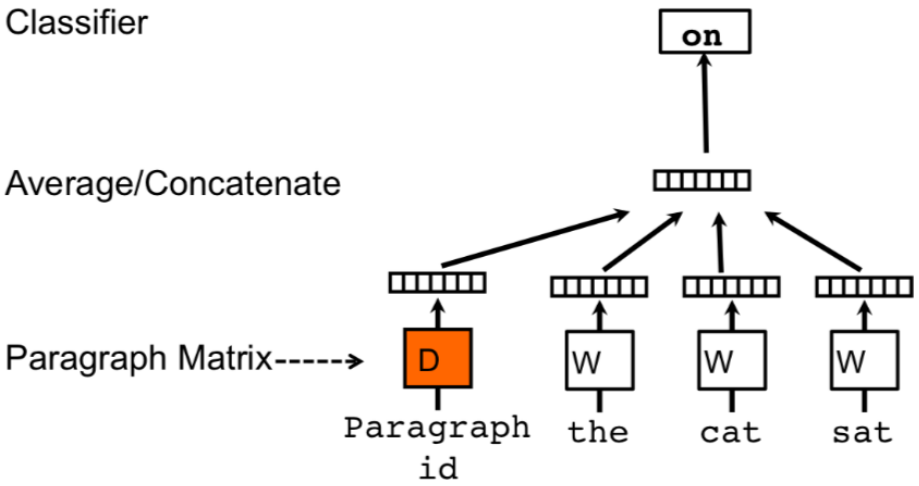
Şekil 2.4: WMD'nin verilen dokümanlarda kelimeleri eşleştirmesi [16]

2.3. Paragraf Vektörleri

Paragraf Vektörleri dokümanları, uzunluklarından bağımsız sabit boyutta bir vektör olarak temsil eden bir tekniktir. Paragraf Vektörleri dokümanların birbirlerine benzerliklerinin tespiti için Word2Vec'e benzer bir neural network yaklaşımı kullanan Mikolov et.al. tarafından geliştirilmiş bir tekniktir[17]. Mikolov et. al çalışmalarında Paragraf Vektörler kullanılarak birçok yazı sınıflandırma ve duygu analizi görevlerinde bilinen en iyi yöntemlerden daha iyi sonuçlar verdiğini göstermiştir.

Paragraf Vektörleri Word2Vec'teki gibi bir Yapay Sinir Ağının eğitilmesiyle elde edilir. Aynı şekilde CBOW ve Skip-Gram mimarilerine benzer Yapay Sinir Ağları kullanılmaktadır. Word2Vec'teki inputa ek olarak doküman vektörü de girilmektedir. CBOW yaklaşımına benzer olan Yapay Sinir Ağı mimarisindeki Paragraf Vektörü "Distributed Memory Model of Paragraph Vectors (PV-DM)" şeklinde ifade edilmektedir. Word2Vec'teki bir kelime etrafındaki diğer kelimeleri içerik kelimeler olarak belirtildiğinde, Paragraf Vektörünün eğitildiği Yapay Sinir Ağında içerik kelimelerine ek olarak her doküman için de one-hot encoded bir vektör kullanılmaktadır. Aynı doküman için oluşturulan içerik vektörleri için bir doküman vektörü kullanılarak Yapay Sinir Ağı eğitilir. Yani doküman vektörlerini de kelimeler olarak düşünebiliriz tek farkı içerik kelimeleri bir doküman için bir pencere boyunca değişirken doküman vektörü değişmeden Yapay Sinir Ağında kullanılır.

Oluşturulan YSA'da paragraf vektörleri ve kelime vektörleri stochastic gradient descent ile eğitilerek elde edilmektedir. Burada gradient Back Propagation ile elde edilmektedir. Yeni bir paragraf için doküman vektörüne yeni bir kolon eklendikten sonra kelime vektörleri ve Softmax ağırlıkları aynı tutularak gradient descent yaklaşımıyla yeni paragraf için bir paragraf vektörü elde edilmektedir. Aşağıdaki figürde PV-DM yaklaşımının Yapay Sinir Ağı mimarisi verilmiştir.

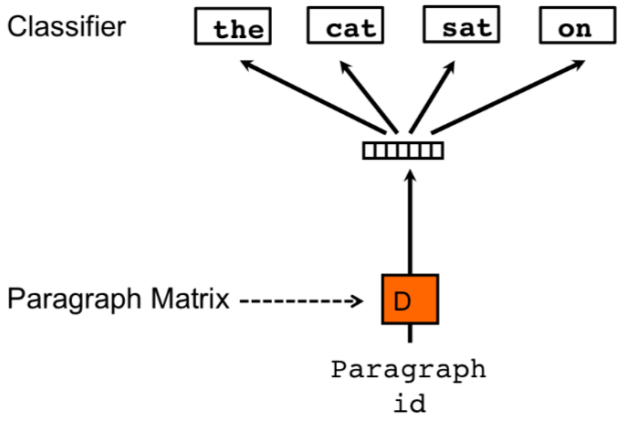


Şekil 2.5: PV-DM paragraf vektör yaklaşımının girdi ve çıktı gösterimi [17]

Figürde paragraf vektörü ile word vektörlerinin ortalamasının alınması veya birleştirilmesi ile elde edilen vektörün Softmax sınıflandırmaya sokulmaktadır. Bu model paragraf vektörünün ve 3 içerik kelimesinin vektörlerinin birleştirilmesi veya ortalamasının alınması ile içerikteki 4. kelimenin tahmin edilmesi için kullanılmaktadır. Backpropagation ile elde edilen ağırlıklar paragraf ve word vektörlerini oluşturmaktadır. Burada paragraf vektörü verilen kelimelerden eksik olan 4. kelimenin tahmin edilmesi için kullanılması bakımından bir çeşit hafıza görevi görmektedir. Bu duruma atfen bu modele Distributed Memory Model of Paragraph Vectors (PV-DM) adı verilmiştir.

"Distributed Bag-of-Words version of Paragraph Vector(DBOW-PV)" yaklaşımında paragraf vektörleri verilen bir doküman IDsinden içerikteki kelimeleri tahmin etmek için

oluşturulmuş bir Yapay Sinir Ağının eğitilmesi sonucu elde edilmektedir. Bu yaklaşım Word2Vec'teki Skip-Gram yaklaşımına benzemektedir. DBOW-PV ile word vektörlerinin tutulması gerekmektedir; sadece Softmax ağırlıklarının tutulması yetmektedir; bu bakımdan bellek kullanımında PV-DM yaklaşımına göre daha verimlidir.



Şekil 2.6: DBOW-PV paragraf vektör yaklaşımının girdi ve çıktı gösterimi [17]

2.4. LDA ile Konu Modelleme

Günümüzde eldeki oldukça büyük boyuttaki yazı verilerini analiz etmek için gözetimsiz makine öğrenmesi tekniklerine olan ihtiyaç daha fazla. Genellikle bir dokümanda veya cümlede önemli olan birkaç kelime var ve bu kelimeler dışındaki kelimeler cümlenin gramatik olarak doğruluğunu sağlamak için kullanılan kelimeler. LDA unsupervised bir konu modelleme yaklaşımıdır, cümleleri oluşturan konuları ve bu konuları oluşturan kelimeleri bulmayı amaçlamaktadır[13].

LDA cümlelerin bir konu dağılımından, konuların da bir kelime dağılımından oluştuğunu kabul ederek konu modelleme yapmaktadır. Konuların kelimelerin Dirichlet dağılımından cümleler de konuların Dirichlet dağılımından oluştuğunu kabul ederek Gibbs Sampling veya variational inference teknikleri ile bir numune bularak dokümandaki konu dağılımı ve konulardaki kelime dağılımlarını yakınsayarak

bilinmeyen değişkenler için bir numune bulunmaktadır. Bu numune de konuların kelime dağılımlarının eğitilmiş hali olarak kullanılmaktadır.

LDA'da her konu kelimelerin dağılımından, her doküman konuların dağılımından elde edildiği varsayılmaktadır. Başlangıçta sadece dokümanları oluşturan kelimeler bilinmektedir. Burdan latent değişkenler ile alakalı çıkarım yapılması gerekmektedir.

LDA modelin açıklaması aşağıdaki gibidir:

1. k = 1… K:
    a. $\phi^{(k)} \sim Dirichlet(\beta)$
2. Her doküman $d \in \mathbf{D}$ icin:
    a. $\theta_d \sim Dirichlet(\alpha)$
    b. Her kelime $w_i \in d$ icin:
        i. $z_i \sim Discrete(\theta_d)$
        ii. $w_i \sim Disctete(\phi^{(zi)})$ [18]

Burada korpustaki K gizli(latent) konu sayısı, $\phi^{(k)}$ k konusunu oluşturan kelimelerin bu konu içindeki dağılımı, θ_d dokümanlardaki konuların dağılımı, z_i ise w_i kelimesinin konu indeksini gösteriyor. α ve β kelimelerin konulara ve konuların dokümanlara olan dağılımının çekildiği Dirichlet dağılımının hyper parametreleri. Eğer çok küçük bir α değeri seçildiğinde, bir doküman tek bir konudan oluşurken α değerinin yüksek seçilmesi durumunda bir doküman birkaç konudan oluşan bir dağılım alır. Aynı şekilde β düşük seçildiğinde bir konu çok az kelime tarafından oluşan bir dağılım alırken yüksek β değerlerinde bir konu daha fazla kelime tarafından oluşturulmuş bir dağılım alıyor.

Bu model bize aşağıdaki bileşik olasılık dağılımını vermektedir[18].

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z)$$

Bu joint probability dağılımdan z, θ ve ϕ gizli(latent) değişkenlerini bulmak LDA'nın amacıdır. Yani aşağıdaki posterior dağılımdan çıkarım yapmak LDA'nın çözmesi gereken asli problemdir.

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

Ancak bu posterior dağılımın hesaplaması intractable zorlu olduğu icin inference teknikleri kullanılması gerekmektedir. Bunun icin Gibbs Sampling veya Variational Inference teknikleri kullanılabilir.

Gibbs sampling kısaca şu şekilde çalışan bir yöntem.

$p(x) = p(x_1, x_2, \ldots, x_m)$ bileşik olasılık dağılımı için

1. Her x_i a rastgele bir deger ata
2. t = 1….T
   - 2.1. $x_1^{t+1} \sim p(x_1| x_2^t, x_3^t, \ldots, x_m^t)$
   - 2.2. $x_1^{t+1} \sim p(x_2| x_1^{t+1}, x_3^t, \ldots, x_m^t)$
   - 2.m. $x_1^{t+1} \sim p(x_m| x_1^{t+1}, x_2^{t+1}, x_3^{t+1}, \ldots, x_{m-1}^{t+1})$

Gibbs Sampling ile joint distribution p(theta, phi, z | w) daki latent değişkenlerin numuneleri alınabilir. Her bir latent değişken için yapılan iterasyonların ardından Gibbs Sampling ile elde edilen numuneler p(θ), p(ϕ), p(z) dağılımlarından elde edilen numunelere yakın sonuçlar elde edilmesine yol açacaktır. Ancak, burada iterasyon sayısı bilinmemektedir. Collapsed Gibbs Sampling yaparak üç tane latent değişken için Gibbs örneklemi yapmamıza gerek kalmadan sadece z üzerinde Gibbs Sampling ile var olan diğer iki latent değişkenden de örneklemler oluşturulabilmektedir. Collapsed Gibbs Sampling'in çalışmasınının detayı [18]'da verilmektedir.

