CSE 484 NATURAL LANGUAGE PROCESSING
HOMEWORK 3: TURKISH CLASSIFIER

ÇAĞRI ÇAYCI
1901042629

1. ABSTRACT

Grammar errors are among the most common mistakes, and they can sometimes lead to the incorrect interpretation of a sentence. One significant example of this occurs with the suffixes '-de' and '-ki' in Turkish. The '-de' suffix in Turkish serves to indicate place or functions as a conjunction. The distinction between these two uses lies in their formatting: in the former case, it is written adjacent to the preceding word, while in the latter case, it is written separately. This project aims to develop a neural network classifier that determines the appropriate placement of '-de' and '-ki' suffixes within sentences. Building a neural network classifier for this grammar rule involves six key steps: data collection, data preprocessing, model definition, training, evaluation, and testing.

2. DATA COLLECTION

To construct a classifier, a substantial dataset must be gathered. After researching available options, the Wikipedia Turkish dump emerged as the most suitable dataset for this purpose. It comprises 4.5 million lines, with each line containing multiple sentences. This vast dataset provides a sufficient number of data samples for this project.

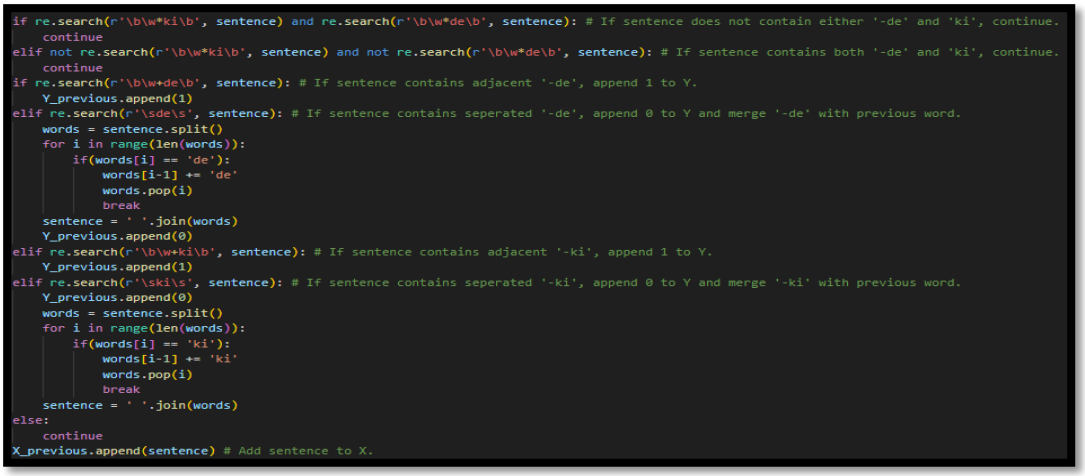
3. DATA PREPROCESSING

a. ADDING SENTENCES

Once all sentences are extracted from the Wikipedia dump, they need to undergo preprocessing before being utilized. Each sentence must meet exactly one of the following conditions to be added to our dataset:

- The sentence contains only one word with the '-de' suffix.
- The sentence contains 'de' solely as a conjunction.
- The sentence contains only one word with the '-ki' suffix.
- The sentence contains 'ki' solely as a conjunction.

If the sentence meets the second or fourth criteria, the conjunction must be merged with the previous word before adding it to the dataset.

