SOSYAL MEDYA İÇERİĞİNDEN VERİSETİ OLUŞTURMA VE VERİLERİ TEMİZLEME

YZ 5521, VERI ANALITIĞINE GIRIŞ

Elif Muslu
elifmuslu2@posta.mu.edu.tr

Saturday 30th January, 2021

Abstract

Gelişen teknoloji ile beraber sosyal medya çok önemli bir veri kaynağı oldu. Sosyal medya ile baraber iletişim tek taraflı boyuttan interaktif bir şekil almış, haberlere ve olaylara olan tepkilerin nabzını tutmak son yıllarda önemli araştırma konusu haline gelmiştir. Bu çalışmada sosyal medya da tarşılan bir konu olan İzmir depremi verileri toplanmış ve üzerinde duygu analizi yapılmıştır. Ayrıca bulunan sonuçlar görselleştirilmiştir.Bu sayede sonuçlar daha anlaşılır bir şekilde ifade edilmiştir.

1 Giriş

Günümüzde sosyal medya vazgeçilmez bir alan kaplamaktadır. Gazete, televizyon, radyo gibi iletişim araçlarına yeni bir alternatif olan sosyal medya, popülerliğini her geçen gün arttırmaktadır. popülerliğini her geçen gün arttırmaktadır. Özellikle Facebook, Twitter gibi sosyal platformlar bilginin oluşması, yayılması, tartışılması, zamanla tekrar ortaya çıkması gibi olayları hiç olmadığı kadar hızlandırmıştır. Ayrıca sosyal medya sayesinde tek taraflı iletişim ve haber paylaşımı yerine daha çok interaktif bir boyut kazanmıştır. Sosyal medya kullanıcının kendilerinin bilgi paylaşması dışında beğeni ya da paylaşım yapan kullanıcının bilgilerinide paylaşabilmektedirler. Herhangi bir konudaki (haber, olay, spor, ekonomi, alışveriş, eğlence gibi) gelişmeyi anlık olarak sosyal medyadan paylaşmaktadırlar ve sosyal medya ile takip etmektedirler. Sosyal medya büyük bir veri havuzunu içermektedir. Bir konu hakkında veri toplayıp analiz etmek için de önemli bir kaynak olduğu ortaya çıkmıştır. Sosyal medya verileriyle ekonomi, hastalık gibi bir çok alanda başarılı sonuçlar alınmıştır. En popüler sosyal medya platformu olan Twitter bir çok araştırmaya öncü olmuştur. Paylaşılan her bir tweetin, kullanıcılarının izni ile konumu, zamanı gibi bilgilere de ulaşılabilir. Ulaşılan bu verilerden de bir çok analiz yapılabilir.Bu kullanılan analizlerden biri duygu analizidir. Duygu analizi bir tweetin,yazının veya durumun olumlu, olumsuz ya da tarafsız yani nötr olup olmadığının belirlenmesidir. Çok önemli sonuçlar veren duygu analizi yazılımları son yıllarda market araştırması açısından önem bir boyut kazanmıştır.

Bu çalışamada 30 Ekim Cuma günü saat 14:51'de İzmir ilinin Seferihisar ilçesinde meydan gelen büyük İzmir depremi hakkında Twitter sosyal meyda platformunda atılan tweetler uygun platformdan çekilmiş ve sonuçlarının tartışılıp görüntülenebileceği bir duygu analizi yapılmıştır.

2 METOD

Bu başlık altında çalışmanın metodu içerisinde yer alan verilerin yapısı ve sayısı, veri toplama ve temizleme aşamaları ve son adımda da duygu analizinin yapılması üzerinde durulmuştur. En son olarak NaiveBayes makine öğrenimi algoritmasının duygu analizi için ne kadar iyi performans gösterdiğine bakıldı. Naive Bayes yöntemi en basit ve kolay makine öğrenimi yöntemlerinden biridir ve belli bir hedefin istenilen sınıf değerine ait olma olasılığını bulmaya yardımcı olur. Değerlendirme sonuçlarını iyileştirmek için cümlenin bağlamını ve anlamına dikkate almaya yardımcı olan LSTM(Long-short term memory) ağı yardımı ile Makine öğrenimi algoritmasına göre kıyaslandı. LSTM de yapay bir tekrarlayan sinir ağıdır. Tekrarlayan verileri kullanarak tahmin de bulunmaya yardımcı olur.