2.5. Adlandırılmış Varlık Tespiti

Adlandırılmış Varlıkların Tespiti (Named Entity Recognition) gorevi bir cümledeki kişi, yer, kuruluş, tarih kelimelerini tespit etme görevidir. AVT icin kullanılan yaygın metodlar Hidden Markov Modelleri ve Conditional Random Fields(CRF)[14] yaklaşımıdır. Bu çalışmada AVT icin HMM'e kıyasla daha fazla nitelik kullanımına imkan verdiği için CRF kullanılmıştır [crf tutorial].

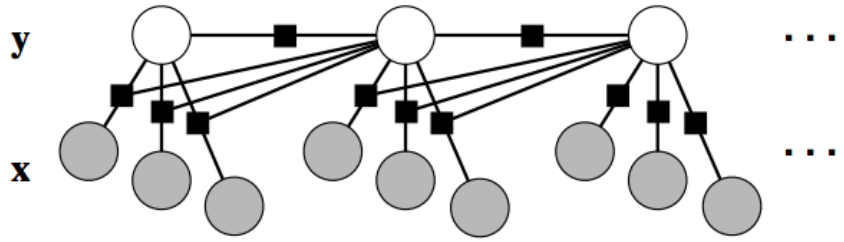
CRF bir Probabilistic Graphical Model(PGM) yaklaşımıdır. Probabilistic Graphical Modeller birbirine bağlı birçok değişkeni alt kümelere ayırıp ilişkileri bu alt kümeler arasına sınırlandırarak hesaplamayı daha verimli ve değişkenlerin ilişkilerini daha anlaşılır bir biçimde temsil etmek için kullanılan bir modelleme biçimidir. PGM'de conditional independence varsayımları ile değişkenin bağlı olduğu diğer değişkenlerin sayısı azaltılır[crf tutorial].

CRF verilen bir dizi içerisindeki elemanların etiketlerini doğru bir şekilde tahmin etmeye çalışır. Bu dizi bir resim, DNA dizisi veya bir cümle olabilir. AVT durumunda cümledeki kelimeler CRF'e dizi olarak verilir. PGM'de değişkenlerin birbirleri arasındaki ilişkiler conditional independence varsayımları ile azaltıldığı daha önce belirtilmişti. Bir cümle içinde geçen her kelime bir bütünsel anlama katkıda bulunduğu için birbirlerine bağlı olsalar da AVT görevi için CRF modellemesi kullanarak kelimelerin birbirleri ile bağlantısı sadece komşu kelimelerle sınırlandırılarak hesaplama kolaylaştırılmıştır. AVT için linear-chain CRF kullanılmaktadır. Bu linear-chain CRF in formülünü aşağıda bulabilirsiniz.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

Burada f_k bir nitelik fonksiyonu ve θ_k bu k niteliğin katsayisi. y kelimenin etiketi ve x kelimenin kendisidir. Z(x) ise bütün olası durumların olasılıklarının toplamı ile elde edilen normalizasyon sabitidir.



Şekil 2.7: CRF yapısının girdi ve çıktıların koşullu olasılıkları arasındaki ilişkiler [19]

3. YAPILAN ÇALIŞMA

Bu çalışmada LDA ve AVT nitelik çıkarımı için kullanılmıştır. Bu tekniklerden elde edilen kelimelerin kelime katıştırmaları(word embedding) kullanılarak WMD ile kıyaslanıp kullanıcıların yazılı paylaşımları arasında benzerlik çıkarımı yapılmıştır. Ayrıca kullanıcıların en sık kullandığı kelimeler de nitelik olarak kullanılmıştır.

Paragraf Vektörlerine kullanıcıların yazılı paylaşımları ön işlemden geçirilerek direk olarak nitelik olarak kullanılmıştır.

Yazıların ön işlemi için noktalama işaretlerinin çıkarılması, Türkçe karakterlerin ascii karakterlerine dönüştürülmesi, linklerin ve sayıların kaldırılması teknikleri kullanılmıştır. AVT'den önce sadece Türkçe karakterlerin ascii karakterlere dönüştürülmesi tekniği kullanılmıştır. Bunun sebebi AVT'de nitelik olarak noktalama işaretleri ve kelimede büyük harf olup olmaması gibi özelliklerin de kullanılmasıdır.

3.1. Latent Dirichlet Allocation(LDA) ile Nitelik Çıkarımı

LDA tekniği ile verilen bir doküman listesinin konularını oluşturan önemli kelimeler bulunabilir. Kullanıcıların gönderilerindeki önemli kelimeleri LDA yardımıyla çıkararak kullanıcıların kıyaslanması esnasında önemsiz olan kelimelerin sistemde değerlendirmesi önlenebilir. Döküman kümesi olarak kullanıcıların yazı olarak paylaştıkları gönderiler kullanılmaktadır. Her kullanıcı için 2 ile 20 konu için LDA modellemesi yapılmaktadır. Bu şekilde elde edilen modellemelerden "coherence" puanı en yüksek olan modelleme seçilmektedir. Coherence puanı Michael Röder et. al. [20] tarafından geliştirilen teknik ile elde edilmektedir. Bu şekilde her kullanıcı için bir LDA modellemesi oluşturulmaktadır.

Çıkan LDA sonuçları kullanıcı gönderilerinin genellikle hangi konulardan oluştuğu ile alakalı bilgi vermesi açısından yararlı bilgiler içereceği ve kullanıcıların birbirine benzerliklerini belirlemek için iyi bir özellik olarak kullanılabileceğini göstermektedir.

3.2. Adlandırılmış Varlıkların Tespiti (AVT) ile Nitelik Çıkarımı

Kullanıcı gönderilerinden bahsedilen varlıkların çıkarılması kullanıcıları birbirlerine karşı kıyaslarken benzerliklerin daha doğru tespit edilmesinde faydalı bir özellik olarak kullanılmıştır.

Adlandırılmış varlıkların tespiti için bir probabilistic graphical model olan Continuous Random Field(CRF) algoritması kullanıldı. Sosyal medya verisi gibi gramer kurallarına genellikle uyulmayan yazılarda AVT görevini yüksek doğrulukta başarabilmek oldukça zor bir görev. En gelişmiş AVT modelleri imla kurallarına uygun yazılmış haber verileri üzerinde eğitilmekte ve bu modeller özellik olarak tokende noktalama işareti, baş harfin büyük olmadığı gibi özellikler kullanılmaktadır.

Ayrıca Türkçe için geliştirilmiş kullanabileceğimiz bir adlandırılmış varlık eğitim verisi bulunmadığı için kendi eğitim verimizi oluşturmamız gerekti. Yaklaşık 2000 tweet i "kişi, tarih, yer, organizasyon ve diğer" olarak etiketledik. bu süreci hızlandırmak için ufak bir web uygulaması geliştirilmiştir. Sonuç olarak bu beş sınıf için yaklaşık %70 lik bir doğruluk başarısı elde edildi. Bu başarının elde edilmesinde normal olarak kullanılan özellikler dışında word2vec embedding ile bir kelime öbekleme işi yapıldı. Öbeklemeden sonra herhangi bir kelimenin hangi öbeğe düştüğünü bularak bu öbeğin ID si de özellik olarak CRF e verildi. Bu şekilde yaklaşık %4 bir doğruluk payı iyileşmesi gerçekleştirilmiştir.

Kullanıcıların gönderilerinde Diğer sınıfı dışındaki Bütün kelimeleri olduğumuz için aslında beşli bir sınıflama görevinden ziyade diğer ve adlandırılmış varlık olarak ikili bir sınıflandırma görevi yerine getirilmektedir. Bu durumda, bir kelimenin adlandırılmış bir varlık olup olmadığı %70'den daha yüksek bir doğruluk payı ile tespit edilmektedir.

3.3. Kelime Katıştırma (Word Embedding) Tekniği ile Anlamsal Çıkarımların Yapılması

Genellikle doğal dil işleme tekniklerinde özellik çıkarmak için bag of words veya TF-IDF gibi kelimeleri bir vektöre çeviren teknikler kullanılmaktadır. Ancak bu durumda kelimelerin sırası göz ardı edilmektedir. Bu durumda benzer içeriklerde kullanılan kelimelerin birbirine yakınlığı vektör gösterimlerinden belirlemek mümkün olamamaktadır. Word embedding teknikleri kelimenin etrafındaki diğer kelimeleri de dikkate alarak kelimelerin bulunduğu içeriği değerlendiren bir vektör olarak kelimeleri tutmaktadır. Bunun etkileyici örnekleri aşağıdaki gibi vektör toplamı ve çıkarmasından elde edilen vektörün anlamsal olarak yakın olan bir kelimeye denk gelmesidir.

$$vector("man") - vector("woman") \cong vector("king") - vector("queen")$$

$$vector("Paris") - vector("France") + vector("Italy") \cong vector("Rome")$$

Bu çalışmada word embedding tekniklerinden word2vec[9]'i kullanarak kullanıcı gönderileri birbirleriyle kıyaslandı. Word2vec'i kullanarak dokümanların birbirine benzerliğini tespit etmek için Word Mover's Distance(WMD)[16] metriği kullanıldı. Bu teknik Earth Mover's Distance dan esinlenilerek geliştirilmiş bir teknik. Bu metriğin önemli özelliklerinden birisi de farklı sayıda kelime içeren dokümanların birbiriyle kıyaslamasını da yapabiliyor olmasıdır. WMD daha önce de belirtildiği gibi bunu bir akış fonksiyonunu minimize ederek başarmaktadır.

Word2Vec gerçeklemesi olarak "gensim" [21] DDİ kütüphanesindeki word2vec modeli kullanıldı. Türkçe Wikipedia korpusu üzerinde word2vec algoritması eğitilerek Türkçe kelimelerin vektörleri elde edildi. Her kullanıcının gönderilerinden nitelikler çıkarıldıktan sonra her kullanıcının gönderileri tek bir doküman gibi değerlendirilerek diğer kullanıcıların gönderileri ile kıyaslandı. Bu kıyaslamada WMD[16] metrik olarak kullanılmıştır. Ortaya çıkan sonuçlar kullanıcıların gönderilerinin benzerliklerini tespit etmek için kullanılabileceğini kanıtlar niteliktedir.

3.4. Paragraf Vektörlerin Çıkarılması

Derin öğrenme teknikleri kullanılarak oluşturulan word-embedding lerden ilham alınarak geliştirilmiş bu teknik ile dokümanlar etiketlenerek Word2Vec'te kullanılan skip-gram veya distributed bag of words şeklindeki Yapay Sinir ağları eğitilerek her doküman için bir vektör belirlenmektedir.

Paragraph2Vec algoritmasında Word2Vec'e ek olarak eğitim esnasında doküman IDsi de girdi olarak verilir; bu şekilde her doküman için de bir vektör oluşturulur.

Bu teknik her kullanıcının gönderileri kullanıcı kimliği etiket olarak kullanılarak paragraph2vec algoritmasına sokulmaktadır. Bu şekilde her kullanıcı için istenen boyutta bir uzayda vektör çıkarılarak kullanıcıların birbirlerine benzerlikleri çıkarılmıştır.