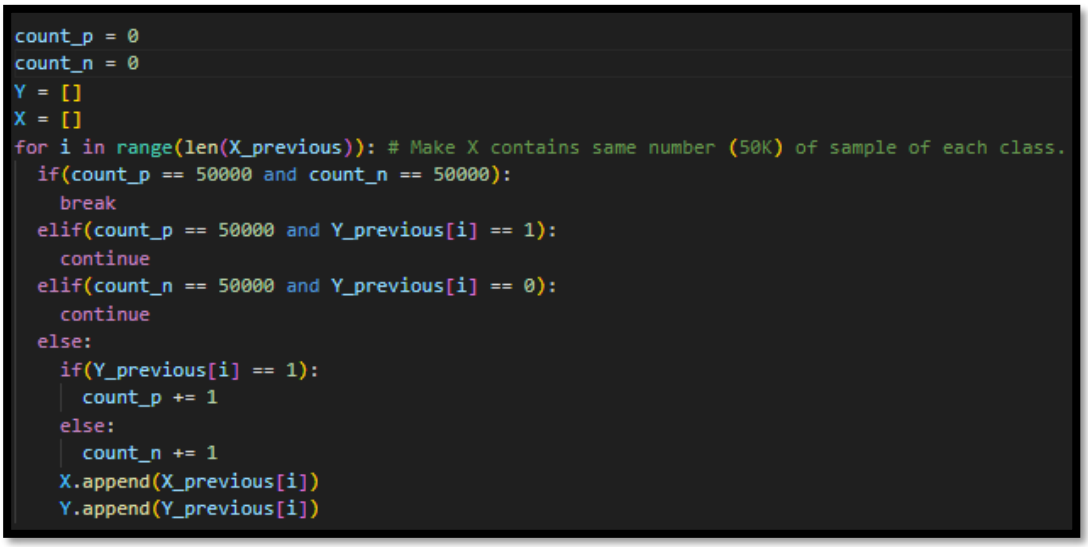


```
if re.search(r'\b\w*ki\b', sentence) and re.search(r'\b\w*de\b', sentence): # If sentence does not contain either '-de' and 'ki', continue.
    continue
elif not re.search(r'\b\w*ki\b', sentence) and not re.search(r'\b\w*de\b', sentence): # If sentence contains both '-de' and 'ki', continue.
    continue
if re.search(r'\b\w+de\b', sentence): # If sentence contains adjacent '-de', append 1 to Y.
    Y_previous.append(1)
elif re.search(r'\sde\s', sentence): # If sentence contains seperated '-de', append 0 to Y and merge '-de' with previous word.
    words = sentence.split()
    for i in range(len(words)):
        if(words[i] == 'de'):
            words[i-1] += 'de'
            words.pop(i)
            break
    sentence = ' '.join(words)
    Y_previous.append(0)
elif re.search(r'\b\w+ki\b', sentence): # If sentence contains adjacent '-ki', append 1 to Y.
    Y_previous.append(1)
elif re.search(r'\ski\s', sentence): # If sentence contains seperated '-ki', append 0 to Y and merge '-ki' with previous word.
    Y_previous.append(0)
    words = sentence.split()
    for i in range(len(words)):
        if(words[i] == 'ki'):
            words[i-1] += 'ki'
            words.pop(i)
            break
    sentence = ' '.join(words)
else:
    continue
X_previous.append(sentence) # Add sentence to X.
```

(Figure 1)

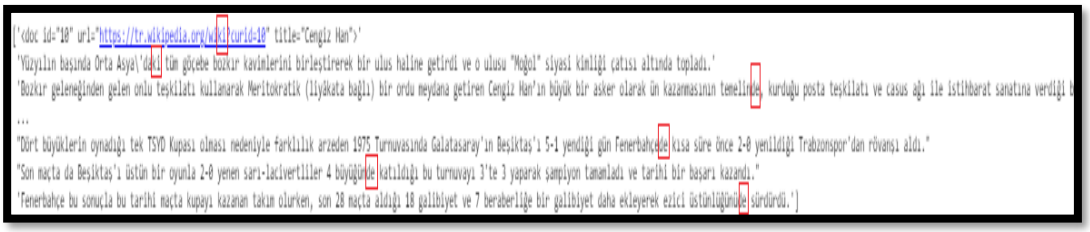
b. PREPARING DATASET

It is considered that having 50,000 data samples for each class (separated and not separated) is sufficient to build a classifier. Therefore, after the sentences are added to the dataset, only 100,000 data samples were retained.



```
count_p = 0
count_n = 0
Y = []
X = []
for i in range(len(X_previous)): # Make X contains same number (50K) of sample of each class.
  if(count_p == 50000 and count_n == 50000):
    break
  elif(count_p == 50000 and Y_previous[i] == 1):
    continue
  elif(count_n == 50000 and Y_previous[i] == 0):
    continue
  else:
    if(Y_previous[i] == 1):
      count_p += 1
    else:
      count_n += 1
    X.append(X_previous[i])
    Y.append(Y_previous[i])
```

(Figure 2)

