Türkçe Mikroblog Metinlerinde Derin Öğrenme Dil Modeli Tabanlı Konu Tespiti

Topic Detection based on Deep Learning Language Model in Turkish Microblogs

Furkan ŞAHİNÜÇ

Elektrik Elektronik Mühendisliği Bölümü
İhsan Doğramacı Bilkent Üniversitesi
ASELSAN Araştırma Merkezi
Ankara, Türkiye
fsahinuc@aselsan.com.tr

Çağrı TORAMAN

ASELSAN Araştırma Merkezi
Ankara, Türkiye
ctoraman@aselsan.com.tr

Aykut KOÇ

Elektrik Elektronik Mühendisliği Bölümü
İhsan Doğramacı Bilkent Üniversitesi
Ulusal Manyetik Rezonans Araştırma Merkezi
Ankara, Türkiye
aykut.koc@bilkent.edu.tr

Özetçe —Mikroblog yazıları insanların sosyal medyada görüşlerini ifade ettikleri ve herhangi bir düzene bağlı olmayan kısa metinlerdir. Sosyal medya mikroblog metinlerinin konularına göre sınıflandırılması anlamsal bir altyapı oluştururken birçok uygulamaya da yardımcı olur. Bu çalışmada, mikroblog metinlerinde konu tespiti problemi için geleneksel kelime torbası ve derin öğrenme tabanlı modellerin karşılaştırıldığı bir analiz sunulmaktadır. Veri kümesinin oluşturulması için Türkiye’de yaşanan güncel olaylarla ilgili mikroblog metinleri içeren Türkçe "tweet"ler toplanmıştır. Oluşturulan veri kümesindeki "tweet"ler içerdikleri "hashtag" ifadelerine göre etiketlenmiştir. Son haline getirilen veri kümesinde bir adet geleneksel kelime torbası (TF-IDF tabanlı SVM) ve iki adet güncel derin öğrenme yöntemi (BERT ve BERTurk) ile eğitim yapılmıştır. Modellerin başarıları ağırlıklı F1 skoru ile ölçülmüştür. TF-IDF tabanlı SVM 0,807, BERT 0,831 ve BERTurk 0,854 F1 skoru elde etmiştir.

Anahtar Kelimeler—BERT, BERTurk, Derin Öğrenme, Kelime Torbası, Konu Tespiti, Mikroblog, Tweet.

Abstract—Microblogs are short and irregular texts in which people express their opinions in social media. While classification of social media microblog texts according to their topics constitutes a semantic substructure, it helps implementation of various applications. In this study, an analysis comparing conventional bag-of-words and deep-learning based models for the problem of topic detection in microblogs is presented. Turkish tweets containing microblog texts related to current events in Turkey are collected for preparation of the dataset. Tweets in dataset are labeled according to the hashtags they contain. One conventional bag-of-words (TF-IDF based SVM) and two deep learning based models (BERT and BERTurk) are trained on dataset. Performances of the models are measured by using weighted F1 score. TF-IDF based SVM model, BERT and BERTurk perform with F1 scores of 0.807, 0.831 and 0.854 respectively.

Keywords—BERT, BERTurk, Deep Learning, Bag-of-Words, Topic Detection, Microblog, Tweet.

I. GİRİŞ

Sosyal ağlar kullanıcıların fikirlerini paylaşma olanağı sağlayan ve birbirleri ile etkileşime girdikleri çevrimiçi platform-

lardır. Bu tarz platformlarda paylaşılan metinler birçok doğal dil işleme ve makine öğrenmesi çalışmaları için potansiyel içermektedir. Bilgi çıkarımı alanında bu tarz metinler, yeni vaka tespiti ya da duygu analizi gibi çalışmalarda sıklıkla kullanılmaktadır [1], [2]. Mikroblog metinlerinin içerdiği konular bakımından sınıflandırılması da hem kendi başına bir problem olup hem de yukarıda bahsedilen çalışmaların yürütülmesini kolaylaştırmaktadır.

