TWITTER'DA GERÇEK-SAHTE HABER TESPIT PROJESİ

Cansu Ayten 171180010

Gamze Aksu 171180005

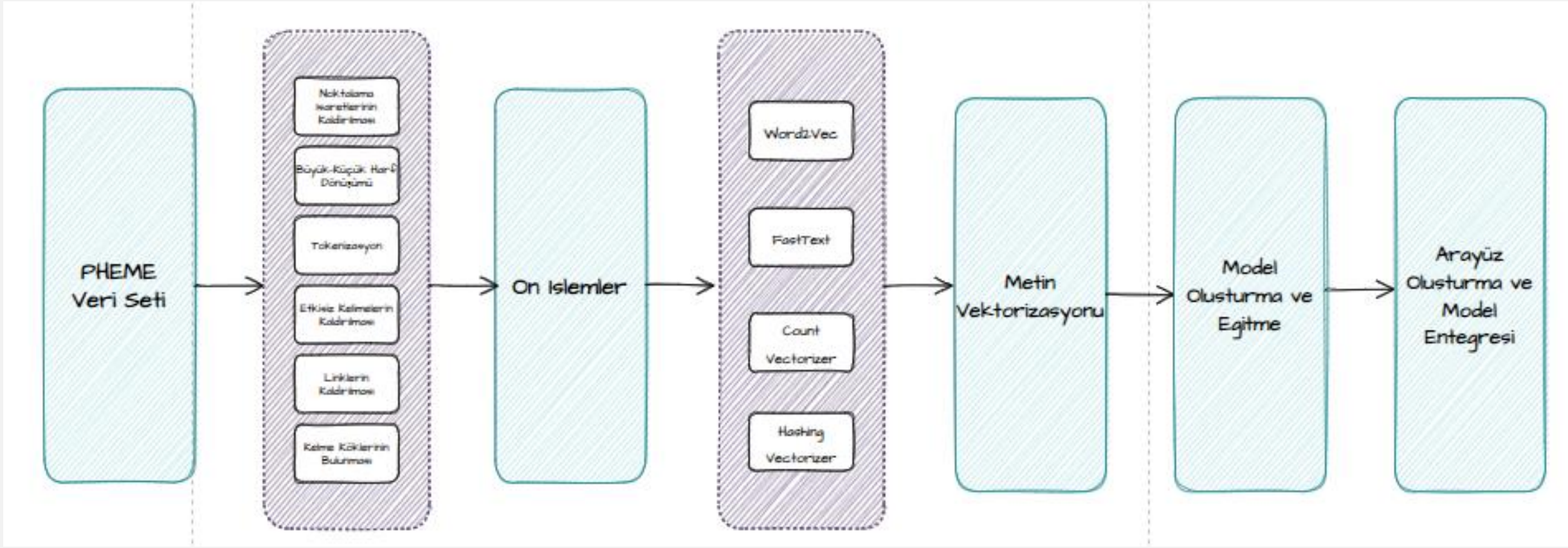
TWITTER'DA GERÇEK-SAHTE HABER TESPIT PROJESİ

- Twitter günümüzde dünyadaki en çok kullanılan sosyal medya platformlarından biridir.
- Çok paylaşılan bir metin Twitter'da gündem olabilir ve böylelikle bu metin eğer yanlış bir bilgi içeriyorsa insanlar arasında yayılması kötü sonuçlar doğurabilir.
- Bilgi kirliliği yanlış, abartılı, yanlı ve doğruluğu kanıtlanmamış bilgilerin kasten, bazen iyi bazen de kötü niyetle yayılması demektir.