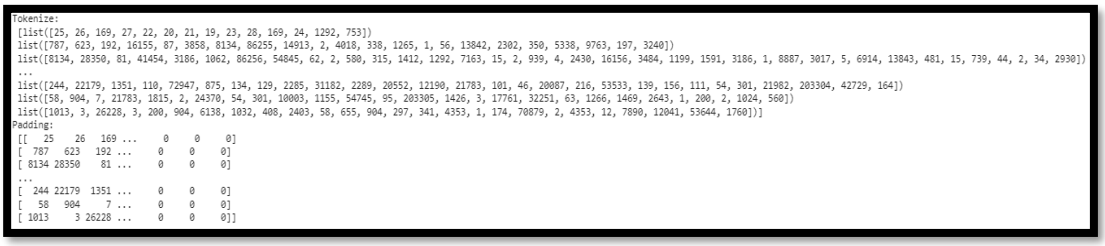


```
['<doc id="10" url="https://tr.wikipedia.org/wiki?curid=10" title="Cengiz Han">'
'Yüzyılın başında Orta Asya'daki tüm göçebe bozkır kavimlerini birleştirerek bir ulus haline getirdi ve o ulusu "Moğol" siyasi kimliği çatısı altında topladı.'
'Bozkır geleneğinden gelen onlu teşkilatı kullanarak Meritokratik (liyakata bağlı) bir ordu meydana getiren Cengiz Han'ın büyük bir asker olarak ün kazanmasının temelinde, kurduğu posta teşkilatı ve casus ağı ile istihbarat sanatına verdiği b
...
"Dört büyüklerin oynadığı tek TSYD Kupası olması nedeniyle farklılık arzeden 1975 Turnuvasında Galatasaray'ın Beşiktaş'ı 5-1 yendiği gün Fenerbahçe'ye kısa süre önce 2-0 yenildiği Trabzonspor'dan rövanşı aldı."
"Son maçta da Beşiktaş'ı üstün bir oyunla 2-0 yenen sarı-lacivertliler 4 büyüğünde katıldığı bu turnuvayı 3'te 3 yaparak şampiyon tamamladı ve tarihi bir başarı kazandı."
'Fenerbahçe bu sonuçla bu tarihi maçta kupayı kazanan takım olurken, son 28 maçta aldığı 18 galibiyet ve 7 beraberliğe bir galibiyet daha ekleyerek ezici üstünlüğünü sürdürdü.']
```

(Figure 3)

c. SENTENCE TOKENIZATION

It's essential to tokenize the sentences for input into the neural network. This process was achieved using the **Tokenizer from keras.preprocessing.text**. As illustrated in the following figure, the sentences were tokenized and sequentially padded. The primary objective of padding is to ensure that inputs are of fixed length. The padding functionality is provided by **pad_sequences from keras.preprocessing.sequence.**



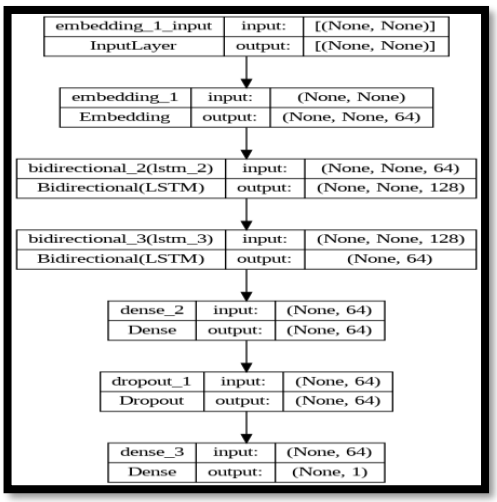
(Figure 4)