Günümüzde kullanılan en popüler mikroblog platformlarından biri de Twitter’dir. Bu sosyal ağda kullanıcılar bir konu hakkındaki düşüncelerini kısa metinler halinde paylaşabilmektedir. Bu paylaşımı yaparken kullanıcılar yazdıkları konuyu tanımlayan "hashtag" adı verilen özel etiketleri metinlerine ekleyebilmektedirler. Aynı etiketleri paylaşan kullanıcı sayısı arttıkça hakkında yazılan konu popüler hale gelmektedir. Bu nedenle, paylaşılan "hashtag" ifadeleri yeni olayların tespiti gibi alanlarda kullanılmaktadır. Bunun nedeni, gündemdeki önemli olayların popüler hale gelmiş "hashtag" ifadelerinden kolaylıkla tespit edilebilmesidir.

"Hashtag" ifadeleri "tweet"lerin konuları hakkında bilgi verebilmektedir. Fakat tamamen "hashtag"lere dayanarak konu tespiti yapmak doğru sonuç vermeyebilir. Mikrobloglarda bir konu için farklı "hashtag"ler olabileceği gibi, farklı konular için de benzer "hashtag"ler kullanılabilir. Ayrıca, "tweet"lerin görünürlüğünü artırmak amacıyla konuyla alakası olmamasına rağmen popüler "hashtag"ler kullanılabilir. Bu bağlamda, "tweet" konu tespitinde "hashtag" kullanımına dikkat edilmeli ve "hashtag" olmadan metin içeriğinden "tweet" konusu tespit edilmelidir.

Twitter metinleri için konu sınıflandırması literatürde çok yeni bir konu değildir. Popüler konuların sınıflandırılması ya da bilgi filtreleme konularında kolaylık sağlaması için çeşitli çalışmalar yapılmıştır [3], [4]. Fakat her geçen gün doğal dili daha iyi temsil edebilen modeller literatüre sunulmaktadır [5]. Geleneksel ve modern modellerin Türkçe metinlerdeki performansını konu tespiti kapsamında ele alan yeterince çalışma bulunmamaktadır. Bu çalışmada COVID-19 pandemisi gibi güncel konuları kapsayan "tweet"lerden oluşan

978-1-6654-3649-6/21/\$31.00 ©2021 IEEE

bir veri kümesi oluşturulmuştur. Daha sonra bu veri kümesinde benzer konularla ilgili olan "tweet"ler içerdikleri "hashtag" ifadelerine göre etiketlenmiştir. Daha sonra kelime torbası yöntemlerinden TF-IDF tabanlı SVM sınıflandırıcısının [6] yanı sıra, çift yönlü dönüştürücü (transformer) modellerinden BERT [5] ve BERTurk [7] ile konu tespiti eğitimi yapılmıştır.

Çalışmada kullanılan modeller Bölüm II'de anlatılacaktır. Veri kümesinin hazırlanışı ve deneylerde takip edilen yöntemler Bölüm III'te incelenecektir. IV. Bölüm'de ise çalışmanın tartışılması yapılarak sonuca bağlanacaktır.

II. KONU TESPİTİ MODELLERİ

Sınıflandırma işlemi için 3 adet makine öğrenmesi yöntemi seçilmiştir ve sonuçları karşılaştırılmıştır. Bu yöntemlerden ilki geleneksel bir yöntem olan kelime torbası (Bag-of-words) yöntemi olarak seçilmiştir [6]. Diğer ikisi modern yöntemler olup çift yönlü dönüştürücü kodlayıcı model BERT (Bidirectional Encoder Representations for Transformers) [5] ve onun Türkçe ön eğitimi (pre-training) yapılmış versiyonu olan BERTurk olarak belirlenmiştir [7].

A. Kelime Torbası

Kelime torbası yöntemi, veri kümesinde bulunan her bir dokümanı sabit uzunluklu bir vektörle ifade eden bir yöntemdir. Bu noktada, söz konusu dokümanlar eğitim kümesinde bulunan her bir "tweet" örneğine tekabül etmektedir. Bu çalışmada vektörlerin boyutlarının ağırlıklandırması için ise TF-IDF (Term Frequency - Inverse Document Frequency) yöntemi kullanılmıştır [6]. Bütün veri kümesinde bulunan özgün kelimeler doküman vektörlerinde bir boyutu temsil etmektedir. TF-IDF yöntemine göre bir dokümanda sayıca fazla olan fakat diğer dokümanlarda az bulunan kelimeler, söz konusu dokümanın ayırt edilmesinde büyük rol oynamaktadır.