2.1 Veriseti

Twitter'dan anlık veri toplamak için Twitter API (Twitter Apps, 2017) kullanılmıştır. Twitter Archiving Google Sheet ile Twitter API kullanılarak gerekli izinler(api key ve access token) oluşturulmuştur. 'izmirdepremi', 'izmirdeprem' ve 'izmir' anahtar sözcüğü içeren tweetler kullanılarak en geç 7 gün geriye dönük veri toplanmıştır. 04-11 Kasım 2020 tarihleri arasında toplam 10878 veri toplanmıştır.

2.2 Veri Temizleme

Twitter mesajları içerisinde et(@) işareti ile başlayan kullanıcı adı, hashtag ile hashtag, URL formatında web sitesi adresleri ve emojileri barındırabilmektedir. Bu karakterler tweetler içerisinden çıkarılarak veri arındırılacaktır.Duygu analizi işlemine geçilmeden önce tweetlerde bulunan bu özelliklerin ön işlemden geçirilmesi gerekmektedir. İlk olarak Python BeautifulSoup Modülü HTML ve XML ayrıştırıcısı kullanıldı. Bunun görevi neredeyse orjinal belgenizle aynı anlama gelen bir ayrıştırma ağacı (parse tree) döndürür ve bu sayede ayrıştırma ağacında kolayca gezinme (traversing), arama ve düzenleme yapmanıza olanak sağlar. HTML'yi genel metne dönüştürmek, veri hazırlamanın ilk adımıdır ve bunun için BeautifulSoup'u kullanılmıştır. Ardından '@bahsetme' yani başka bir kullanıcının bahsettiği bahsetme durumu kaldırılmıştır çünkü bu bilgi duygu analizi oluşturmak için bir bilgi katmaz.Temizlemenin diğer bir kısmı @bahsetme ile aynı olan URL bağlantılarını temizlemektir. Bu URL'lerin bazıları bilgi içermesine rağmen duygu analizi yapılacağı için göz ardı edilecektir. Bazen etiketler ile kullanılan metin tweet hakkında faydalı bilgiler sağlayabilir. Bu nedenle etiketler ile birlikte tüm metinden kurtulmak biraz riskli olabilir. Bu yüzden metin olduğu gibi bırakıldı ve sadece hashtag() işaretini kaldırmaya karar verildi. Bu işlemleri tüm verisetine uygulamak için bir fonksiyon yazıldı ve Python içerisinde bulunan NLTK kütüphanesi kullanılmıştır. Bu sayede tüm veri setinden istenilen durumlar temizlendi ve tüm harfler küçük harfe dönüştürüldü. Verisetinde benzer veri birden fazla olduğu için bunları temizlemek gerekti. Pandas kütüphanesi yardımıyla 'text' sütunu içerisindeki aynı tweetler temizlendi. Bu temizlemenin sonucunda NaN veriler ortaya çıktı ve bunlarda dolduruldu. Veriler temizlendikten sonra elimizde 6420 adet temiz veri kaldı. Verileri temizledikten sonra .csv dosyasına kaydedilip .csv formatında veri dışa artarıldı. Artık elimizde 3577 adet negatif tweet, 1123 adet pozitif tweet ve 1719 adette nötr duygu durumuna sahip tweet bulunmaktadır.

2.3 Duygu Analizi

Dünya da duygu analizi için kullanılan HowNet , Senti WordNet , MPQA vb. sözlükler kullanılmaktadır. Bu çalışma da, Türkçe duygu analizi için verisetindeki tweetleri manuel olarak