d. SHUFFLING DATASET

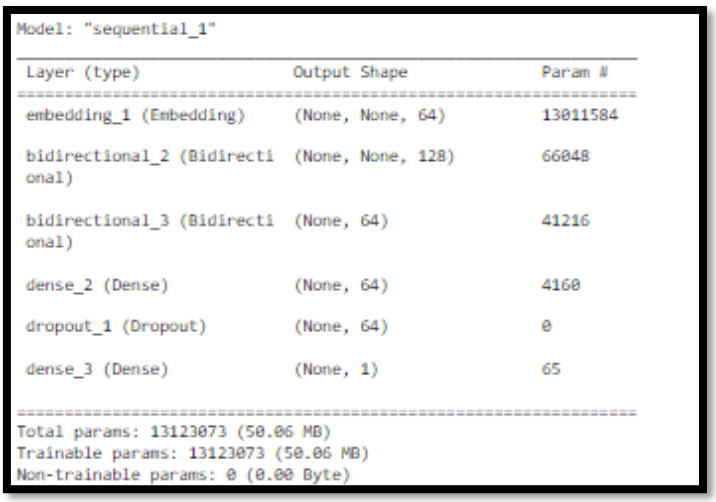
After the dataset preparation step, it was discovered that the data points belonging to one class were more numerous than the other class among the first 50,000 data points. To mitigate the impact of this situation, the dataset was shuffled by **shuffle from sklearn.utils**.

4. MODEL DEFINITION

To create a model, I conducted a literature review and experimented with various models. After testing, I selected the following model to use in the classifier.



(Figure 5)



(Figure 6)

Text data is inherently high-dimensional, sparse, and discrete. An embedding layer helps in transforming this data into dense, continuous vector representations where similar words or phrases are closer together in the embedding space. This transformation enables the model to learn more meaningful representations from the text data, capturing semantic and syntactic similarities.

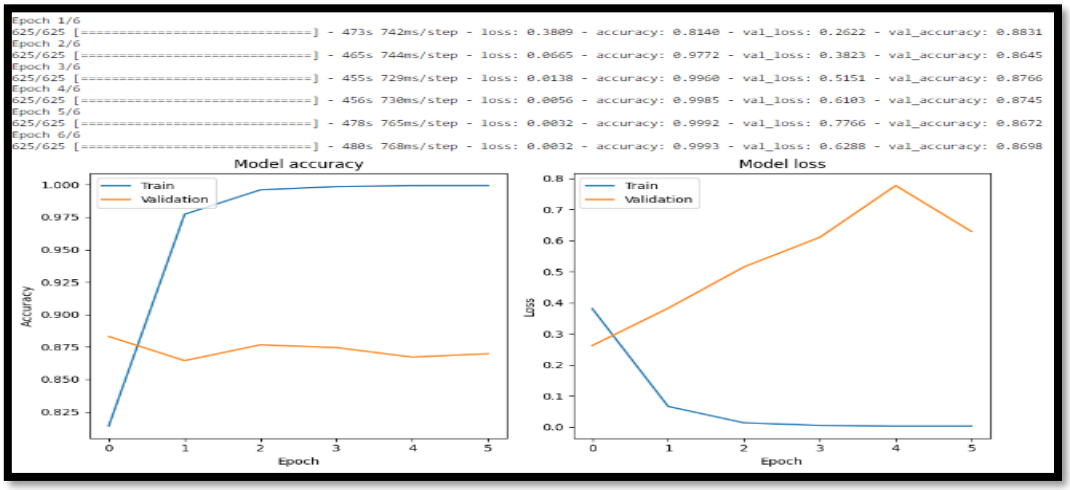
LSTMs are designed to handle sequential data, making them well-suited for processing text, which is inherently sequential in nature. They can capture dependencies and relationships between words across varying distances within a text, allowing them to understand context and meaning more effectively.

Two Dense layers are added to the model. The first Dense layer has 64 units and uses the ReLU (Rectified Linear Activation) activation function, which introduces non-linearity to the model. The second Dense layer has 1 unit and uses the sigmoid activation function. This layer serves as the output layer for binary classification tasks, where the output value (between 0 and 1) represents the probability of the input belonging to the positive class.

A Dropout layer is added after the first Dense layer. Dropout is a regularization technique used to prevent overfitting by randomly setting a fraction of input units to zero during training. In this case, 50% of the input units are randomly dropped out during each training iteration.

Finally, the model is compiled using the Adam optimizer, binary cross-entropy loss function (suitable for binary classification problems), and accuracy as the evaluation metric.

5. TRAINING



(Figure 7)

The model is trained over 6 epochs with batches of size 64. Half of the dataset is allocated for training, and 20% of the training set is used for validation.

During training, the loss decreases significantly from 0.3809 to 0.0032 over the epochs, indicating that the model is effectively learning and improving its predictions on the training data. The training accuracy also increases substantially from 0.8140 to 0.9993, demonstrating that the model's predictions are becoming increasingly accurate on the training data.