B. BERT

İkinci kullanılan model ise çift yönlü dönüştürücü kodlayıcı olan BERT modelidir [5]. Bu model dikkat (attention) mekanizmasını kullanarak metinde geçen kelimelerin içinde bulundukları bağlamları da eğitim süresince dikkate almaktadır [8]. Bağlamsal özellikler dikkate alındığından dolayı geleneksel kelime temsillerinden (örneğin Word2Vec, GloVe [9]–[11]) ayrılmakta ve birçok doğal dil işleme problemine daha başarılı çözümler sunmaktadır.

BERT modellerinin eğitim süreci iki aşamadan oluşmaktadır. İlk aşama etiketsiz büyük bir derlem üzerinden, dili modelleme çalışması olan ön eğitim (pre-training) safhasıdır. Bu ön eğitimde çift yönlü dönüştürücü yapı ön planda olup, metindeki gizlenmiş kelimeyi bulma ya da bir sonraki cümleyi tahmin etme problemleri üzerine eğitim yapılmaktadır. Yapılan eğitimin sonunda modele verilen herhangi bir kelime sekansı için vektörel bir temsil elde edilebilmektedir. Bu temsil [CLS] olarak adlandırılmaktadır ve girdi olarak verilen sekansın tamamını vektörel olarak yansıtmaktadır.

İkinci aşama ise eğitimin verilen probleme göre değiştiği ve parametrelerin ince ayarlı olarak hesaplandığı aşamadır (fine-tuning). Metinlerin sınıflandırması problemi özelinde yukarıda bahsedilen [CLS] temsillerine ihtiyaç duyulmaktadır. Çünkü

metinlerde geçen kelimelerin bir bütün olarak temsili, sınıflandırma işleminde büyük fayda sağlamaktadır. Bu temsiller parametrelerin ince ayarı için tekrar bir sınıflandırma katmanına girdi olarak verilmektedir. [CLS] temsili, vektörel olarak $C \in \mathbb{R}^H$ şeklinde tanımlanabilir. Burada H gizli parametre sayısını ifade etmektedir. Sınıflandırma için yeni oluşturulan katman da $W \in \mathbb{R}^{K \times H}$ şeklinde tanımlanmaktadır. K ise toplam sınıf sayısını belirtmektedir. Sınıflandırma kayıp fonksiyonu standart şekilde $\log(\text{softmax}(CW^T))$ ile hesaplanmaktadır [5].

C. BERTurk

BERT modelinin önerilmesinden sonra literatüre birçok farklı varyasyonlarla eğitimi yapılan benzer modeller kazandırılmıştır [12], [13]. Bu modellerden bazıları ön eğitim tarzını (örneğin gizlenmiş kelimeyi bulma) değiştirmektedirler. Öte yandan model büyüklüğü ile ilgili çalışmalar da yapılmıştır. Temel modelin eğitilmesi maliyetli olduğu için daha küçük modellerle de benzer performans alan çalışmalar literatüre kazandırılmıştır.

BERT tarzı dönüştürücü modellerin sadece Türkçe'ye has versiyonları literatürde yeteri kadar bulunmamaktadır. Bu eksikliği gidermek adına literatüre kazandırılan Türkçe BERTurk modeli ön eğitimi Türkçe derlem üzerinden yapılmış bir modeldir [7].

III. DENEYLER

A. Veri Kümesi

Çalışmada kullanılacak olan "tweet" veri kümesi 7 Nisan 2020 ve 15 Haziran 2020 arasında Twitter API¹ vasıtasıyla bir araya getirilmiştir. Toplanan "tweet" veri kümesinin 7.743.706 elemanı Türkçe dilindedir. Sınıflandırma çalışmasında "tweet" metinlerini gruplandırmak için "tweet"lerde bulunan "hashtag" özelliği taşıyan kelime ya da kelime gruplarından yararlanılmıştır. Sınıfları belirlemek amacıyla Türkçe "tweet"lerde en çok kullanılan 100 "hashtag" sırayla çıkarılmıştır. Daha sonra bu "hashtag" ifadelerinden birbirleri ile ilgili olanlar aynı sınıf altında toplanmıştır. Örneğin, COVID-19 pandemisi ile ilgili olan "tweet"ler #covid19, #evdekal, #sokagacikmayasagi gibi çeşitli "hashtag" ifadelerini bünyelerinde barındırmaktadır. Şekil 1'de çıkarılan 100 "hashtag" ifadesinin %25'lik kısmı gösterilmiştir.