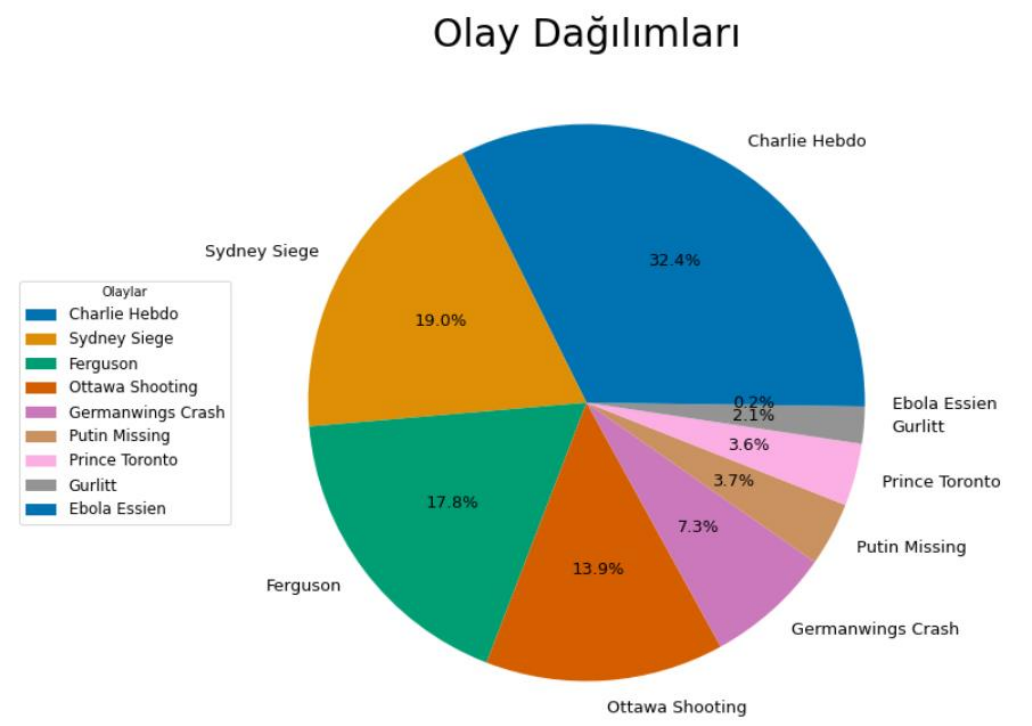
https://sarkac.org/2021/05/bilim-akademilerinden-bilgi-kirliligine-dair-bir-rapor/