skorladık. Negatifler için 0, pozitifler için 4 ve nötr tweet için 2 skorunu girdik. Stopwords ile etkisiz kelimeleri atma işlemi yapılmaktadır. Metinlerle çalışırken, daha iyi bir analiz veya işlem yapmak için etkisiz kelimeler metinden atılır. Stopwords değişkeni Türkçe(ya da İngilizce)'deki tüm etkisiz kelimeleri bir liste halinde tutar. Ardından WordCloud ile de kelimeleri estetik bir görüntü sağlamak için görselleştirdik. Temizlenen veri, test ve eğitim seti olarak ikiye ayırıldı. Ardından nötr duygu olarak işaretlenen 2 skoruna sahip tweetler çıkarıldı. Temizlenmiş verilerin içerdikleri kelimeleri kolayca görselleştirmek için eğitim seti negatif ve pozitif olarak ayırıldı. Artık sadece pozitif ve negatif tweetlerin en empatik kelimelerini yani vurgulanmış kelimelerini bir Worldcloud yardımıyla görselleştirdik. Devamında da nltk lib yardımıyla ilk kelimelerin sık dağılımına bakıldı ve ortaya çıkan anahtar kelimelere bakılarak sözde özellikleri çıkarıldı. Bu şekilde İzmir depremi için atılan tweetlerde pozitif, negatif ve en sık kullanılan kelimeleri ortaya çıkarmış olduk. Son olarakta NaiveBayes Makine Öğrenimi algoritmasının duygu analizi için ne kadar iyi performans gösterdiğine bakıldı. Değerlendirme sonuçlarını iyileştirmek için cümlenin bağlamını ve anlamına dikkate almaya yardımcı olan LSTM(Long-short term memory) ağı yardımı ile Makine öğrenimi algoritmasına göre kıyaslanmıştır.

3 Tartışma

Yapılan duygu analizi ile alakalı çıktılar ve ekran görüntüleri aşağıdaki gibidir.

Figure 1: Çıkarılan Pozitif Kelimeler

Pozitif olumlu kelimelerin görselleştirilmiş haline bakılınca beraber atlacatacağız, İzmir, İzmirdepremi hisseden oldu, pelerinsiz kahramanlar,depremzadeler için, depremzadelere kebap, destek olun, farkındalık oluşturmak, izmirin yanındayız, ikram, kahraman,aydabebek,aydagezgin, kontrol, ihtiyaç, son dakika, sağlık,teşekkür ederiz, kızılay, urla, bornava, gönüllü,afad,sevgi,ziyaret, elif bebek,yarım kalan,hayallere dokunmak ve dini terimler içermektedir. İnsanların olaylar hakkında atmış olduğu tweetlerde destek kelimeleri, depremin olduğu yer İzmir ve depremzadelere ikramlar ve destek tweetleri şeklindedir.

Figure 2: Çıkarılan Negatif Kelimeler

Figure 2'de Negatif olumsuz kelimelerin görselleştirilmiş haline bakılınca kardeşlik, izmir depremi, izmir, itfaiye, sessizlik, atlatacağız, pelerinsiz kahramanlar,furkan gönüllüleri, arama kurtarma, razı olsun, karşıyaka, ruhsat verende, destek olmak, afad,gıda, kardeşlik, saat, bayraklı, televizyonda haber, enkaz altında, ayda bebek, alan buz, bütün sorumlusu, deprem, izmir deprem,lütfen,covid, depremdeğilcinayet,son dakika, yardım,enkaz altında,deprem verigileri nerede,rıza bey apartmanı ve dini terimler yer almaktadır. Burda gördüğümüzde İzmir depreminde çoğu insanın tek yürek olduğu, benzer tweetler attığı görülmektedir. İnsanlar umut etme eğilimindeler ve depremin verdiği üzüntülü durumdan yakınılmaktadırlar.