Paragraph2Vec uygulaması olarak gensim kütüphanesindeki[21] Doc2Vec gerçeklemesi kullanıldı.

Paragraf vektörlerinin çıkarılmasında kullanıcıların gönderi sayıları fazla olmadığı durumlarda korpus küçük olacağından Doc2Vec'ten verim almak mümkün olmamaktadır. Bu nedenle kullanıcıların postlarından Adlandırılmış Varlıklar çıkarılıp bu varlıkların geçme frekansına göre önemli olan 100 tanesi ile tweetler toplanmaktadır, bu şekilde yaklaşık 100*200 = 20000 tweet(doküman) elde edilmiş oluyor. Ayrıca son tarihlerdeki varlıkların da önemli olabileceği göz önüne alınarak son 1 aydaki bütün varlıklarla Twitter da search yapılıp gene her bir entity için en fazla 200 tweet olacak şekilde bir veri toplama faslı gerçekleştirilmektedir. Bu şekilde en az 20000 lik bir doküman kümesi oluşturularak Doc2Vec ile kullanıcıların vektörleri çıkarılır. Kullanıcı vektörleri 50 boyutlu bir uzayda temsil edilerek birbirine yakınlıklarına göre kullanıcılar arasındaki benzerlik değerleri elde edilmektedir.

3.5. Kullanıcı Erişimlerinin Kıyaslanması

Kullanıcıların platformdaki kullanım biçimleri iki kullanıcının benzer zamanlarda SMP leri kullanmaları konusunda bilgi vermektedir. Bu bilgiden faydalanmak için

kullanıcıların paylaşım yaptıkları saatlere ihtiyaç duyulmaktadır. Bu tarihlerden birkaç farklı şekilde yararlanabileceği öngörülmüştür. Bunlardan birincisi, kullanıcının haftalık kullanım davranışıdır. Haftanın hangi günleri hangi saatlerde paylaşım yapıldığına bakılarak kullanıcıların haftalık kullanım davranışları elde edilmektedir. Bu şekilde, iş dönüş saatlerinde yada belli bir günün akşam veya sabah saatlerinde paylaşımda bulunulması bu kişilerin benzer olabileceği yönünde bir nitelik olarak değerlendirilmektedir.

İkincisi, kullanıcıların son 3 ayda her gün için ayrı olarak hangi gün ve hangi saatlerde kaç paylaşım yaptığına bakılarak 3 aylık kullanım alışkanlıkları çıkarılmıştır.

Bu iki nitelik çıkarımı için de bir metriğe ihtiyaç vardı. İki durum için de kullanıcıların aktivite vektörleri binary vektörlere dönüştürülmüştür. Bu vektöre A diyelim aşağıda A vektörünün oluşturulma biçimi verilmiştir

$$foreach\ x\ in\ Activity \quad A_i = \begin{cases} 1\ if\ x > 0 \\ 0\ otherwise \end{cases}$$

Elde edilen vektördeki değerlerin ikinci kullanıcının kullanım vektöründeki 1 olan indeksi en yakın olan eleman ile aralarındaki indeks farkı o aktivite için oluşan uzaklık değeri olarak kabul edilir. Bu şekilde A vektöründeki 1 den büyük her değer (her aktivite) için bir uzaklık çıkarımı yapılmaktadır. Bu şekilde, her aktivite için hesaplanan uzaklıkların medyanı birinci kullanıcının aktivitesinin ikinci kullanıcının aktivitesine olan uzaklığı olarak kabul edilmiştir.

İkinci aşamada aynı adımlar, ikinci kullanıcının birinci kullanıcıya olan uzaklığını bulmak için kullanılmıştır. Sonuç olarak, bu iki uzaklık değerinin ortalaması iki kullancının birbirine aktivite uzaklıkları olarak değerlendirilmiştir.

Burada, ikinci aşamayla devam edilmesinin nedeni farklı aktivite sayıları olması durumuda farklı değerlerin çıkmasıdır.

3.6. Arkadaşların Kıyaslanması

Arkadaş listeleri sosyal medya platformu tarafından gizlendiği durumda platformdaki etkileşim verilerinden yola çıkılarak elde edildi. örneğin Facebook'ta bir kullanıcının gönderilerine yorum yazan ve ifade bırakan kişiler de bu kullanıcının arkadaş ağında olarak değerlendirilmiştir.

Farklı profillerin arkadaş ağındaki kişi sayısı farklı olduğundan WMD dekine benzer bir şekilde EMD tekniği kullanılarak bir kıyaslama yapılmıştır. EMD tekniğine girdi uyarlamak için her arkadaş ismi bir histogram olarak değerlendirilerek histogramlar arası uzaklıklar bir string edit distance tekniği olan Levenshtein Distance ile hesaplanmıştır. Sonuç olarak, oluşturulan arkadaşlık ağlarının benzerliği EMD ile elde edilmektedir.

4. DENEYSEL ÇALIŞMALAR

4.1. Deney 1: Farklı Platform Kullanıcılarının Benzerlikleri Tespit Edilerek Aynı Kişiye ait Sosyal Medya Profillerinin Tespit Edilmesi

Çıkardığımız metriklerle farklı sosyal medya platformlarındaki kullanıcıların birbirlerine benzerliklerini elde etmek mümkün. Elimizdeki metrikleri değerlendirmek için aynı kişinin iki platformdaki hesaplarını eşleştirebilmesi elimizdeki bu metriklerin verimliliklerini ortaya koyacaktır. Bu nedenle bu deneyde iki sosyal medya platformundan profiller alarak elimizdeki metriklerle iki platformdaki aynı kişiye ait profillerin ne kadar eşleştirilebildiği denenmiştir.

Bu deneyde birkaç profil elde etmek için tek bir isim kullanarak hem Twitter'da hem de Facebook'ta arama yapıldı. Bu aramalar sonucunda listelenen ilk 10 profildeki bilgiler çekilmiştir. Bu profillerden 15 gönderinin altında paylaşımı olan profiller elendi ve kıyaslamalar sadece kalan profiller arasında yapılmıştır. Kıyaslamalar sadece farklı platform profilleri arasında yapılarak gerçekleştirilmiştir.

Data toplanmasındaki akış aşağıdaki gibidir:

"Mustafa Tuna" ismiyle yapılan bir arama sonucu hem Twitter'dan hem de Facebook'tan 10 ar kullanıcı profili tespit ediliyor. Ardından bu kullanıcı profillerinin bilgileri platformun sağladığı API'lar kullanılarak toplanır. Ardından herhangi bir profilde 15 ten az paylaşım varsa bu profil çıkarılıyor ve kalan profiller deney verisi olarak kullanılır.

Adlandırılmış varlıklar ve LDA ile çıkarılan kelimeler WMD için bir feature reduction hazırlığı olarak değerlendirilebilir. WMD $O(n^3)$ zamanda çalıştığından kelime sayısı arttıkça ciddi anlamda yavaşlayan bir algoritma olduğu için kelime sayısı fazla olduğunda sonuç alması uzun sürmektedir. Gönderilerdeki bütün kelimelerden WMD ile dokümanlar arasında benzerlik çıkarmak oldukça yavaş sonuç vereceğinden LDA, AVT

ve en fazla geçen kelimelerle her kullanıcı için bir doküman oluşturarak WMD ile bu dokümanlar arasındaki benzerlikler alındı. Bu benzerlikler ile oluşturulan bipartite çizgeler figürde incelenebilir.

Figürlerdeki düğümler kullanıcı profillerini, çizgiler iki profil arasındaki benzerliği göstermektedir. Benzerlik arttıkça çizgiler koyulaşmaktadır. Şekildeki çizimde benzerlikler çizgilerin renkleri ile ayrıştırılabilmesi için çizgedeki en yüksek benzerlik ağırlığına sahip çizgilerin %30'u renkli bir sekilde gösterilmiş az benzerlik içeren çizgiler silinmiştir.

Şekillerde kullanıcı isimlerinin hepsi arama sonucu elde edilen kişiler olduğundan çoğu aynı isme sahip birkaç tanesi benzer farklı isimlerdedir. Bu nedenle Facebook profillerini gösteren düğümler f_x, Twitter profillerini gösteren düğümler t_x formatında isimlendirilmiştir.

Kullanıcıların profil linkleri aşağıda verilmiştir.

- t_0: twitter.com/mustafatuna0606
- t_1: twitter.com/mustafatuna2014
- t_2: twitter.com/bat_mustafa
- t_3: twitter.com/Mtuna_alanya
- t_4: twitter.com/mustafatuna2206
- t_5: twitter.com/Mustafa32636798
- t_6: twitter.com/mtunaxx
- t_7: twitter.com/mustafaaaatuna
- t_8: twitter.com/mstftuna
- t_9: twitter.com/Mustafa92709286
- f_0: www.facebook.com/mtuna0606
- f_9: www.facebook.com/mustafa.tuna.946

- f_8: www.facebook.com/mustafatuna03

t_0 ve f_0 Ankara Büyükşehir Belediye Başkanı Mustafa Tuna'nın Twitter ve Facebook profillerine denk geliyor. t_1 Mustafa Tuna'nın destekçileri tarafından oluşturulmuş bir Twitter hesabıdır. Geri kalan hesaplar kişisel paylaşımlarda bulunan profillerdir.

4.1.1. Kullanıcının Kelimelerinden Çıkarılan Nitelikler

4.1.1.1. LDA Nitelikleri

LDA ile kullanıcıların gönderilerindeki konular modellendikten sonra konuları oluşturan kelimeler nitelik olarak kullanılmaktadır. Bu şekilde yapılan benzerlik kıyaslamasında Belediye Başkanı Mustafa Tuna'nın Twitter ve Facebook hesapları birbirlerine diğer kullanıcılara göre daha yakın çıktığı figürde gözlenebilir.

4.1.1.2. AVT Nitelikleri

Kullanıcıların gönderilerindeki adlandırılmış varlıkları çıkardıktan sonra çıkan bu varlıkların WMD ile kıyaslanması sonucu bir yakınlık ölçümü yapılmıştır. Bu ölçüm ile farklı sosyal medya platformlarındaki aynı isimle yapılan arama sonucu oluşan profil kümesi arasındaki benzerlik yukarıdaki figürde "AVT Benzerlikleri" başlığı altında verilmiştir.

Ankara Büyükşehir Belediye Başkanı Mustafa Tuna'nın Twitter ve Facebook profilleri arasındaki benzerliğin en fazla olduğu figürde de görülmektedir. Bu deney bize profillerdeki Adlandırılmış Varlıklar arasındaki benzerliğin aynı kişiye ait farklı sosyal medya platformlarındaki hesapların tespiti için kullanılabileceğini göstermektedir.