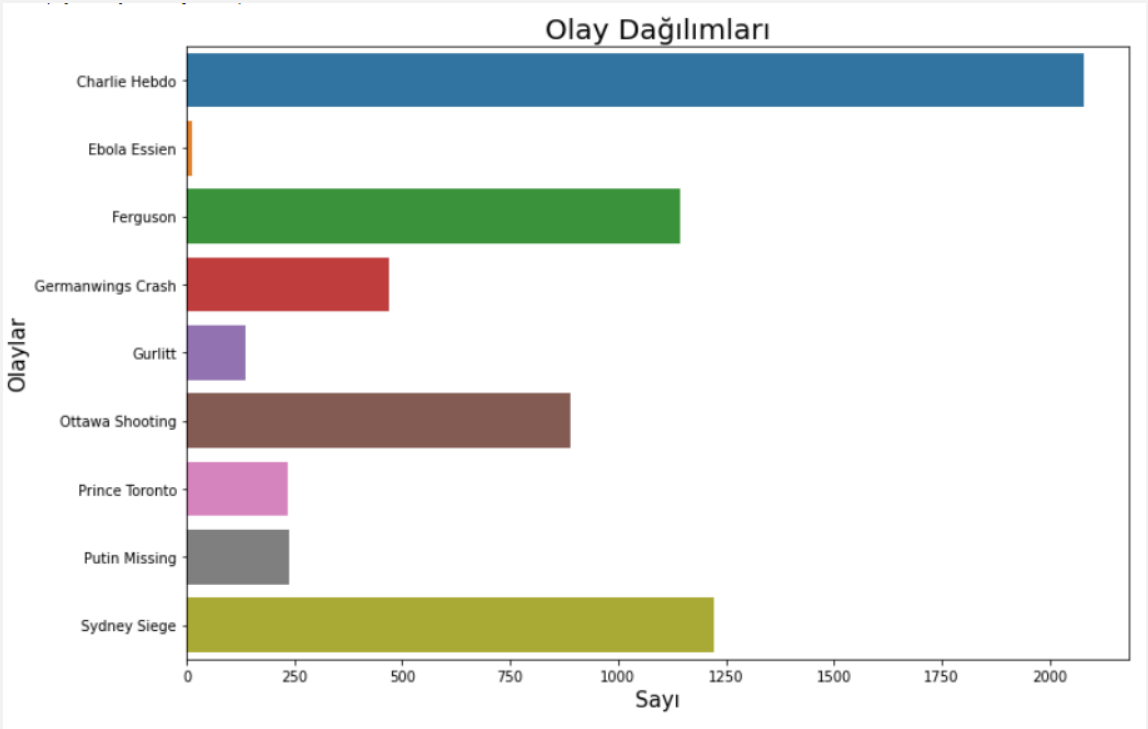
PHEME VERİ SETİ

- 9 farklı olaya ait haber başlığına sahiptir.
- 6425 tweet bulundurmaktadır.
- Bu tweetlerden 4022 gerçek tweetlerden oluşur.
- Bu tweetlerden 2402 sahte tweetlerden oluşur.
- 1 belirsiz tweet bulunmaktadır.
- .json → .csv

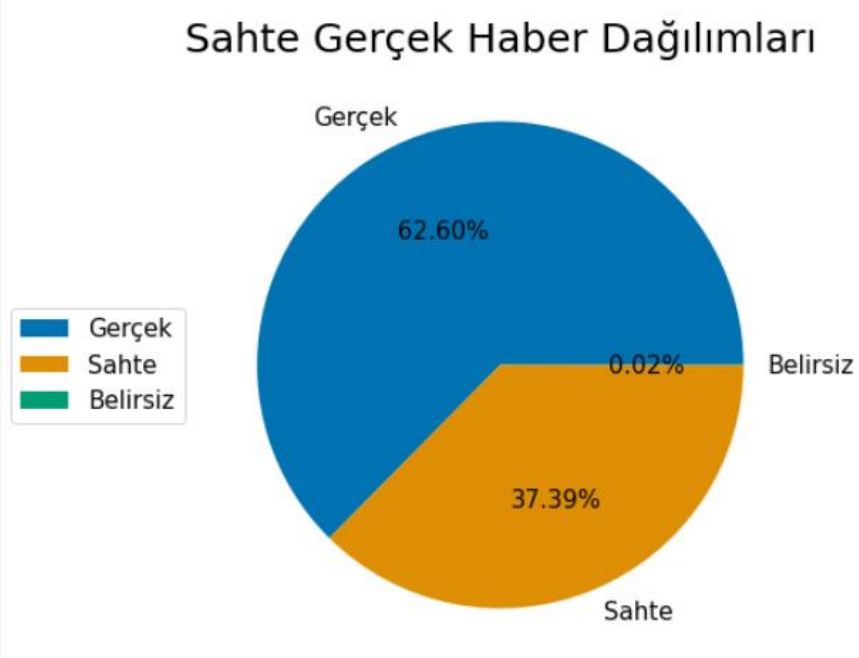

PHEME
Veri Seti
Noktalama
Karakterlerinin
Kaldırılması
Büyük-Küçük Harf
Dönüşümü
Tokenizasyon
Etkisiz Kelimelerin
Kaldırılması
Linklerin
Kaldırılması
Kelime Köklerinin
Bulunması
On Islemler
Word2Vec
FastText
Count
Vectorizer
Hashing
Vectorizer
Metin
Vektorizasyonu
Model
Olusturma ve
Egitme
Arayüz
Olusturma ve
Model
Entegresi