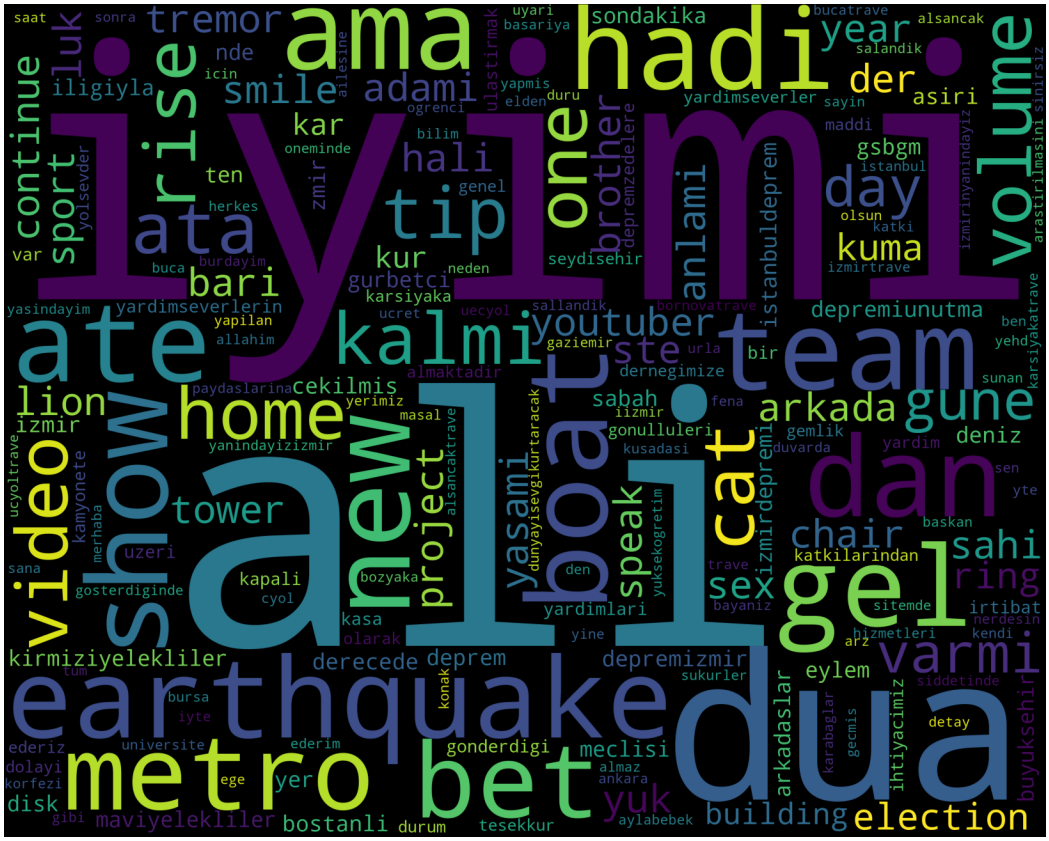

Figure 3: Ortaya Çıkan Anahtar Kelimeler ile Çıkarılan Özellikler

Figure 3'de nltk lib ile, önce sık bir dağılımı ölçerek ve ortaya çıkan anahtarlar kelimeleri seçerek, sözde özellikleri çıkarıldı. Bu şekilde en sık dağıtılan kelimeler ortaya çıkarıldı. Çoğu kelime, earthquake,dua, yük,victim(kurban),hadi, help, disaster(felaket), vefat, ev, sıcak, building(bina), tremor(titreme), iyileşeceğiz ve death(ölüm), speak, acil, var mı,tıp,metro, kontrol,iyimi, cadde, çökebilir, sağlammı, yaşam, acili, var mı, alsancak, aynıçatıaltındayız, dakikalar, yardım, izmir depremi ve İzmir depremi etrafında toplanmaktadır.

4 Sonuç

Bu çalışmada Twitter üzerinden toplanan veriler 30 Ekim 2020 günü Türkiye saati ile 14.51'de İzmir Seferihisar da gerçekleşen İzmir depremi ile alakalı olup 10878 tweet içeren bir veri seti elde edilmiş. Temizlenen verilerin sonucunda 6420 adet veri kullanılmıştır. Verisetini temizleme aşamasında Python programlama dili kullanılmıştır ve NLTK (Natural Language Tool Kit), Pandas, Numpy ve matplotlib kütüphaneleri kullanılmıştır. Cümlelerin pozitif, negatif ve nötr oluşu manuel olarak etiketlenmiştir. Figure 1 ve Figure 2 de atılan tweetlerdeki pozitif ve negatif kelimeler görselleştirilmiştir.Ardından sınıflandırıcı algoritma olan NaiveBayes ile negatif tweetlerde 363/328 bir skor, pozitif tweetlerde 105/42 bir skorlama yapılmıştır. Bu skorlamaya bakıldığında atılan tweetlerin çoğunluğunun İzmir depremi hakkında negatif duygular içeren tweet olduğu görülmektedir.

```
[Negative]: 363/328
[Positive]: 105/42
```