However, the validation loss fluctuates and generally increases from 0.2622 to 0.6288, suggesting that the model's performance on unseen data deteriorates over the epochs. This indicates potential overfitting as the model becomes overly specialized to the training data. Despite this, the validation accuracy fluctuates but remains relatively stable around 0.86 to 0.88, indicating that the model's performance on unseen data is consistent but not improving significantly.

Overall, while the model achieves very high accuracy on the training data, its performance on the validation set suggests overfitting, as indicated by the increasing validation loss and relatively stable validation accuracy. This suggests that the model may benefit from regularization techniques or adjustments to improve its generalization performance.

6. TESTING

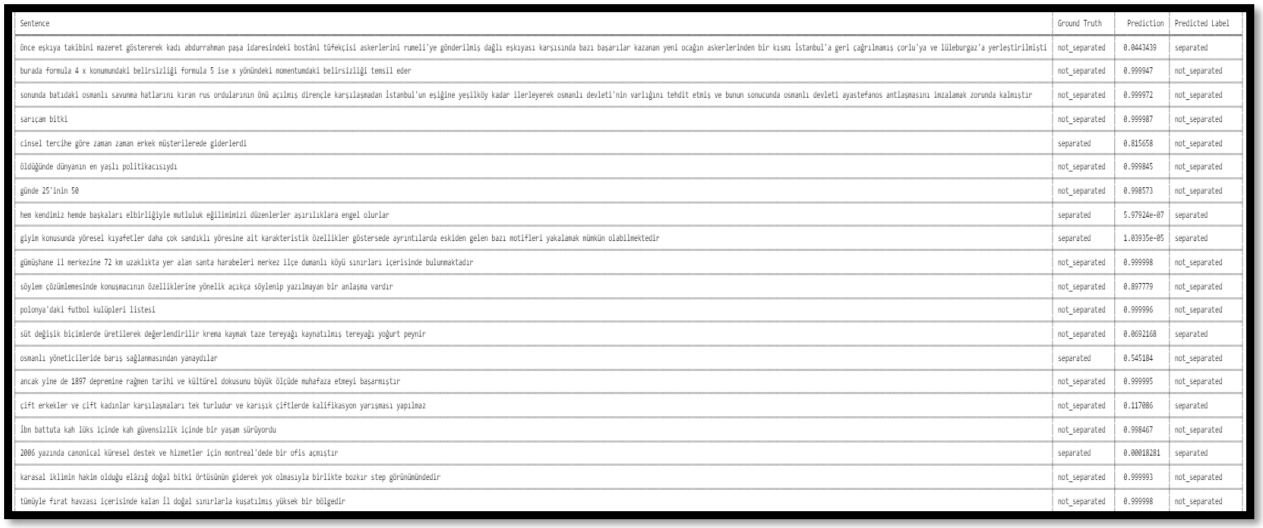
The model is tested on both unseen 50,000 data samples and seen 50,000 data samples. The results of this test are shown in the following figure.

DATA TYPE	ACCURACY	LOSS
UNSEEN	0.8699	0.629187
SEEN	0.97326	0.128456

(Figure 8)