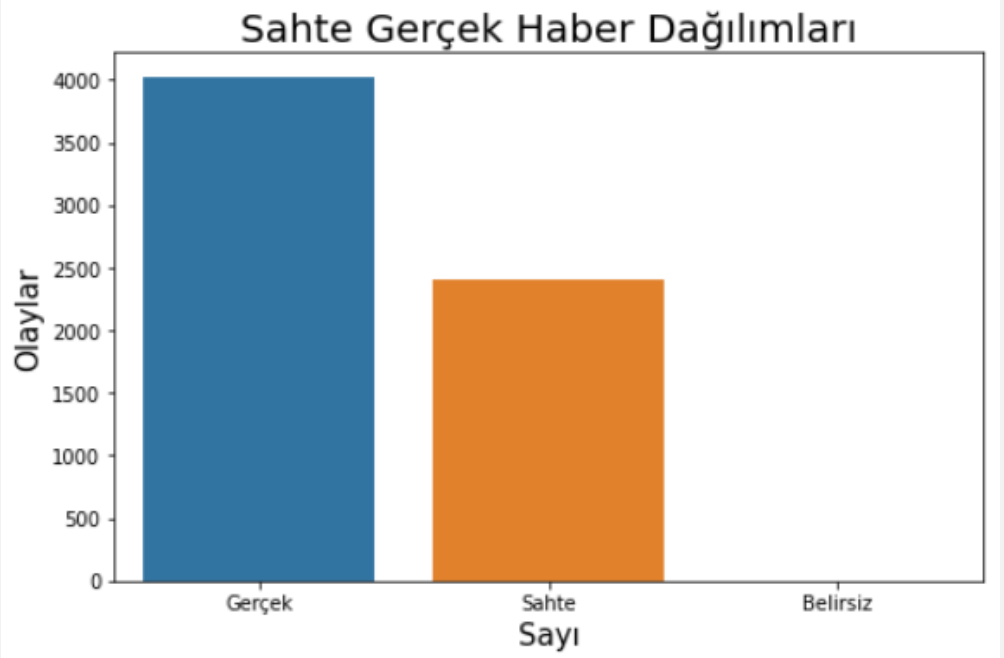
OLAY DAĞILIMLARI

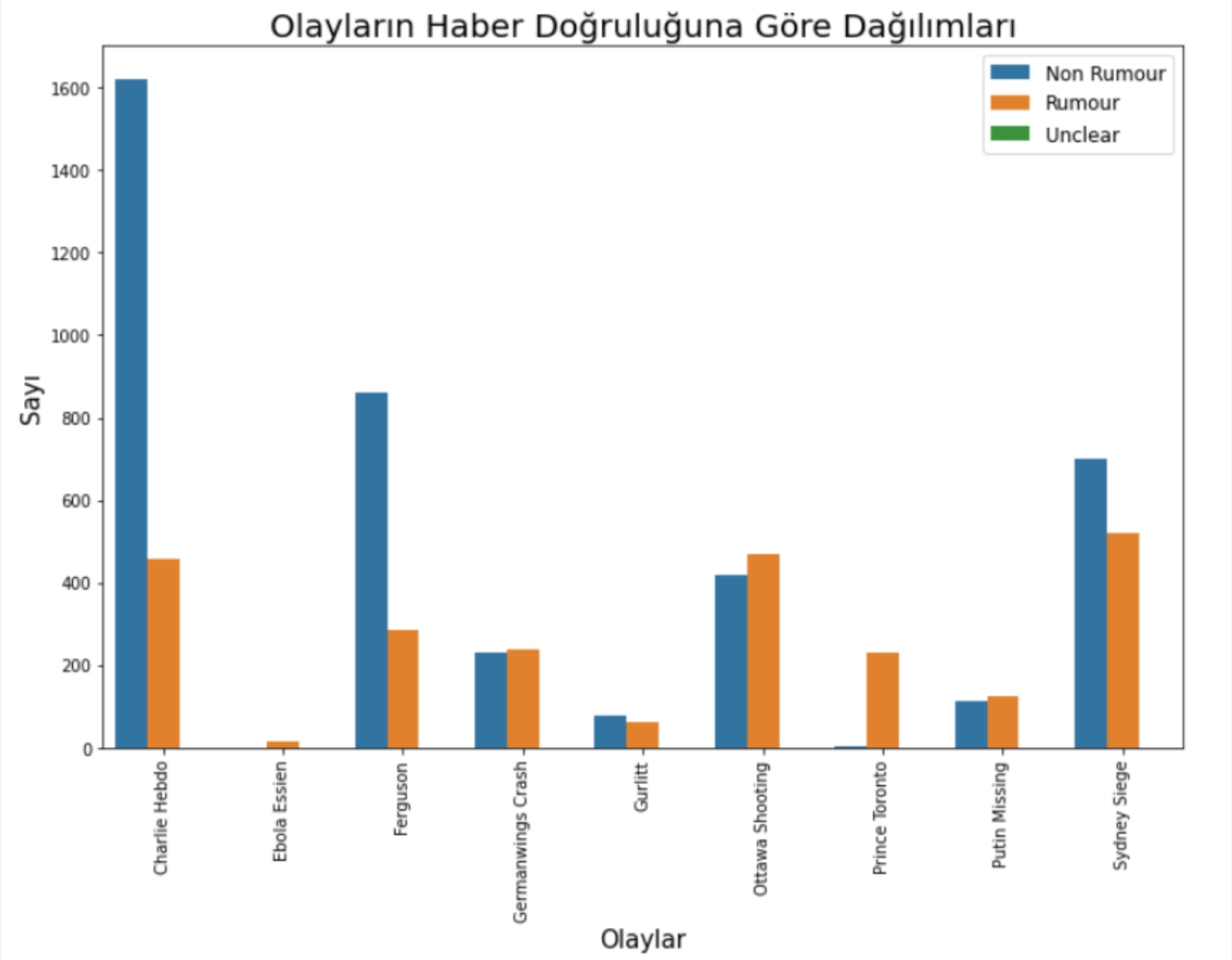


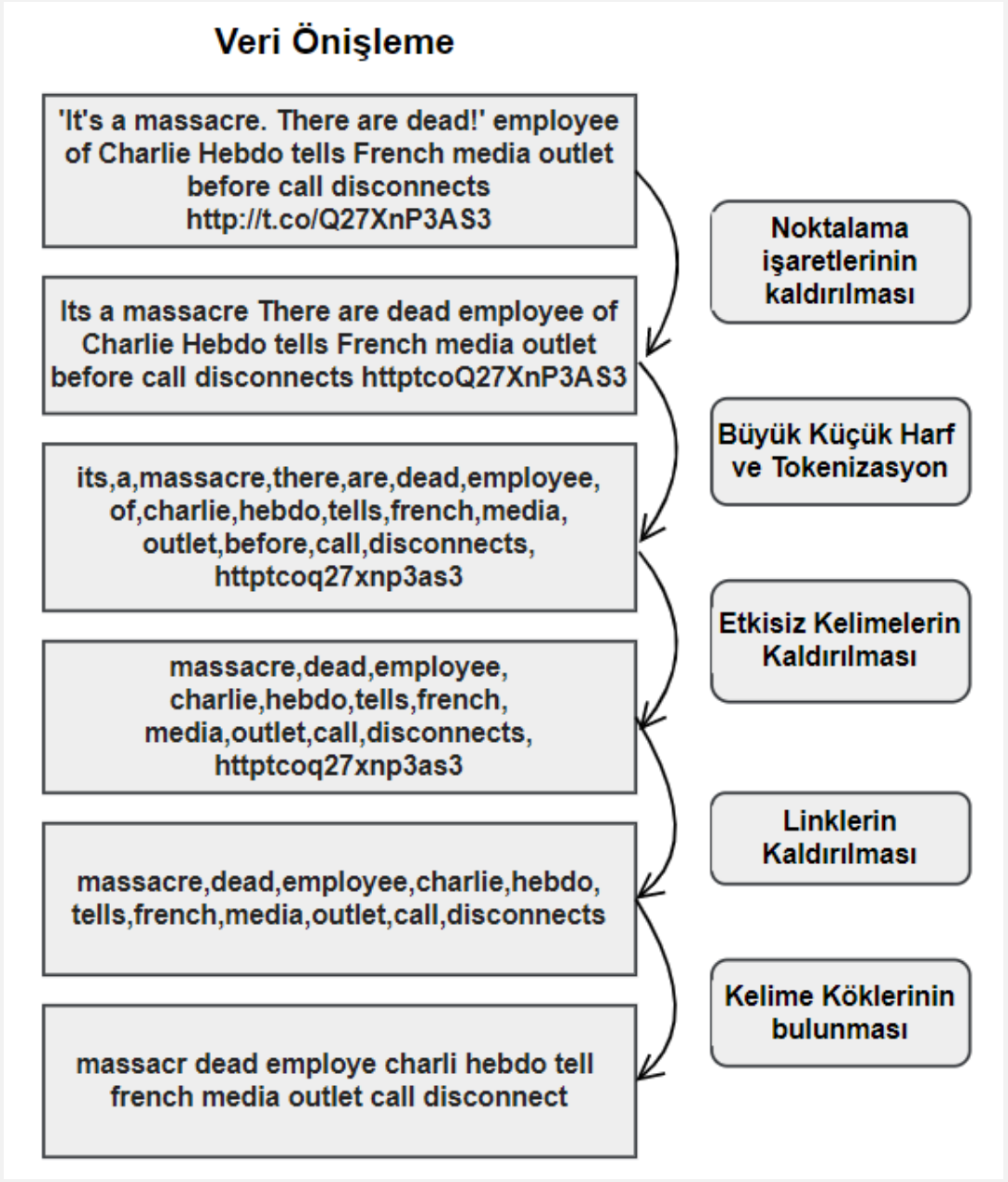


SAHTE GERÇEK HABER DAĞILIMLARI






Olayların Haber Doğruluğuna Göre Dağılımları
Non Rumour
Rumour
Unclear
Sayı
0
200
400
600
800
1000
1200
1400
1600
Charlie Hebdo
Ebola Essien
Ferguson
Germanwings Crash
Gurlitt
Ottawa Shooting
Prince Toronto
Putin Missing
Sydney Siege
Olaylar


Veri Önişleme
'It's a massacre. There are dead!' employee of Charlie Hebdo tells French media outlet before call disconnects http://t.co/Q27XnP3AS3
Noktalama işaretlerinin kaldırılması
