Makine Öğrenmesi Proje Ara Raporu Sercan Sözen

21501071

*Proje Özeti*

Proje Adı : Natural Language Processing with Disaster Tweets

Projenin amacı : Hangi tweet’lerin gerçek felaketler hakkında olup hangilerinin gerçek olmadığını tespit etmek

*Veri Seti*

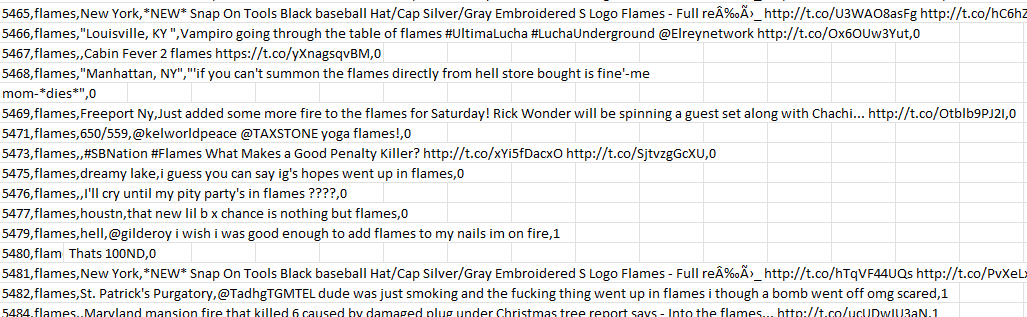
Eğitim verisetinin ayıklanmamış hali 8545 satır içermektedir. Özellik olarak 5 sütun vardır.Bunlar id, keyword, location, text ve target’tir.

Keyword sütunu tweet içinde geçen ve muhtemelen bir felaketi ifade edebilecek sözcüklerin toplamı olan bir sözlüktür. (apocalypse, ambulance, accident, attack, bomb, burning vs.)

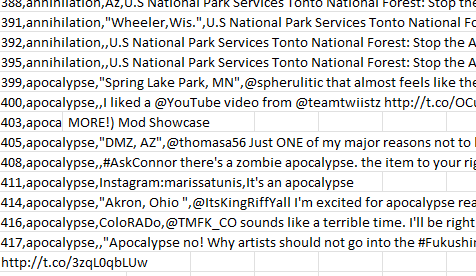
Text sütunu tweetleri barındırır.

Target sütunu ise tweet’in hangi sınıfa dahil olduğunu ifade eder. 1 değerinde ise gerçek bir felaket hakkında atılan bir tweet’tir.

Test veri seti ise 3689 ayıklanmamış satır içerir. Özellik olarak 4 sütundan oluşur. Bunlar id, keyword, location ve text’tir.



Train Dataset



Test Dataset

*https://ieeexplore.ieee.org/document/9136918*

*https://ieeexplore.ieee.org/document/8864964*

*https://ieeexplore.ieee.org/document/9617586*

*https://ieeexplore.ieee.org/document/9536745*

*https://ieeexplore.ieee.org/document/8129679*

***https://link.springer.com/chapter/10.1007/978-3-030-57805-3\_22***

***https://link.springer.com/article/10.1007/s11042-020-10183-2#Sec12***

*Yöntem*

Verinin önişlenlemsi : NLTK (natural language tool kit) ve string benzeri kütüphaneler kullanılarak veri setleri; noktalama işaretleri, stopwordler ve hmtl etiketleri gibi gereksiz kelimelerden ve boş satırlardan ayıklanacak.

Ayıklanan veriden oluşan dataframe, Countvectorizor (bag of words) ve TF-IDF vectorizor fonksiyonları ile tabloya dökülecek. Uygulanacak model bu iki vektörizasyon yöntemiyle test edilecek. Başarıları kıyaslanacak.

*Validasyon:* Eldeki veri halihazırda train-test şeklinde bölünmüş fakat ek olarak train veriseti kullanılarak k-fold cross validation (k=1,5,10) yöntemi ile yeni bir test veri seti için bölümleme yapılacak.

*Kullanılacak sınıflandırıcılar :*

-Naive Bayes

Gaussian NB ve multinomial NB

-DecisionTree

-KNN classification

KNN sınıflandırması için K değerleri (1-31) aralığında denenerek en uyumlu K değeri alınacak

-Logistic Regression

TF-IDF ve bag of words yöntemleriyle başarı kıyaslanacak.

-BERT

Tweetlerdeki kelimelerin kullanılma bağlamı da önemlidir. Önceden eğitilmiş bir derin öğrenme algoritması BERT, kelimeleri bağlamlarını (context) göz önünde bulundurarak birleştirebilir. Her kelime için ayrı bir vektör oluşturup vekörlerdeki feature’ların değerlerinin benzerliğini kıyaslayarak iki ayrı kelimeyi eşleştirebilir.

*Başarı ölçümü :*

Başarı ölçümü için F-measure (F1) kullanılacak. Buna ek olarak accuracy değerleri de kıyaslanacak.

Referanslar

[Natural Language Processing with Disaster Tweets](https://www.kaggle.com/c/nlp-getting-started)

[GaussianNB or MultinomialNB - Gowtham CH](https://towardsdatascience.com/how-i-was-using-naive-bayes-incorrectly-till-now-part-1-4ed2a7e2212b)

[Illustrated BERT - Jay Alammar](http://jalammar.github.io/illustrated-bert/)

[Logistic Regression and it’s applications in Natural Language Processing - Rithwik Kukunuri](https://medium.com/@rithwikkukunuri30/logistic-regression-and-its-applications-in-natural-language-processing-5f835c901fa7)