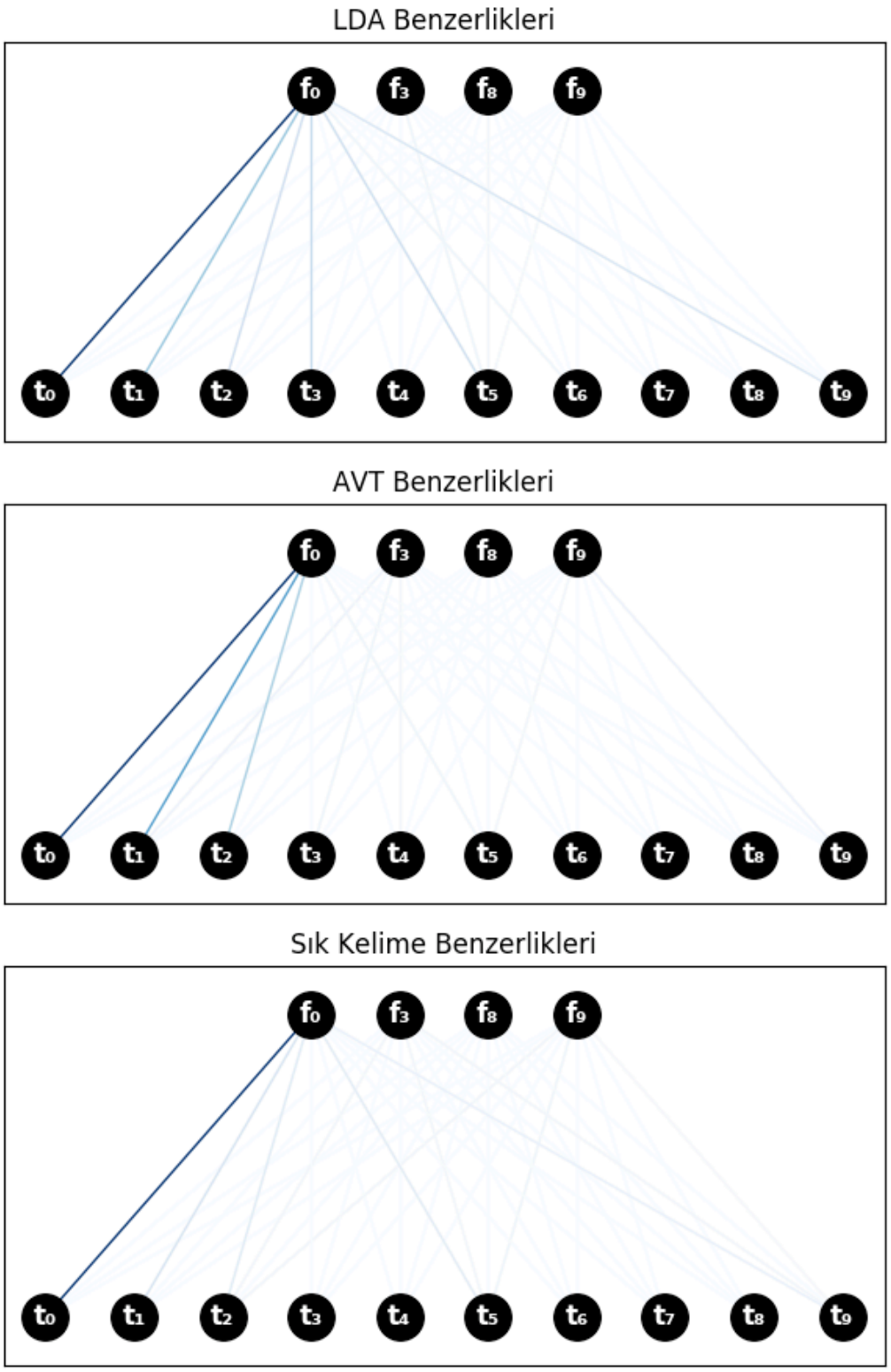
4.1.1.3. En Sık Kullanılan Kelimeler

WMD ile bir kullanıcının bütün gönderilerini kullanarak sonuç elde etmek WMD'nin $O(n^3)$ zamanda çalışması nedeni ile pek pratik olmamaktadır. Bu nedenle kullanıcının paylaşımlarında en fazla kullandığı 150 kelime ile bir kullanıcının nitelikleri çıkarılarak bu niteliklerle kullanıcılar arasındaki benzerlik elde edilmiştir. WMD'yi bu şekilde

kullanarak elde edilen yakınlık kıyaslamasını figürde “Sık Kelime Benzerlikleri” başlığı altındaki bipartite çizgede görebilirsiniz.

En sık kullanılan kelimeler ile de hem AVT hem de LDA da olduğu gibi en yakın çıkan kullanıcı hesapları Ankara Büyükşehir Belediye Başkanı Mustafa Tuna kişisinin Twitter ve Facebook hesapları olmuştur.

Burada basit bir nitelik çıkarımı ile AVT ve LDA gibi karmaşık sayılabilecek nitelik çıkarımı tekniklerine benzer sonuç elde edildiği görülse de AVT ve LDA kelime sayısı anlamında az geçebilecek kelimeleri de nitelik olarak seçebilmesi için farklı durumlarda etkili niteliklerin kaçırılmasına neden olabilir. En sık geçen kelimeler olarak çıkarılan nitelikler bir kullanıcının az da olsa bahsettiği yer ve kişi isimlerini veya konuları oluşturan kelimeleri kaçırabilir. Bu nedenle bu nitelik çıkarımları birbirlerini tamamlayıcı nitelikler olarak kullanılması daha doğru karar verilmesine yardımcı olacaktır.



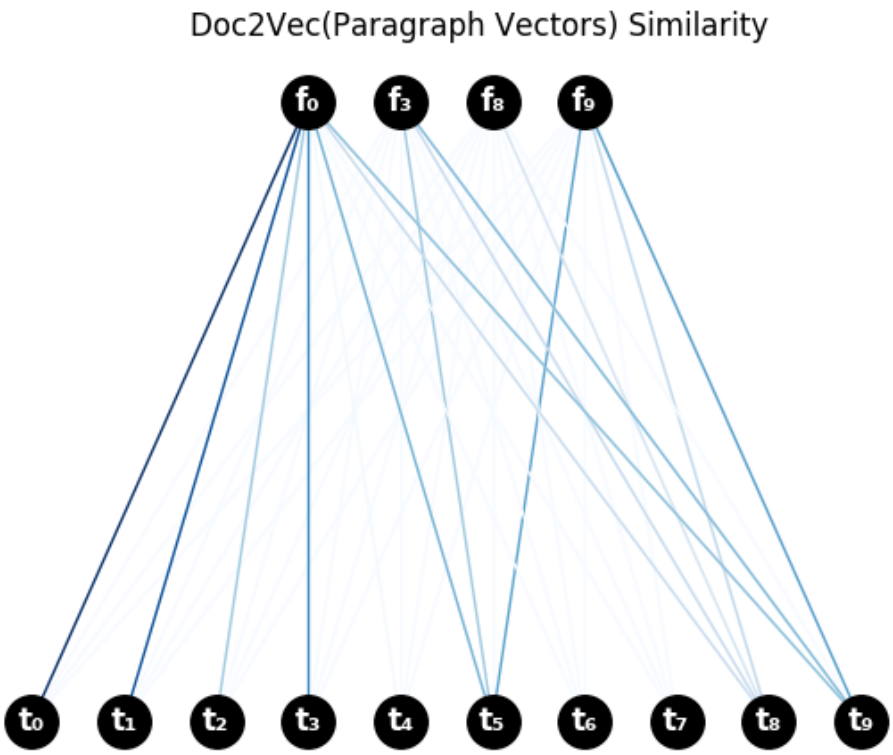
Şekil 4.1: Farklı platformlardaki kullanıcı hesaplarının LDA, AVT ve Sık Kelimeler ile elde edilen benzerliklerin gösterimi

4.1.2. Paragraph2Vec ile Elde Edilen Sonuçlar

Paragraf Vektör uygulaması olarak gensim kütüphanesinin Doc2Vec gerçeklemesi kullanıldı. Doc2Vec normalde her dokümana başka bir ID verilerek oluşturulmaktadır. Ancak paragraf vektörlerinde daha önce de belirtildiği her doküman ID'si ekstra bir kelime gibi YSA'ya girdi olarak veriliyor. Doküman IDsi aynı dokümandan oluşturulan içerik için tekrarlı bir şekilde YSA'ya girdi olarak veriliyor. Aynı mantıkla her kullanıcıya bir ID verilip bu ID'ler de doküman ID'leri gibi YSA'ya girdi olarak verildiğinde kullanıcılar için de bir vektör oluşturulabilmektedir Burada her kullanıcının gönderileri ile birlikte kullanıcı ID'si de YSA'ya verilerek her kullanıcı için 50 boyutlu bir vektör oluşturuldu. Bu şekilde kullanıcılar arasındaki benzerlik bu vektörlerin uzaklıklarına göre elde edilmiştir.

Burada büyük bir korpus oluşturulmadan elde edilen sonucu şekilde görebilirsiniz. Korpus oluşturulurken, bir kullanıcı ismiyle aratılıp elde edilen Facebook ve Twitter hesaplarında geçen adlandırılmış varlıklardan en fazla geçen 600 tanesi alınmıştır. Ardından Twitter üzerinde bu kelimelerle arama yapılarak her adlandırılmış varlık için 200 civarı Tweet çekilmiştir. Bu şekilde 12000 civarı Tweet ve kıyaslama yapılacak kullanıcıların gönderileri ile Doc2Vec için bir korpus oluşturulmuştur.

Doc2Vec ile elde edilen sonuçlarda kullanıcıya benzer diğer kullanıcılarda benzer aday profiller birbirlerine çok yakın çıkabiliyor. Kelimelerden elde edilen niteliklerle yapılan karşılaştırmalarda benzer aday kullanıcılardan öne çıkan kullanıcı profili diğerlerine göre daha büyük benzerliğe sahip olduğu görülüyordu. Ancak Doc2Vec'te bu fark anlamlı bir şekilde elde deilememiştir. Yine de Mustafa Tuna kullanıcıları arasında Belediye Başkanı Mustafa Tuna'nın Twitter ve Facebook profilleri arasında benzerlik tespit edilmiştir. Doc2Vec daha büyük bir korpusta daha belirgin sonuçlar çıkarabilecek bir yaklaşım olabileceği değerlendirilmiştir.



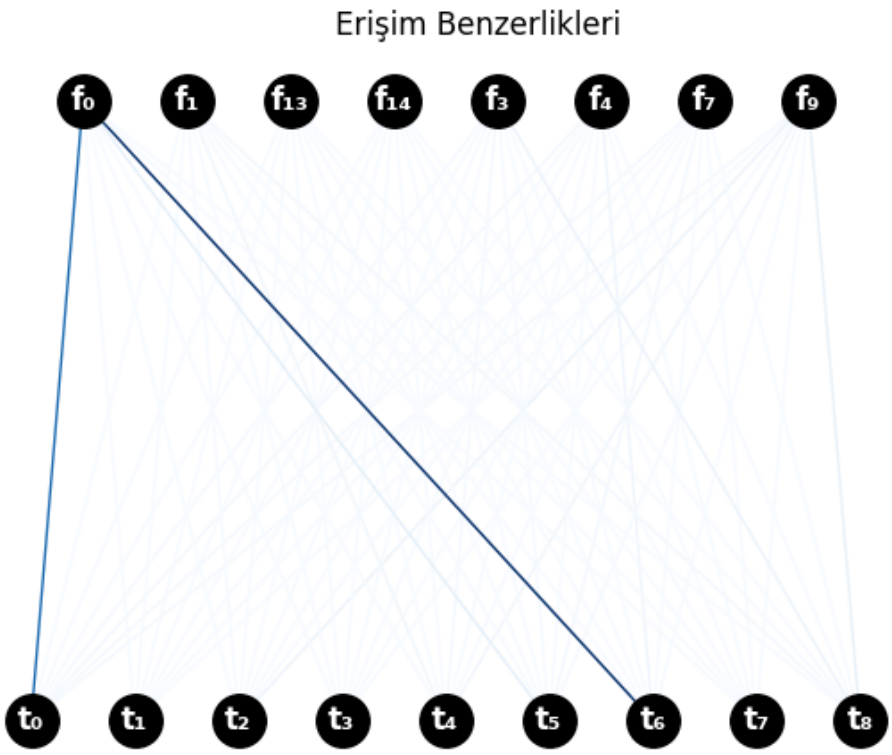
Şekil 4.2: Farklı platformlardaki kullanıcı hesaplarının Doc2Vec benzerliklerinin gösterimi

4.1.3. Kullanıcı Aktivitesi ile Elde Edilen Sonuçlar

Bir kullanıcının bir olayla alakalı yaptığı paylaşımlarda diğer sosyal medya hesaplarında da benzer zamanlarda paylaşım yapacağı göz önüne alınarak kullanıcıların benzerliklerini tespit etmek için geliştirilen teknik 3.5. bölümde açıklanmıştır. Bu metrikle kullanıcı birbirine yakın saatlerde paylaşım yaptığı takdirde aynı kişiye ait hesapların tespit edilebilmesi denenmiştir.