Its a massacre There are dead employee of Charlie Hebdo tells French media outlet before call disconnects httptcoQ27XnP3AS3
Büyük Küçük Harf ve Tokenizasyon
its,a,massacre,there,are,dead,employee, of,charlie,hebdo,tells,french,media, outlet,before,call,disconnects, httptcoq27xnp3as3
Etkisiz Kelimelerin Kaldırılması
massacre,dead,employee, charlie,hebdo,tells,french, media,outlet,call,disconnects, httptcoq27xnp3as3
Linklerin Kaldırılması
massacre,dead,employee,charlie,hebdo, tells,french,media,outlet,call,disconnects
Kelime Köklerinin bulunması
massacr dead employe charli hebdo tell french media outlet call disconnect

WORD2VEC

Word2Vec temel olarak birer adet girdi, çıktı ve gizli katmandan oluşan bir yapay sinir ağıdır.

CBOW

- Merkezde bulunan kelime tahmin edilmeye çalışılır.
- Merkezde bulunmayan kelimeler girdi olarak alınır.
- Büyük veri kümelerinde daha iyi çalışır.

SKİP-GRAM

- Merkezde bulunmayan kelimeler tahmin edilmeye çalışılır.
- Merkezde bulunan kelimeler girdi olarak alınır.
- Küçük veri kümelerinde daha iyi çalışır.

Proje içerisinde Word2Vec modelinde CBOW kullanılmıştır.

FASTTEXT

- Word2Vec'in bir uzantısıdır.
- Word2Vec'ten hızlı çalışmaktadır.
- Kelimeleri tek tek işlemek yerine n-gram tekniği kullanılarak kelimeler bölünür.