Toplamda 5 adet "tweet" konu sınıfı belirlenmiştir. Bu sınıflar, 23 Nisan Ulusal Egemenlik ve Çocuk Bayramı, Survivor televizyon programı, ulusal sınavlar, COVID-19 pandemisi ve dizi-film-ünlüler'dir. Sınıfların içerdikleri "hashtag" ifadeleri Tablo I'de verilmiştir.

Tablo I'de görüldüğü üzere, bir sınıf için birçok farklı "hashtag" kullanılabilmektedir. Aynı sınıf içindeki "hashtag" ifadeleri birebir eşleşmese de içinde bulundukları bağlam aynı olmaktadır. Hatta bazı "hashtag" ifadeleri Türkçe karakter içerirken, aynı kelimenin Türkçe karakter içermeyen hali de kullanılabilmektedir. Ek olarak bazı "hashtag" ifadelerinde yazım yanlışları da bulunabilmektedir.

Şekil 1'den de görülebileceği üzere, veri kümesi ile ilgili dikkat edilmesi gereken bir nokta da elde edilen Türkçe

¹<https://developer.twitter.com/en/docs/twitter-api>

Şekil 1: Toplanan Türkçe veride en çok geçen 25 "hashtag".

"tweet"lerde en çok kullanılan "hashtag"lerin dizi-film-ünlüler kategorisine ait olmasıdır. Bu sınıfa ait "tweet"leri incelediğimizde, atılan birçok "tweet"ın aslında "retweet" paylaştığı olduğu ve çoğunlukla popülerliği artırmak amacıyla otomatik hesaplar tarafından paylaşıldığı düşünülmektedir. "Retweet", başka bir kullanıcının yaptığı paylaştığı olduğu gibi tekrar paylaşmaktır. Kullanıcı tekrar paylaştığı "tweet"e herhangi bir katkı yapmamaktadır. Bu durum, sınıflandırmada kullanılacak modellerin, "tweet" metinlerini yanlış bir şekilde öğrenme olasılığını artırmaktadır. Bu olasılığı en aza indirmek için "Retweet" özelliği taşıyan "tweet"ler sınıflandırma için kullanılacak eğitim kümesine dahil edilmemiştir. Sadece kullanıcıların kendi ifadelerini içeren orijinal "tweet"ler kullanılmıştır. Bu işlemden sonra "tweet" sayısı en yüksek sınıf Survivor programı olmuştur. Kullanıcıların orijinal "tweet"leri sınıflandırmanın güvenilir bir şekilde yapılmasına da olanak sağlamaktadır. "Retweet" elemelerinden sonra veri kümesinin son halinde toplam 115.241 tweet bulunmaktadır. Sınıf bazında istatistikler Tablo II'de verilmiştir.

Eğitimde kullanılacak olan "tweet" kümesi belirlendikten sonra "tweet"lerin temizlenme aşamasına geçilmiştir. "tweet"lerde bulunan bütün "hashtag" ifadeleri, diğer kullanıcılara yapılan atıflar ve internet bağlantıları kaldırılmıştır. Ayrıca noktalama işaretleri, emojiler gibi alfasayısal olmayan

TABLO I: Konu sınıfları ve hashtag grupları.

Konu	Hashtag
Survivor Programı	#survivor2020, #survivorekstra, #survior2020, #survivor, #survivior2020, #survivo2020, #survivorpanorama
Sınav	#ykstarihinedokunma, #yks2020, #sınavhayatçalmasin, #ykstarihinedokunma, #yksertelevirüsüyayma, #lgs2020, #turkishstudentlivesmatter
COVID-19	#sokagacikmayasagi, #evdekal, #covid19, #coronavirus, #covid_19, #koronavirüs, #corona, #sokagacikmayasagi
23 Nisan	#23nisan kutlu olsun, #23nisan
Dizi Film Ünlüler	#ozandolunay, #cencem, #berkerguven, #cemned, #ayker, #urazkaygilaroglu, #narged, #caglarertugrul, #keremiyigiter, #neslihanatagül, #enginakyurek, #barisarduc, #kiralıkkaşk, #nedimkaraçay, #nehnaz, #narsan, #reymir, #cenkkaraçay, #cihangircayhan, #hercai, #serakutlubey, #sancarefeoglu, #cembar, #ebruşahin

TABLO II: Deneylerde kullanılan tweet sayıları.