Aşağıdaki şekilde Mustafa Tuna ile elde edilen profillerin benzerlikleri verilmiştir. Şekilde de görüldüğü gibi bu metrik Belediye Başkanı Mustafa Tuna'nın Facebook ve Twitter profilleri arasında diğer birçok profile göre belirgin bir benzerlik yakalayabilmiştir. Ancak Belediye Başkanı Mustafa Tuna'nın "twitter.com/mtunaxx" kullanıcısı ile aktivite benzerliği kendi Twitter profiline göre daha fazladır. Yine de diğer

birçok kullanıcı arasından aynı kişiye ait belirgin bir benzerlik elde edebilmesi bakımından bu metriğin de faydalı olabileceği durumlar olacaktır.



Şekil 4.3: Kullanıcı aktivitesi ile elde edilen benzerlikler

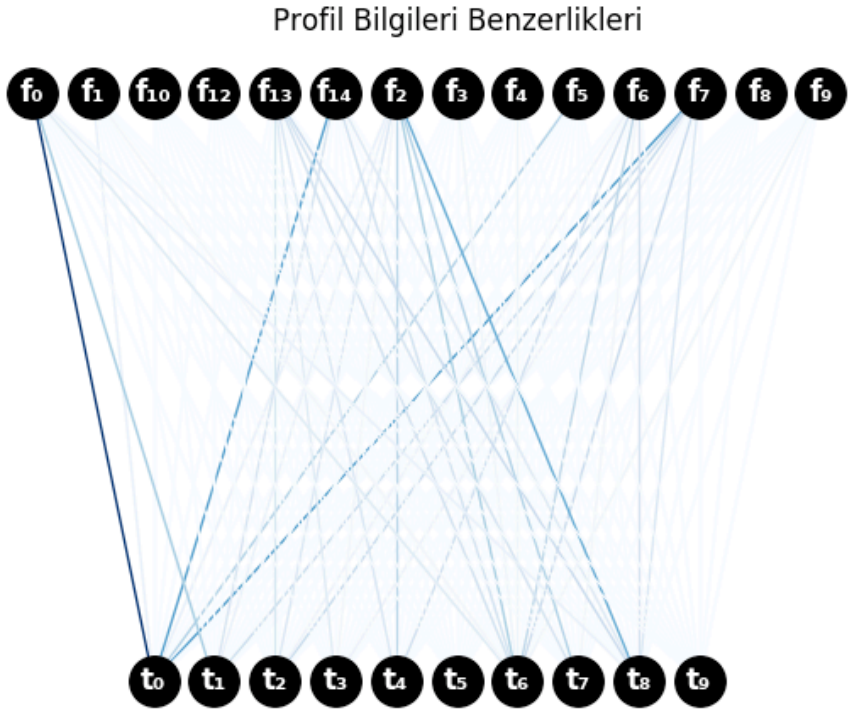
4.1.4. Kullanıcı Arkadaşları ile Elde Edilen Sonuçlar

Kullanıcıların arkadaş listelerindeki ortak isimlere göre iki profilin aynı kişiye ait olabileceği fikri ile bu deney gerçekleştirilmiştir. Burada kullanıcıların arkadaş listeleri WMD'dekine benzer bir şekilde EMD yaklaşımı ile kıyaslanarak kullanıcıların birbirine yakınlığı tespit edilmiştir. WMD bir word embedding modeline göre kelimeler arasında kıyaslama yaparken bu teknikte kullanıcı isimleri Levenshtein Distance alınarak birbirlerine benzerlikleri değerlendirilmiştir.

4.1.5. Kullanıcı Profil Bilgileri ile Elde Edilen Sonuçlar

SMP'lerde kullanıcıların profil bilgileri genellikle kişinin yaşadığı şehir, çalıştığı kurum gibi kısa biyografik bilgilerden oluşmaktadır. Ancak çoğu zaman bu bilgi düzensiz bir şekilde elde edilebilmektedir. Herhangi bir SMP'de çalışılan bir yer verildiğinde diğer SMP'de bu bilgi saglanmamış olabilir. Bu deneyde kullanıcıların profil bilgileri toplandıktan sonra profil bilgilerini oluşturan kelimeler arasında Earth Mover's Distance(EMD) tekniği kullanılarak bir benzerlik elde edilmiştir. EMD için kelimeler arası uzaklık temsil edilmesi gerekmektedir. Bunun için Levenshtein uzaklığı kullanılmıştır.

EMD bilgilerin düzensiz olmasından etkilenmeden birbirine yakın kelimeler arasında akışı oluşturarak sonuca ulaşabilmektedir. Bu şekilde kullanıcı profil bilgilerini oluşturulan kelimeler arasında bir ilişki kurmaya gerek kalmadan sonuç alınabilmektedir. Bu benzerlik metriği ile de Belediye Başkanı Mustafa Tuna'ya ait profillerin en benzer profiller olduğunu aşağıdaki şekilde görebilirsiniz.



Şekil 4.4: Profil bilgileri ile elde edilen benzerlik sonuçları

4.2. Deney 2: Sosyal Medya Profillerinin Doğal Dil İşleme Teknikleri ile Hiyerarşik Kümelenmesi

Gözetimli makine öğrenmesi tekniklerinin performanslarını ölçmek gözetimsiz tekniklerin performansını ölçmeye göre eğitim verilerinin etiketleri olduğu için göreceli olarak daha kolay bir iştir. Veri noktalarının etiketi kullanılarak kesin yargılara ulaşılabilecek metrikler geliştirilebilir. Ancak eğitim verisinde etiket olmadığı durumda modellerin performansını değerlendirmek pek kolay bir görev olmamaktadır. Bu bakımdan daha önceden farklılıkları ve yakınlıkları arasında bir fikre sahip olduğumuz veri noktalarının, geliştirilen DDİ ölçütleri kullanılarak kümelenmesi, geliştirilen DDİ ölçütlerinin etkinliği hakkında fikir verecektir.

Bu fikirden yola çıkarak yeni bir veri seti oluşturulmuştur. Veri seti hazırlanırken mantıksal olarak ayrı gruplardan kişilerin Twitter ve Facebook hesaplarındaki gönderiler toplanmıştır. Tablo 4.1'de kullanılan veri seti hakkında bilgi verilmiştir. Bu kullanıcıların 1 Haziran 2018 tarihinden 31 Temmuz 2018 tarihine kadar gönderileri kullanılarak bu veri seti oluşturulmuştur.

Veri setinin kümelenmesi, hiyerarşik kümeleme yöntemiyle gerçekleştirildi. Öbeklerin oluştuğu adımları anlaşılır biçimde takip etmemize ve görselleştirmemize olanak sağladığı için hiyerarşik öbekleme seçildi. Metrikler çalışıyorsa siyasetçileri, gazetecileri ve diğer ünlü isimleri kendi aralarında mantıklı öbeklere ayırmasını bekliyoruz. Ancak kimi ünlü isimler aynı gazeteci veya politikacılar gibi gündemle alakalı olaylar hakkındaki yorumlarını da paylaştıkları için direk olarak 3 öbekten oluşmasını beklemek yerine, kümelerin oluşma adımlarını incelemek ölçütlerimizin etkinliğini anlamada daha mantıklı bir yaklaşım olacaktır.

Tablo 4.1: Hiyerarşik kümelemede kullanılan profiller

Kategori	Kişi	Facebook Profili	Twitter Profili
Ünlü/Magazinsel Kişilik	Acun Ilıcalı	fb/acunilicali	tw/acunilicali
	Bengu	fb/benguofficial	tw/bengu
	Gülben Ergen	fb/GulbenErgenOfficial	tw/GulbenErgen
	Hadise	fb/Hadise	tw/Hadise
	Kenan Doğulu	fb/kenandogulu	tw/kenandogulu
	Mustafa Ceceli	fb/mustafaceceli	tw/mustafaceceli
	Sıla	fb/Sila.Gencoglu	tw/silagencoglu
Gazeteci/Yazar	Banu Güven	fb/banuguven	tw/banuguven
	Bekir Coşkun	fb/ BEKIRCOSKUNveYAZILARI	tw/cosknbekr2
	Can Ataklı	fb/AtakliCan	tw/can_atakli_
	Cüneyt Özdemir	fb/cuneyt.ozdemir.98	tw/cuneytozdemir
	Ergun Diler	fb/ergundileryazar	tw/ergn_diler
	Fehmi Koru	fb/ fehmikorunungunlugu	tw/fkoru
	Hasan Cemal	fb/hsncml	tw/hsncml
	Nazlı Çelik	fb/nazlicelikofficial	tw/nazlicelik_
Siyasetçi	Kemal Kılıçdaroğlu	fb/K.Kilicdaroglu	tw/kilicdarogluk
	Koray Aydın	fb/korayaydintr	tw/korayaydintr
	Mahir Ünal	fb/MahirUnal46	tw/mahirunal
	Meral Akşener	fb/meral.aksener.9	tw/meral_aksener
	Muharrem İnce	fb/muharrem.ince77	tw/vekilince
	Mustafa Tuna	fb/mtuna0606	tw/mustafatuna0606
	Recep Tayyip Erdoğan	fb/ RecepTayyipErdogan	tw/RT_Erdogan
	Süleyman Soylu	fb/suleymansoylu	tw/suleymansoylu
	Tuncay Özkan	fb/ATuncayOzkan	tw/ATuncayOzkan

Hiyerarşik kümelemede en yakın kümelerden başlanarak kümeler birleştirilerek bir ağaç yapısı oluşturulur. Oluşturulan ağaçtan kümelerin çıkarılması, bir kesit alınarak dalların birbirlerinden ayrılması ile elde edilir. Bu kesitin alınacağı nokta genellikle bu kümeler arasındaki uzaklıkların ortalamasının %70'ine kadar olan kısımdır, bu çalışmada da bu oran kullanılmıştır. Bu oran değiştirilerek kümelerin ayrımı daha detaylı veya daha genel bir hale getirilebilir. Kümeler hiyerarşi ağacının kökünden yapraklarına doğru renklendirilerek gösterilmiştir.