COUNT VECTORIZER

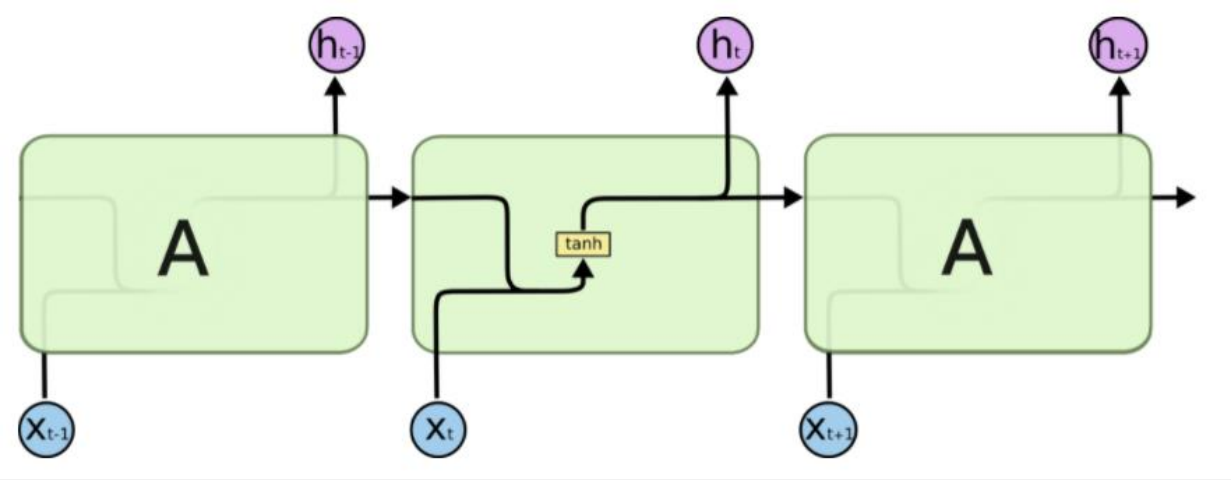
- One Hot Encoding ile benzer olarak her bir kelimenin cümle geçme sıklığı ile kelimeler temsil edilir.
- Oluşturulan matriste her bir sütun bir eşsiz kelimeye karşılık gelir.
- Matrisin boyutu veride bulunan benzersiz kelime sayısı ile verideki tweet sayısı kadardır.

HASHING VECTORIZER

- Count Vectorizer ile benzerdir.
- Aralarındaki fark kelimelerin metin içerisinde geçme sıklığına bir hash fonksiyonu uygulanmasıdır.
- Benzersiz kelimeler bellekte tutulmaz.

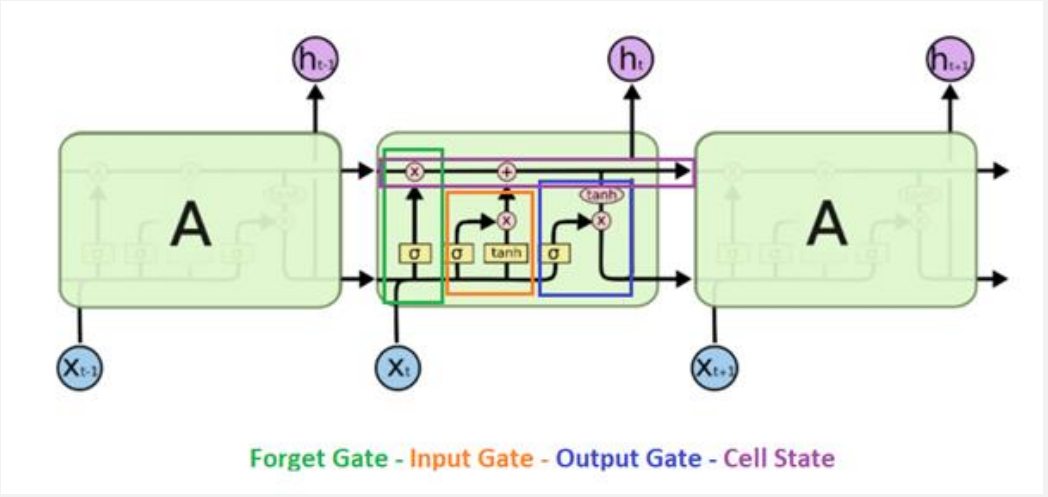
RNN

- NLP problemlerinde geçmiş verileri hatırladığı için genellikle iyi sonuçlar verir.
- RNN kısa vadeli bir hafızaya sahiptir.
- Standart bir RNN tek bir tanh katmanı içerir.
- RNN kısa vadeli hafızaya sahip olduğu için yetirince uzun cümlelerin olduğu bir girdi geldiğinde geçmiş verilerini hatırlamakta başarısız olur.