Figure 4: NaiveBayes Sınıflandırma Skorlaması

Daha sonra cümlenin bağlamını dikkate alan Long-Short Term Memory (LSTM) ile performansına bakıldı. Maksimum özellik sayısı tanımlandı ve metni vektörleştirmek ve dizi haline getirmek için tokenizer kullanıldı.4 katmandan oluşan sıralı bir LSTM modeli oluşturuldu ve aktivasyon fonksiyonu olarak softmax kullanıldı. Dropout değeri 0.2, batch size 32 ve optimizasyon algoritması Adam seçildi.Toplam parametre sayısı 62,691 elde edildi. Model 5 epoch ile eğitildiğinde loss değeri 0.4709'e kadar düştü ve accuracy değeri 0.8007 elde edildi. Modelin hesaplanan Score değeri 0.50 ve acccuracy değeri de 0.83 başarı değeri elde etti.

```
Epoch 1/5
100/100 - 99s - loss: 0.5176 - accuracy: 0.7663
Epoch 2/5
100/100 - 98s - loss: 0.4913 - accuracy: 0.7802
Epoch 3/5
100/100 - 98s - loss: 0.4813 - accuracy: 0.7846
Epoch 4/5
100/100 - 98s - loss: 0.4757 - accuracy: 0.7941
Epoch 5/5
100/100 - 98s - loss: 0.4709 - accuracy: 0.8007
<tensorflow.python.keras.callbacks.History at 0x7fad513a9240>
```

Figure 5: Modelin Eğitiminin Sonucu

```
2/2 - 1s - loss: 0.5034 - accuracy: 0.8254
score: 0.50
acc: 0.83
```

Figure 6: LSTM Score ve Accuracy Değerleri

```
pos_acc 16.86046511627907 %
neg_acc 97.23183391003461 %
```

Figure 7: LSTM ile Doğru Tahminlerin Oranı

Figure 7'de, doğru tahmin sayısı değerlerlerine bakılırsa modelin negatif tweetleri bulmada başarılı olduğu görülmektedir. Ancak tweetin olumlu olup olmadığa karar vermede iyi olmadığı ortadadır. Sonuç olarak eğitimli tahminin burada pozitif eğitim setinin olumsuz eğitim seti olandan önemli ölçüde daha küçük olduğu, dolayısıyla olumlu tweetler için kötü sonuçlar verdiği görülmektedir. Eğitim verileri çok dengesiz olduğu için(pozitif:2220, negatif:7088) güvenilir veriler elde etmek için daha çok veri elde edilmelidir. Verisetinin eksikliği ve yetersizliği tutarsız sonuçlara sebep olduğu görülmektedir.Ne kadar çok veri olursa o kadar çok yüksek oranda doğruluk elde etmek mümkündür. Ayrıca daha iyi sonuçlar da elde etmek için LSTM modelinin daha yüksek epoch değerleri ile eğitilmesi ve kullanılan parametrelerde en doğru sonuç verenlerin kullanılması gerekir.

References

[1] BOLLEN, J., MAO, H. ZENG, X.-J. (2011). Twitter mood predicts the stock market. J. Comput. Science, 2, 1- 8.

[2] DEHKHARGHANI, R., YANIKOGLU, B., SAYGIN, Y. OFLAZER, K. (2017).

Sentiment analysis in Turkish at different granularity levels. Natural Language Engineering.
[3] NLTK 3.2.5 Kütüphanesi. http://www.nltk.org/ adresinden alındı.

[4] HEALEY, C. G. (2017, 08 25). Visualizing Twitter Sentiment.
https://www.csc2.ncsu.edu/faculty/healey/tweet$_viz$/*adresindenalınd*ı.

[5]Mehmet A. , Kamil T., Volkan A.(2017). DATA ANALYSIS ON SOCIAL MEDIA: TWITTER The Journal of Faculty of Economics and Administrative Sciences Vol.22, Special Issue on Kayfor15, pp.1991-1998.

[6] KAYA, M., FİDAN, G., TOROSLU, I. H. (2012). Sentiment Analysis of Turkish Political News. Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (s. 174- 180). Washington, DC, USA: IEEE Computer Society

[7]Pandarachalil, R., Sendhilkumar, S., Mahalakshmi, G. S. (2015). Twitter sentiment analysis for large-scale data: an unsupervised approach. Cognitive computation, 7(2), 254-262.

[8] Twitter veri çekmek için https://tags.hawksey.info/ kullanıldı.

[9]Twitter Apps. (2017, 8 1). https://apps.twitter.com/ adresinden alındı.