Deneyler SciPy [22] kütüphanesindeki 'hierarchical clustering' uygulaması kullanılarak gerçekleştirilmiştir. Kümelerin ağaç şekilleri için matplotlib [23] kütüphanesinden yararlanılmıştır.

4.2.1. LDA Nitelikleri ile Kümeleme

LDA ile konuları oluşturan kelimelerin çıkarımı yapıldıktan sonra elde edilen konu kelimelerinin birbirlerine yakınlıkları ile oluşan öbekleri aşağıdaki tabloda ve şekilde görebilirsiniz.

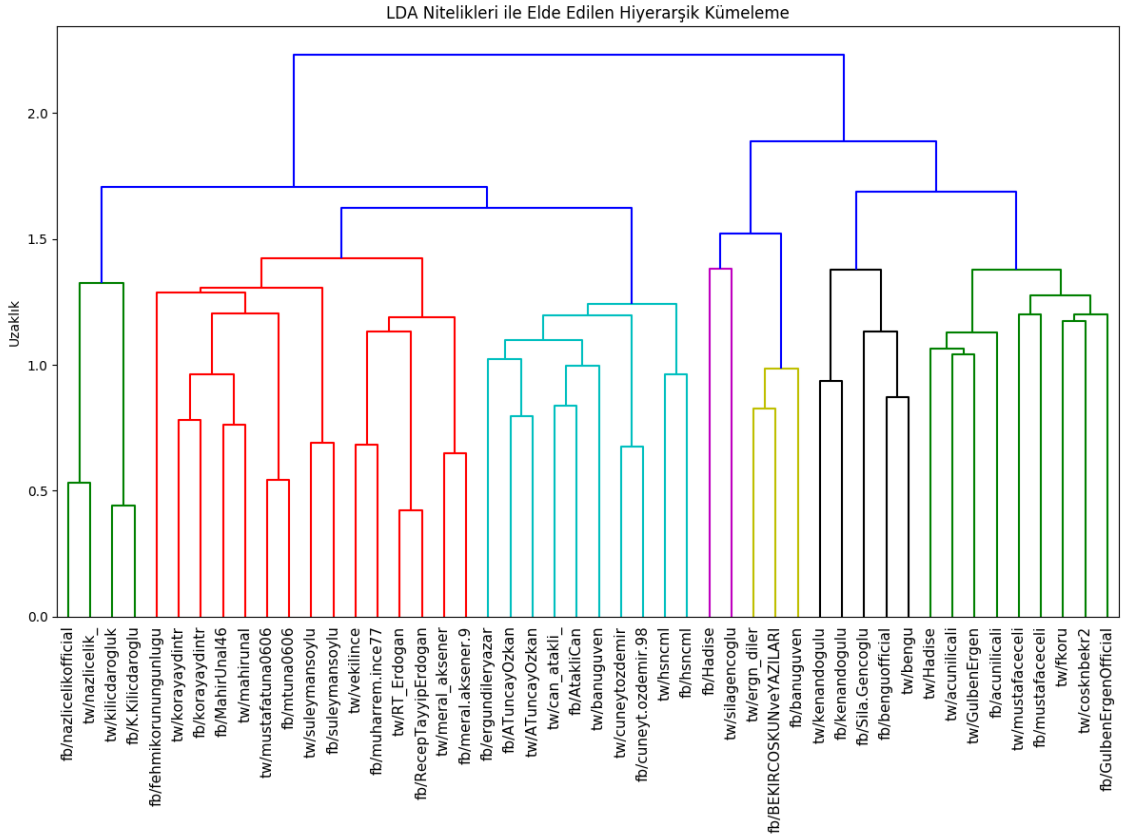
Tablo 4.2'ye bakıldığında 1. ve 2. kümeler dışındaki kümelerin, bir gruba ait kişilerden oluştuğu görülmektedir. 1. küme gazeteci Nazlı Çelik ve siyasetçi Kemal Kılıçdaroğlu'nun Twitter ve Facebook hesaplarından oluşmaktadır. 2. kümedeyse gazeteci Fehmi Koru dışındaki bütün kullanıcı hesapları siyasilere aittir. 3. Kümede de benzer şekilde siyasi bir kişilik olan Tuncay Özkan'ın profilleri gazetecilerin olduğu bir kümede yer almıştır. Geri kalan 4 kümede farklı yazarlar ve ünlü kişiler kendi aralarında kümelere ayrılmıştır. Gazetecilerin ve siyasilerin genellikle siyasi gündem hakkında paylaşımlarda bulunmaları göz önüne alındığında 1., 2. ve 3. kümedeki durumun oluşması kabul edilebilir bir durum olarak görülebilir. Sonuç olarak LDA ile elde edilen niteliklerle mantıklı bir kümeleme işlemi gerçekleştirildiği söylenebilir.

Hiyerarşik kümeleme ile birleşen iki elemanlı ilk kümelerin 16 tanesi aynı kişiye ait sosyal medya profillerinden oluşmaktadır. Veri setimiz küçük de olsa bu gözlemden yola

çıkarak aynı kişiye ait farklı sosyal medya profillerini belirlemede %66.6'lık bir doğruluk payı elde ettiğimizi söyleyebiliriz.

Tablo 4.2: LDA ile elde edilen nitelikler ile oluşan öbekler

Küme 1	fb/nazlicelikofficial, tw/nazlicelik_, fb/K.Kilicdaroglu, tw/kilicdarogluk
Küme 2	fb/fehmikorunungunlugu, fb/korayaydintr, tw/korayaydintr, fb/MahirUnal46, tw/mahirunal, fb/mtuna0606', tw/mustafatuna0606, fb/suleymansoylu, tw/suleymansoylu, fb/ muharrem.ince77, tw/vekilince, fb/RecepTayyipErdogan, tw/RT_Erdogan, fb/meral.aksener. 9, tw/meral_aksener
Küme 3	fb/ergundileryazar, fb/ATuncayOzkan, tw/ATuncayOzkan, fb/AtakliCan, tw/can_atakli_, tw/ banuguven, fb/cuneyt.ozdemir.98, tw/cuneytozdemir, fb/hsncml, tw/hsncml
Küme 4	fb/Hadise, tw/silagencoglu
Küme 5	tw/ergn_diler, fb/BEKIRCOSKUNveYAZILARI, fb/banuguven
Küme 6	tw/kenandogulu, fb/kenandogulu, fb/Sila.Gencoglu, fb/benguofficial, tw/bengu
Küme 7	tw/Hadise, tw/acunilicali, tw/GulbenErgen, fb/acunilical, tw/mustafaceceli, fb/mustafaceceli, tw/fkoru, tw/coskunbekr2, fb/GulbenErgenOfficial



Şekil 4.5: LDA ile elde edilen nitelikler ile oluşan kümeler

4.2.2. AVT Nitekileri ile Kümeleme

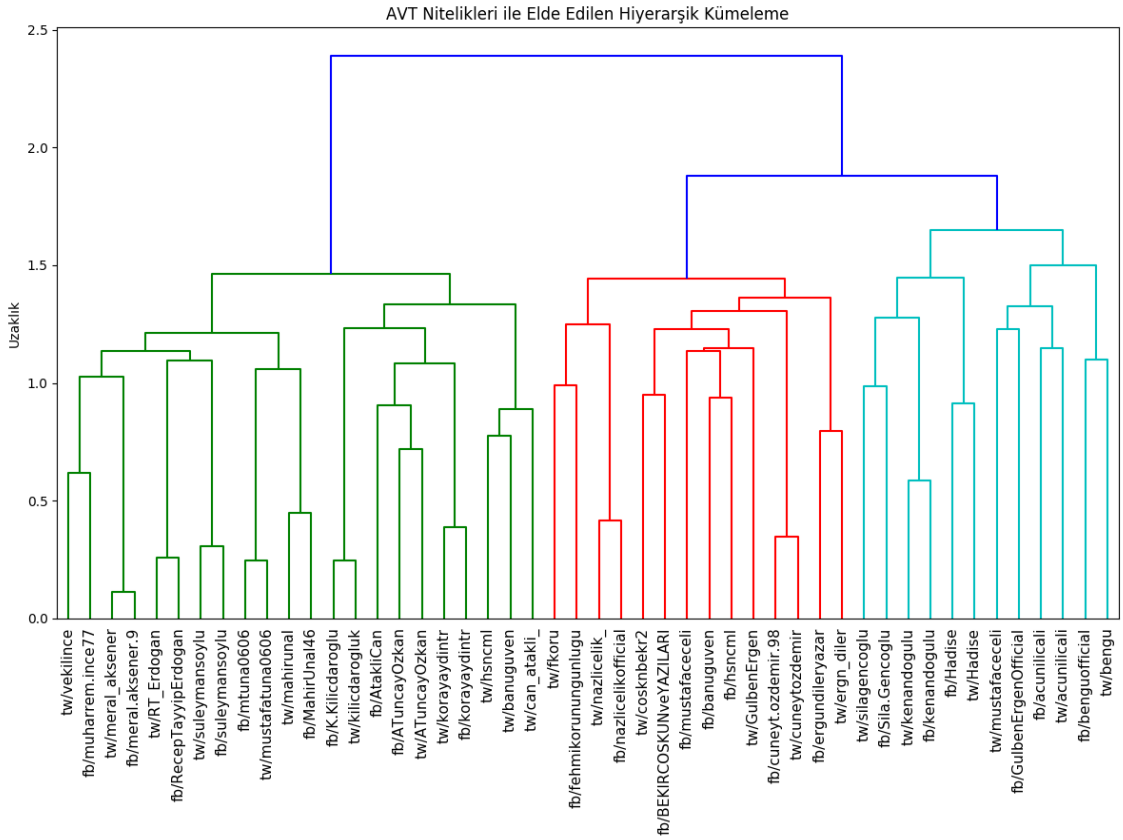
AVT ile adlandırılmış varlıkların çıkarımı sonucu, varlıkların birbirlerine yakınlıkları kullanılarak elde edilen öbeklerin hiyerarşik öbeklemesi şekil 4.6'da ve oluşan öbekler tablo 4.3'te verilmiştir.

Tablo 4.3'e bakıldığında 1. kümedeki kullanıcı hesaplarının 22 sinden 4'ü gazetecilere 18'i siyasi kişilere ait olduğu görülmektedir. 2. Kümedeki kullanıcılarda Mustafa Ceceli ve Gülben Ergen ünlü kişiliklerinin Twitter hesapları dışındaki hepsi gazetecilerden oluşmaktadır. 3. Kümedeki kullanıcıların tamamı ünlü kişilerin Facebook ve Twitter hesaplarından oluşmaktadır. 1. kümede gazetecilerden bazılarının siyasilerle aynı kümede olması bakımından LDA ile elde edilen kümelemeye benzer bir sonuç bulunmaktadır. 2. kümede ise Gülben Ergen ve Mustafa Ceceli Twitter kullanıcılarını gazetecilerle aynı kümede yer almıştır.