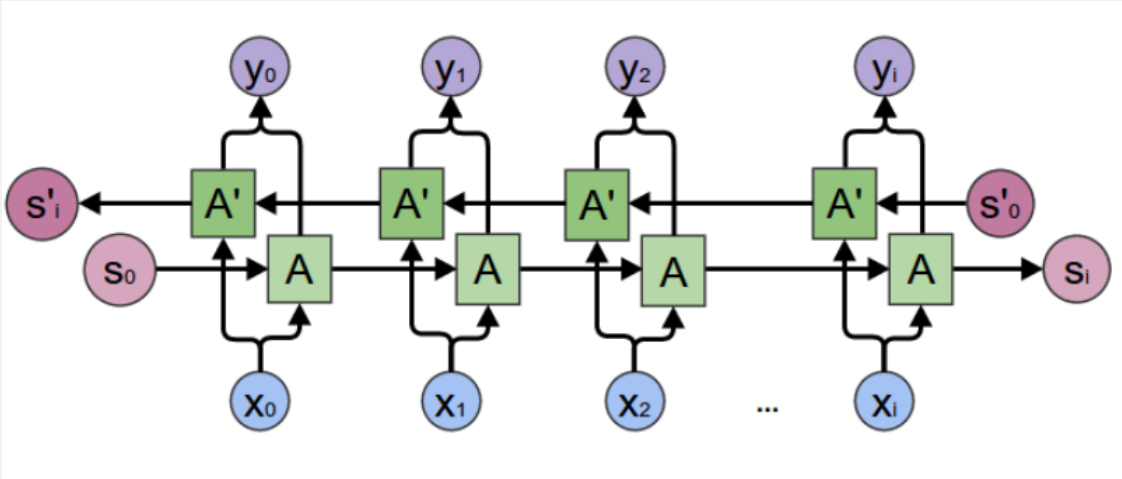
LSTM

- LSTM özel bir RNN türüdür.
- RNN'den farklı olarak uzun vadeli hafızaya sahiptir.
- 3 farklı katman içerir.
- Giriş Kapısı: Hücre durum güncellemesi yapar.
- Unutma Kapısı: Bilginin unutulup unutulmayacağına karar verir.
- Çıkış Kapısı: Hücrenin hangi kısımlarının çıktı olacağına karar verir.



BIDIRECTIONAL LSTM

- Bi-LSTM iki ayrı RNN bir araya getirmektedir.
- Aynı anda çift yönde ilerleme yaparlar.
- Çift yönlü işleme mantığı sayesinde hem önceki hem de sonraki zamana ait bilgilere sahip olabilmektedirler.



ARA YÜZ


tk
Twitter Platformunda Sahte Haber Tespiti
Tweet giriniz...
Tahmin
Sayfayı Temizle

SONUÇ

- Veriler eğitim ve test verisi olarak eğitilmeden önce 80,20 olacak şekilde ayrılmıştır.
- Metrik olarak doğruluk göz önüne alınmıştır.
- Eğitimler sonucu en yüksek eğitim doğruluk değerine Embedding katmanında FastText kullanılmış olan LSTM modeli sahiptir.
- En yüksek validasyon doğruluk değerine Embedding katmanında Word2Vec kullanılmış olan Bidirectional LSTM modeli sahiptir

Name	Acc	Loss	Val_Acc
bidirect_lstm_w2v	0.974	0.075	0.857
dropout_w2v	0.974	0.072	0.855
count_vec	0.940	0.167	0.855
lstm_fast	0.979	0.060	0.851
dropout_fast	0.907	0.228	0.843
bidirect_lstm_fast	0.975	0.070	0.840
lstm_w2v	0.710	0.575	0.715
hash_vec	0.650	0.624	0.626

Dinlediğiniz için teşekkür ederiz.