Sınıflar	Tweet Sayıları
Survivor Programı	58.412
Sınav	16.559
COVID-19	15.478
23 Nisan	2.051
Dizi Film Ünlüler	22.741
Toplam	115.241

karakterler de temizlenmiştir. Son olarak tüm harfler küçük harflere çevrilmiş ve 2 karakterden kısa, 16 karakterden uzun kelimeler "tweet" metinlerinden çıkartılmıştır. Çok kısa kelimeler öğrenme modellerine fazla bir bilgi sağlayamamaktadır. Çok uzun kelimeler ise çoğunlukla rastgele yazılan ifadelerden oluşmaktadır. Böyle ifadeleri eğitim kümesine dahil etmeyerek eğitim kümesindeki gürültüyü azaltmak hedeflenmiştir.

B. Deney Yöntemi

TF-IDF yöntemi ile oluşturulan öznitelikler, klasik bir makine öğrenmesi yöntemi olan SVM (Support Vector Machine) sınıflandırma modeli ile sınıflandırılmıştır. Çekirdek fonksiyonu olarak doğrusal yöntem kullanılmıştır. Özniteliklerin çıkartılması ve sınıflandırma modelinin eğitiminde *scikit-learn* kütüphanesinden yararlanılmıştır [14]. Veri kümesinin %80'lik kısmı eğitim, %20'lik kısmı test olarak ayrılmıştır. Veri kümesindeki sınıfların dağılımı dengeli olmadığından dolayı sınıflandırma modelinin başarısı ağırlıklı F1 skoru ile ölçülmüştür.

Derin öğrenme yönteminde model olarak ön eğitimi yapılmış BERT-cased modeli kullanılmıştır ve ince ayar aşaması oluşturduğumuz veri kümesi ile gerçekleştirilmiştir. Ön eğitimi yapılmış modeli kullanırken Huggingface kütüphanesinden yararlanılmıştır [15]. Veri kümesi Türkçe olduğu için BERT çok dilli modeli (mBERT) kullanılmıştır ve bu modelin son katmanındaki parametre sayısı $H = 768$ 'dir. TF-IDF tabanlı SVM modeli eğitiminde olduğu gibi veri kümesinin %80'i eğitim, %20'si test için ayrılmıştır. Modelin başarısı da yine ağırlıklı F1 skoru ile ölçülmüştür.

BERTurk parametrelerinin ince ayar eğitimi yapılan kısmı, BERT modelinde anlatılan yöntemin aynısı izlenerek gerçekleştirilmiştir. İki modelin de parametre sayısı aynı tutulmuştur. Bu yöntemde de veri kümesinin %80'i eğitim, %20'si test olarak ayrılmıştır ve model başarısı ağırlıklı F1 skoru ile ölçülmüştür.

C. Deney Sonuçları

Test için ayrılmış veri kümesindeki model başarıları Tablo III'te verilmiştir. Modellerin içinde en yüksek başarıyı BERTurk modeli elde etmiştir. TD-IDF SVM, BERT ve BERTurk modellerinin karmaşa matrisleri sırasıyla Tablo IV, V ve VI'da verilmiştir.

TABLO III: Modellerin sınıflandırma başarısı.

Model	F1 Skoru (%)
TF-IDF SVM	0,807
BERT	0,831
BERTurk	0,854

Eğitilen üç model arasında en düşük F1 skorunu TF-IDF tabanlı SVM modeli elde etmiştir. Fakat diğer modellerle karşılaştırıldığında aradaki farkın çok fazla olmadığı görülmektedir. TF-IDF yöntemi sınıflandırılacak metinler için öznitelik çıkarımında genel olarak tercih edilen bir yöntemdir. "Tweet"ler kısa metinler oldukları için az kelime ile fazla açıklayıcılık sağlamaları gerekmektedir. Bu nedenle uzun bir metne göre konuya spesifik kelimelerin yoğunluğu daha fazla olmaktadır. Bu durum, TF-IDF ile çıkarılan özniteliklerin belirleyici olmalarını sağlamaktadır.

TABLO IV: TF-IDF tabanlı SVM modelinin karmaşa matrisi.