Hiyerarşik kümelemede oluşan iki elemanlı ilk kümelerin 19'u aynı kişiye ait sosyal medya profillerinden oluşmuştur. Yani elimizdeki veri seti ile aynı kişiye ait sosyal medya platformlarını %79.1'lik bir doğruluk payı ile elde edilmektedir. LDA ile elde edilen niteliklerde bu oranın %66.6 olduğu düşünüldüğünde AVT ile çıkarılan özniteliklerin aynı kişiye ait sosyal medya profillerini tespit etmede LDA ile elde edilen sonuçlara göre daha başarılı olduğu söylenebilir.

Tablo 4.3: AVT ile elde edilen nitelikler ile oluşan öbekler

Küme 1	tw/vekilince, fb/muharrem.ince77, , tw/meral_aksener, fb/meral.aksener.9, tw/RT_Erdogan, fb/RecepTayyipErdogan, tw/suleymansoylu, fb/suleymansoylu, fb/mtuna0606, fb/mustafatuna0606, tw/mahirunal, fb/MahirUnal46, fb/K.Kilicdaroglu, tw/kilicdarogluk, fb/AtakliCan, fb/ATuncayOzkan, tw/ATuncayOzkan, tw/korayaydintr, fb/korayaydintr, tw/hsncml, tw/banuguven, tw/can_atakli_
Küme 2	tw/fkoru, fb/fehmikorunungunlugu, tw/nazlicelik_, fb/nazlicelikofficial, tw/cosknbekr2, fb/BEKIRCOSKUNveYAZILARI, fb/mustafaceceli, fb/banuguven, fb/hsncml, tw/GulbenErgen, fb/cuneyt.ozdemir.98, tw/cuneytozdemir, fb/ergundileryazar, tw/ergn_diler
Küme 3	tw/silagencoglu, fb/Sila.Gencoglu, tw/kenandogulu, fb/kenandogulu, fb/Hadise, tw/Hadise, tw/mustafaceceli, fb/mustafaceceli, fb/GulbenErgenOfficial, fb/acunilicali, tw/acunilicali, fb/benguofficial, tw/bengu



Şekil 4.6: AVT ile elde edilen nitelikler ile oluşan kümeler

4.2.3. En Sık Kullanılan 180 Kelime ile Kümeleme

Kullanıcıların gönderilerinde en sık kullandıkları 180 kelime nitelik olarak kullanıldığında oluşan öbekleri Şekil 4.6 ve Tablo 4.4'de gösterilmiştir.

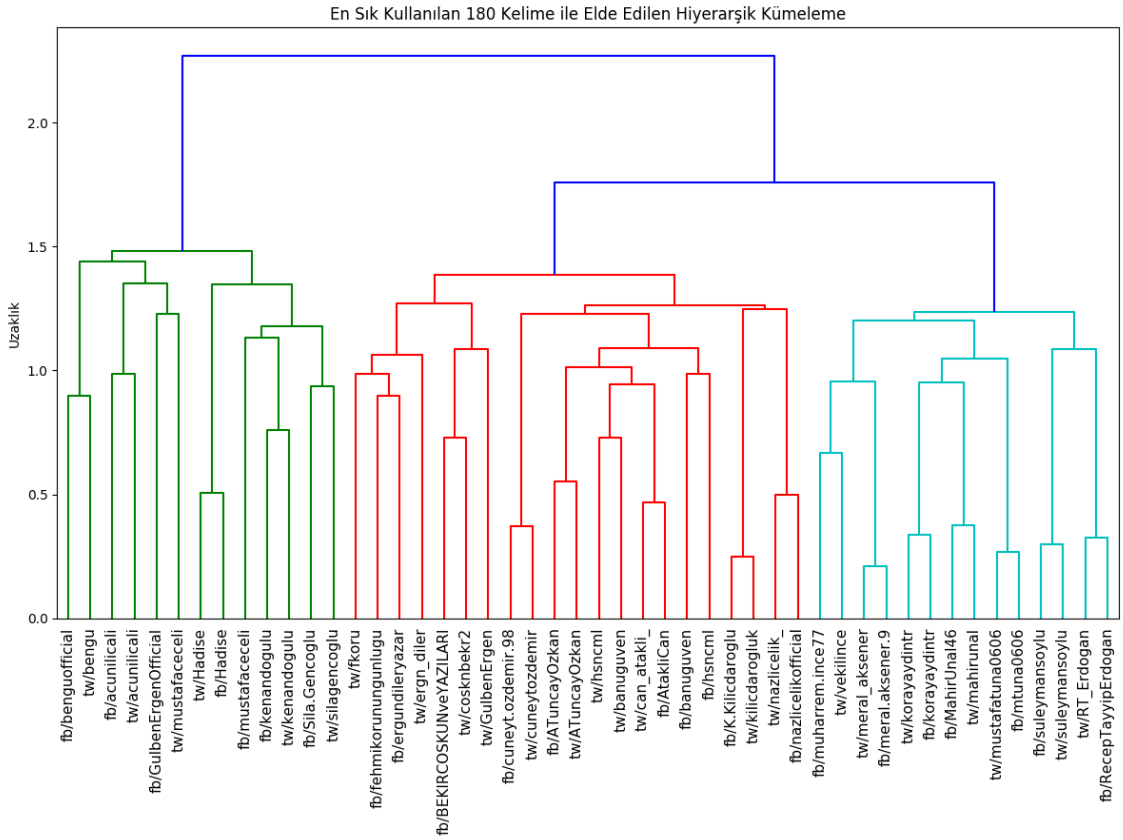
Tablo 4.4'te görüldüğü gibi 1. küme tamamen ünlü kişilere ait hesaplardan oluşmaktadır. 2. küme Tuncay Özkan, Kemal Kılıçdaroğlu siyasilerinin Twitter ve Facebook hesapları ve Gülben Ergen Twitter kullanıcısı dışında gazetecilerden oluşmaktadır. Gülben Ergen Twitter kullanıcısının AVT nitelikleri ile elde edilen kümelemede olduğu gibi burada da gazetecilerle aynı kümede çıkması Gülben Ergen'in gündem ile alakalı paylaşımlarda bulunmasından kaynaklandığını teyit etmektedir. Son küme tamamen siyasilerin Twitter ve Facebook hesaplarından oluşmaktadır. En sık kelimelerin nitelik olarak kullanılması ile elde edilen kümeler de 1. ve 3. kümeler için %100 doğrulukta 2. küme için %76.1 doğrulukta bir kümeleme sonucu verdiği söylenebilir. Ancak daha önce de belirtildiği

gibi gazeteci ve siyasilerin benzer konulardan bahsetmeleri 2. kümelemede ortaya çıkan durumun sebebi olarak değerlendirebilir.

Diğer niteliklerle elde edilen kümelemelerde yaptığımız hesabı yapacak olursak en sık kullanılan kelimeler ile oluşan iki elemanlı ilk kümelerin aynı kişiye ait sosyal medya profillerinden oluşma sayısı 18 olmuştur. Yani aynı kişiye ait hesapları tespit etmedeki doğruluk payı bu niteliklerle %75 olarak gerçekleşmiştir.

Tablo 4.4: En sık kullanılan 180 kelime ile oluşan öbekler

Öbek 1	fb/benguofficial, tw/bengu, fb/acunilicali, tw/acunilicali, fb/GulbenErgenOfficial, tw/mustafaceceli, tw/Hadise, fb/Hadise, fb/mustafaceceli, fb/kenandogulu, tw/kenandogulu, fb/Sila.Gencoglu, tw/silagencoglu
Öbek 2	tw/fkoru, fb/fehmikorunungunlugu, fb/ergundileryazar, tw/ergn_diler, fb/BEKIRCOSKUNveYAZILARI, tw/cosknbekr2, tw/GulbenErgen, fb/cuneyt.ozdemir.98, tw/cuneytozdemir, fb/ATuncayOzkan, tw/ATuncayOzkan, tw/hsncml, tw/banuguven, tw/can_atakli_, fb/AtakliCan, fb/banuguven, fb/hsncml, fb/K.Kilicdaroglu, tw/kilicdarogluk, tw/nazlicelik_, fb/nazlicelikofficial_
Öbek 3	fb/muharremince77, tw/vekilince, tw/meral_aksener, fb/meral.aksener.9, tw/korayaydintr, fb/korayaydintr, fb/MahirUnal46, tw/mahirunal, tw/mustafatuna0606, fb/mtuna0606, fb/suleymansoylu, tw/suleymansoylu, tw/RT_Erdogan, fbRecepTayyipErdogan



Şekil 4.7: En sık kullanılan 180 kelime ile oluşan öbekler

Tablo 4.[2,3,4]'e bakıldığında siyasilerin genellikle aynı öbeklerde, gazetecilerin kendi öbeklerinde veya siyasilerle aynı öbekte, magazinsel kişilerin de kendi içlerinde bir öbek oluşturduğu görülebilir. Bu sonuçlar bize geliştirdiğimiz kıyaslama ölçütlerimizin mantıklı sonuçlar verdiğini teyit etmektedir.

Basitçe en sık kullanılan kelimeler ile de mantıklı sonuçlar alınabilse de AVT ve LDA teknikleri ile kelime sayısı bakımından az geçebilecek kelimeleri nitelik olarak kullanabilmektedir. Bu nedenle farklı durumlarda sadece en sık geçen kelimelerin nitelik olarak kullanılması etkili kelimelerin kaçırılmasına neden olabilir. Bu nitelik çıkarımı tekniklerinin birbirini tamamlayıcı olarak değerlendirilmesi analizler esnasında daha faydalı olacaktır.

5. PROFİL BENZERLİK GÖSTERİM ARACI

Geliştirilen metriklerin farklı kullanıcı verileri ile daha kolay sonuç alınması amacıyla bir web uygulaması geliştirildi. Bu uygulama sunucu tarafında benzerlik metrikleri ile sonuçları alarak bu sonucu web arayüzünde çizge yapısında gösterilmektedir. Çizgedeki düğümler kullanıcıları düğümler arasındaki çizgiler kullanıcılar arasındaki benzerliği belirtmektedir. Benzerlikler çizgilere ağırlık olarak atanmıştır. Bu şekilde benzer kullanıcılar arasındaki çizgiler daha koyu renkle gösterilmektedir. Bu gösterim aracında çizgelerin oluşturulmasında networkx[24] kütüphanesi kullanılmış ve çizgelerin gösterimi için D3.js kütüphanesinden[25] yararlanılmıştır.

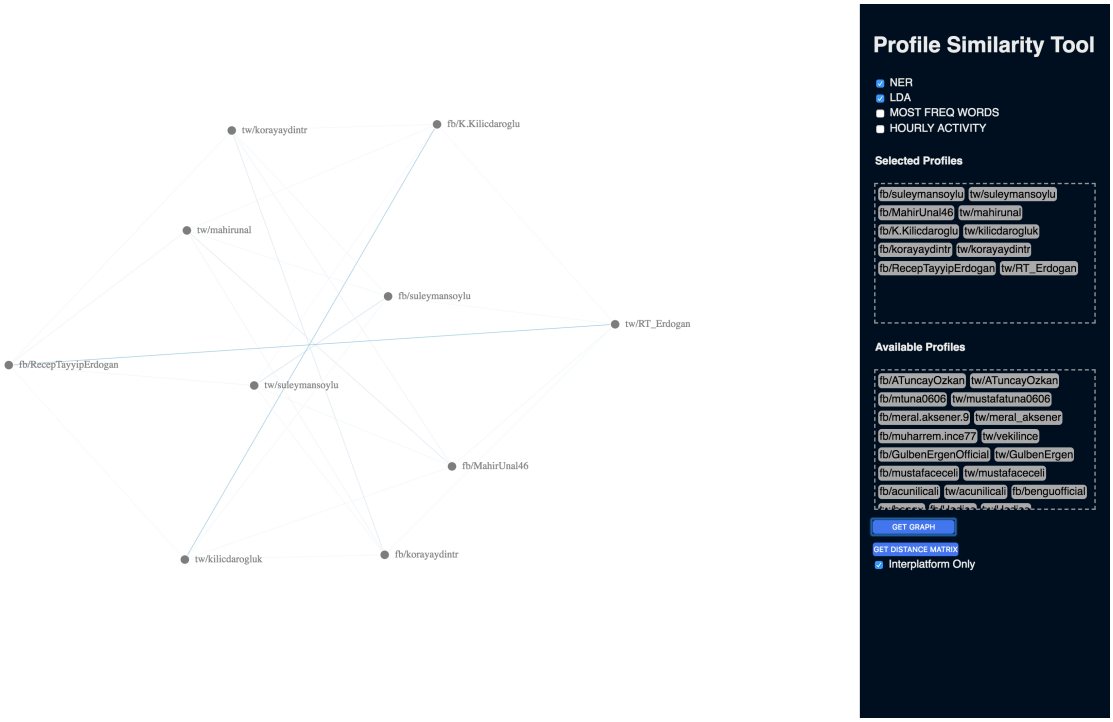
Aşağıdaki resimde uygulamaya ait bir ekran alıntısını görebilirsiniz. Uygulama daha önceden verileri çekilmiş bir kullanıcı kümesinden kullanıcılar seçilmesine olanak sağlamaktadır. Kullanıcılar seçildikten sonra "GET GRAPH" tuşuna basılarak bu kullanıcılar arasındaki benzerliklerin gösterildiği bir çizge ekranın sol tarafında kullanıcıya sunulmaktadır. Kullanıcının bu arayüz ile farklı kullanıcı grupları için farklı metrikleri deneyebilmesi ve analiz yapması kolaylaşacaktır.

Sunulan metriklerin ayrı ayrı sonuçları da elde edilebileceği gibi kullanıcı seçilen profiller için farklı metriklerin bu araç sayesinde birleştirilmiş bir sonucunu alarak analizler yapabilir. Örneğin, NER(AVT) ve LDA birlikte seçildiği durumda bu teknikler ile elde edilen benzerliklerin ortalaması alınarak seçilen kişiler arasındaki benzerlikler değerlendirilebilmektedir.

"Interplatform Only" seçeneği ile sadece Facebook ve Twitter profilleri arasındaki benzerlikler alınabilmekte bu seçenek boş bırakıldığında bütün seçilen kişiler arasındaki benzerlik oluşturularak ekranda gösterilmektedir.

"Get Graph" ile kullanıcılar arasındaki benzerlik gösteriminin alındığı çizgeyi aşağıdaki Resim 5.1'de görebilirsiniz. Bir kişinin farklı platformlardaki hesapları arasındaki benzerlik diğer düğümlere göre daha fazla olduğu resimde görülebilmektedir.

“Get Distance Matrix” tuşuna basıldığında kullanıcı seçtiği profiller arasındaki benzerlikleri bir CSV dosyası formatında alabilmektedir. Bu şekilde kullanıcının bu metriklerle oluşan uzaklık matrisi ile kendi analizlerini yapabilmesine olanak sağlanmıştır.



Resim 5.1: Gösterim aracında kişilerin farklı platformlardaki hesaplarının eşleşmesi

6. SONUÇ

Bu tez çalışmasında sosyal medya kullanıcıları arasındaki benzerliklerin tespit edilmesi amacıyla geliştirilen teknikler ve yapılan deneylerle bu tekniklerin etkinlikleri sunuldu. Nitelik çıkarmak için teknikler geliştirilmiştir.

Deneysel çalışmalar bölümünde iki farklı şekilde veri setleri oluşturularak geliştirilen teknikler denenmiştir. Birinci deneyde bir kullanıcı ismi için yapılan arama ile elde edilen profilleri arasındaki benzerliklerin tespiti ile aynı kişiye ait farklı sosyal medya platformlarındaki profillerin tespit edilmesi denenmiştir. Geliştirilen tekniklerden AVT, LDA, sık kullanılan kelimeler ve profil bilgileri ile anlamlı ayırımlar yapılabileceği bu deneyde gösterilmiştir. Paragraph2Vec eğitim verisinin fazla olmaması nedeniyle iyi sonuçlar verememiştir. Kullanıcıların erişim benzerlikleri ile de kesin bir ayrıma varmak mümkün olmamıştır.

Bu tekniklerden DDİ teknikleri farklı bir senaryoda ikinci bir veri seti ile denenerek performansları değerlendirilmiştir. İkinci deneydeki veri seti farklı topluluklar oluşturulacağı düşünülen 9 siyasi, 8 gazeteci ve 7 ünlüden oluşmaktadır. Deneysel çalışmalar kısmındaki ikinci deneyde kullanılan göreceli olarak daha geniş veri setinde AVT, LDA ve en sık kelimelerden elde edilen özniteliklerin WMD ile etkinlikleri hem aynı kişiye ait farklı sosyal medya hesaplarının tespit edilmesinde hem de farklı tiplere ait kullanıcıların kümelenmesinde etkili sonuçlar vermiştir. Aynı kişiye ait farklı SMP'lerdeki profilleri tespit etmede AVT ile elde edilen nitelikler %79.1, en sık kelimeler kullanılarak elde edilen nitelikler %75, LDA ile elde edilen nitelikler %66.6'lık bir doğruluk payı yakalamıştır. DDİ teknikleri ile yapılan kümelemelerde bazı gazetecilerin veya siyasilerin azınlık olarak birbirlerinin kümelerinde yer almaları dışında hiyerarşik kümelemelerde geliştirilen metriklerin mantıklı kümelemeler yaptıkları gösterilmiştir. Siyasilerin ve gazetecilerin benzer konulardan bahsetmelerinin bu duruma neden olduğu söylenebilir.

Geliştirilen tekniklerin birkaç farklı kullanım durumunda etkinlikleri deneylerde gösterildi. Geliştirilen teknikler aynı kişinin farklı sosyal medya hesaplarının tespitinde kullanılabileceği gibi verilen bir kullanıcı topluluğundan kullanıcılar arasındaki benzerliklerin tespit edilmesi ile yapılan öbekleme sonucu topluluktaki öbekler belirlenebilmektedir.

Bunların dışında bu kıyaslama tekniklerinin uygulandığı bir gösterim aracı web ortamında geliştirilmiştir. Bu araç yardımıyla bu birbiri arasında benzerlik tespiti yapılmak istenen kullanıcılar uygulamada seçilerek seçilen kullanıcılar arasında istenen metriklerle benzerlik tespiti yapılabilmektedir. Ayrıca bu uygulama yardımıyla kullanıcılar arası uzaklık matrisi de dosya formatında indirilebilmektedir. Bu şekilde kullanıcı uzaklık matrisi ile farklı kümeleme algoritmaları veya farklı sınıflandırma teknikleri ile analizler yapılmasına olanak sağlanmıştır.

7. GELECEKTEKİ ÇALIŞMALAR

Veri toplama kısmının zorluğu sebebiyle en fazla 48 sosyal medya profili ile deneyler gerçekleştirilmiştir. Veri setinin az olmasına rağmen kullanılan tekniklerin etkileyici sonuçlar vermesi daha büyük bir veri setinde sonuçlar almak, geliştirilen tekniklerin verimliliğini daha iyi ortaya çıkaracaktır.

Bu çalışma Türkçe dili için yapılsa da kullanılan teknikler bakımından herhangi bir dil kısıtına maruz kalmamaktadır. Farklı dillerle de benzer sonuçlar alınabileceği ilerki çalışmalarda gösterilebilir.

Bu çalışmada, deneylerde sadece Twitter ve Facebook sosyal medya platformlarındaki profiller kullanılarak deneyler gerçekleştirilmiştir. Bunların dışında kullanıcılarının yazıyla gönderilerde bulunabildiği Instagram[28], LinkedIn[29] gibi SMP'ler de deneyler için kullanılabilir. Bu şekilde kullanılan tekniklerin platformlardan bağımsız bir şekilde etkinliği hakkında fikir edinilebilir.

KAYNAKLAR

[1] Schneier, Bruce. "A taxonomy of social networking data." *IEEE Security & Privacy* 8.4 (2010): 88-88.

[2] Millham, Mary Helen, and David Atkin. "Managing the virtual boundaries: Online social networks, disclosure, and privacy behaviors." *New Media & Society* 20.1 (2018): 50-67.

[3] Jain, Paridhi, Ponnurangam Kumaraguru, and Anupam Joshi. "@ i seek'fb. me': Identifying users across multiple online social networks." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

[4] Getoor, Lise, and Ashwin Machanavajjhala. "Entity resolution: theory, practice & open challenges." *Proceedings of the VLDB Endowment* 5.12 (2012): 2018-2019.

[5] Wilder, Nathan, Jared M. Smith, and Audris Mockus. "Exploring a framework for identity and attribute linking across heterogeneous data systems." *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*. ACM, 2016.

[6] Malhotra, Anshu, et al. "Studying user footprints in different online social networks." Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012.

[7] Vosoughi, Soroush, Helen Zhou, and Deb Roy. "Digital stylometry: Linking profiles across social networks." International Conference on Social Informatics. Springer, Cham, 2015.

[8] Akcora, Cuneyt Gurcan, Barbara Carminati, and Elena Ferrari. "Network and profile based measures for user similarities on social networks." Information Reuse and Integration (IRI), 2011 IEEE International Conference on. IEEE, 2011.

[9] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[10] McCormick, Chris. "Word2vec tutorial-the skip-gram model." (2016).

[11] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[12] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *arXiv preprint arXiv:1607.04606* (2016).

[13] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.

[14] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[15] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research*9.Nov (2008): 2579-2605.

[16] Kusner, Matt, et al. "From word embeddings to document distances." International Conference on Machine Learning. 2015.

[17] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International Conference on Machine Learning. 2014.

[18] Darling, William M. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling." Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011.

[19] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." Foundations and Trends® in Machine Learning 4.4 (2012): 267-373.

[20] Röder, Michael, Maximilian Speicher, and Ricardo Usbeck. "Investigating Quality Raters' Performance Using Interface Evaluation Methods." *GI-Jahrestagung*. 2013.

[21] Rehurek, R., and P. Sojka. "Gensim–python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3.2 (2011).

[22] Jones, Eric, Travis Oliphant, and Pearu Peterson. "{SciPy}: open source scientific tools for {Python}." (2014).

[23] Hunter, John D. "Matplotlib: A 2D graphics environment." *Computing in science & engineering* 9.3 (2007): 90-95.

[24] Hagberg, Aric, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Almos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[25] Bostock, Michael. "D3. js." Data Driven Documents 492 (2012): 701.

[26] facebook. https://www.facebook.com/ Accessed 1 Aug. 2018.

[27] twitter. https://www.twitter.com Accessed 1 Aug. 2018.

[28] instagram. https://www.instagram.com Accessed 1 Aug. 2018.

[29] linkedin. https://www.linkedin.com Accessed 1 Aug. 2018.