		Tahmin Edilen Sınıf				
		Survivor	Sınav	COVID-19	23 Nisan	Ünlüler
Doğru Sınıf	Survivor	11.043	198	160	7	204
	Sınav	718	2.308	257	2	55
	COVID-19	667	225	2.284	13	45
	23 Nisan	86	16	52	232	10
	Ünlüler	1.427	103	65	3	2.869

TABLO V: BERT modelinin karmaşa matrisi.

		Tahmin Edilen Sınıf				
		Survivor	Sınav	COVID-19	23 Nisan	Ünlüler
Doğru Sınıf	Survivor	10.839	213	290	13	257
	Sınav	488	2.432	370	7	43
	COVID-19	336	202	2.649	20	27
	23 Nisan	31	24	80	252	9
	Ünlüler	1.131	139	139	3	3.055

TABLO VI: BERTurk modelinin karmaşa matrisi.

		Tahmin Edilen Sınıf				
		Survivor	Sınav	COVID-19	23 Nisan	Ünlüler
Doğru Sınıf	Survivor	10.752	209	151	9	491
	Sınav	400	2.615	192	19	114
	COVID-19	337	245	2.518	44	90
	23 Nisan	32	18	30	289	27
	Ünlüler	773	104	48	9	3.533

Diğer kullanılan BERT modelleri geleneksel yöntemlerden daha başarılı olup daha karmaşık yapılar içermektedir. Metinlerdeki kelimelerin içinde bulundukları bağlamları da dikkate aldıklarından dolayı metinleri daha iyi temsil edebilmektedirler. Bunun yanında metin uzunluğu arttıkça bu tarz modellerin kendi performanslarını artırma eğilimi de bulunmaktadır. İki BERT tabanlı model de geleneksel yöntemli sınıflandırmayı geçmeyi başarmıştır. En iyi F1 skorunu BERTurk modeli elde etmiştir. BERTurk modelinin Türkçe bir sınıflandırma probleminde orijinal BERT modelini geçmesi makul bir sonuçtur. Çünkü çok dilli BERT modeli Türkçe dahil birçok farklı dil içermektedir. Buna karşın BERTurk modeli ise ön eğitimde sadece Türkçe metinleri kullanmıştır. Bu nedenle daha iyi bir sınıflandırma performansı göstermiştir.

Karmaşa matrisleri incelendiğinde TF-IDF modelinin en kalabalık olan Survivor programı sınıfında en fazla doğru tahmini yaptığı görülmektedir. Fakat benzer başarıyı diğer sınıflarda gösterememektedir. Bu da diğer modellere kıyasla daha düşük F1 skoru elde etmesine sebep olmaktadır. Öte

yandan BERTurk modeli Survivor ve COVID-19 sınıflarında BERT modeline göre daha düşük bir başarı elde etmesine rağmen, özellikle dizi-film kategorisinde çok ciddi bir başarı sağladığı için daha yüksek bir F1 skoru elde etmiştir.

IV. SONUÇ

Sosyal medya mikroblog örneklerinden olan "tweet"lerde bulunan "hashtag" ifadeleri, "tweet"lerin sınıflandırılmasında kolaylık sağlamaktadır. Bu çalışmada toplanan Türkçe "tweet"ler benzer "hashtag" gruplarına göre 5 sınıfa ayrılmıştır. Daha sonra "hashtag"ler "tweet"lerden atılarak üç farklı yöntemle sınıflandırma işlemi yapılmıştır. Bu yöntemler geleneksel kelime torbası (TF-IDF tabanlı SVM) ve derin öğrenme tabanlı (BERT ve BERTurk) modellerdir.

Sonraki çalışmalarda daha başarılı skorlar elde etmek için Türkçe dilinde eğitilmiş farklı derin öğrenme dil modelleri ve Twitter dışındaki sosyal medya platformlarından elde edilen veri kümeleri ile de çalışılabilir.

KAYNAKLAR

- [1] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 1998, p. 37–45.
- [2] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. ACL: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 151–160.
- [3] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *IEEE 11th Int. Conf. on Data Mining Workshops*, 2011, pp. 251–258.
- [4] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in Twitter to improve information filtering," in *Proceedings of the 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2010, p. 841–842.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
- [6] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [7] S. Schweter, "BERTurk - Bert models for Turkish," Apr. 2020.
- [8] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.
- [9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NeurIPS*. Curran Associates, Inc., 2013, pp. 3111–3119.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [11] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.
- [12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *ICLR*, 2020.
- [13] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [15] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.