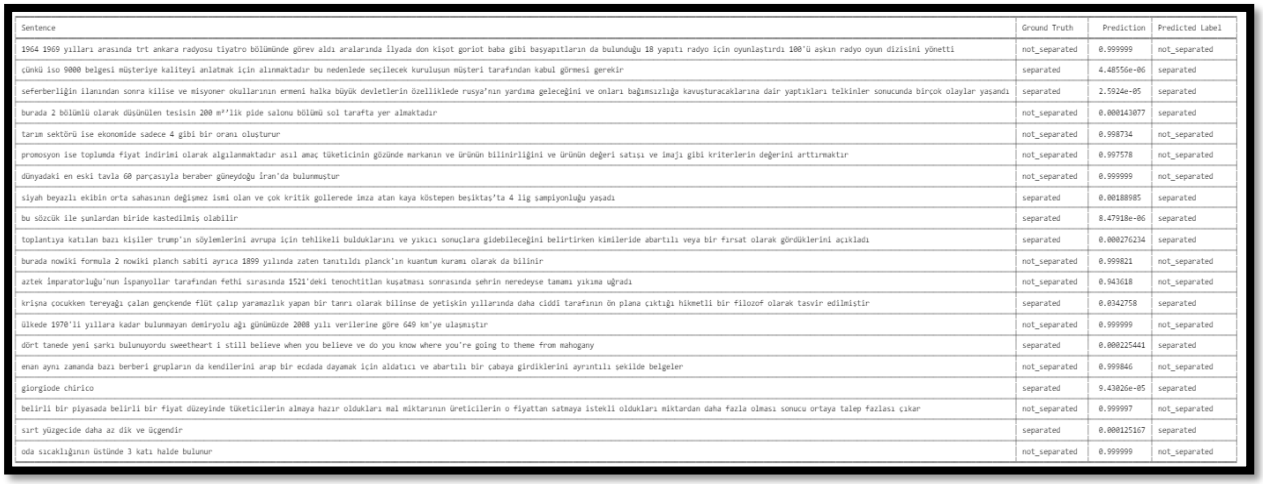
The system exhibits better performance on seen data samples, as expected. As a next step, additional regularization techniques such as Dropout, Early Stopping, etc., can be applied to the model to mitigate overfitting. These techniques help in improving the model's generalization ability, ensuring that it performs well on unseen data samples as well. These techniques may avoid overfitting and provide better results on unseen data samples.

Some unseen sentences along with their original and predicted labels are shown in the following figures.



Sentence	Ground Truth	Prediction	Predicted Label
önce eşkıya takibini mazeret göstererek kadı abdurrahman paşa idaresindeki bostâni tüfekçisi askerlerini rumeli'ye göndermiş dağlı eşkıyası karşısında bazı başarılar kazanan yeni ocağın askerlerinden bir kısmı İstanbul'a geri çağrılmamış çorlu'ya ve lüleburgaz'a yerleştirilmişti	not_separated	0.0443439	separated
burada formula 4 x konumundaki belirsizliği formula 5 ise x yönündeki momentumdaki belirsizliği temsil eder	not_separated	0.999947	not_separated
sonunda batıdaki osmanlı savunma hatlarını kıran rus ordularının önü açılmış dirençle karşılaşmadan İstanbul'un eşiğine yeşilköy kadar ilerleyerek osmanlı devleti'nin varlığını tehdit etmiş ve bunun sonucunda osmanlı devleti ayastefanos antlaşmasını imzalamak zorunda kalmıştır	not_separated	0.999972	not_separated
sarıçam bitki	not_separated	0.999987	not_separated
cinsel tercihe göre zaman zaman erkek müşterilerede giderlerdi	separated	0.815658	not_separated
öldüğünde dünyanın en yaşlı politikacısıydı	not_separated	0.999845	not_separated
günde 25'inin 50	not_separated	0.998573	not_separated
hem kendiniz hemde başkaları elbirliğiyle mutluluk eğitiminizi düzenlerler aşırılıklara engel olurlar	separated	5.97924e-07	separated
giyim konusunda yöresel kıyafetler daha çok sandıklı yöresine ait karakteristik özellikler göstersede ayrıntılarda eskiden gelen bazı motifleri yakalamak mümkün olabilmektedir	separated	1.03935e-05	separated
gümüşhane il merkezine 72 km uzaklıkta yer alan santa harabeleri merkez ilçe dumanlı köyü sınırları içerisinde bulunmaktadır	not_separated	0.999998	not_separated
söylem çözümlemesinde konuşmacının özelliklerine yönelik açıkça söylenip yazılmayan bir anlaşma vardır	not_separated	0.897779	not_separated
polonya'daki futbol kulüpleri listesi	not_separated	0.999996	not_separated
süt değişik biçimlerde üretilerek değerlendirilir krema kaymak taze tereyağı kaynatılmış tereyağı yoğurt peynir	not_separated	0.0692168	separated
osmanlı yöneticileride barış sağlanmasından yanaydılar	separated	0.545184	not_separated
ancak yine de 1897 depremine rağmen tarihi ve kültürel dokusunu büyük ölçüde muhafaza etmeyi başarmıştır	not_separated	0.999995	not_separated
çift erkekler ve çift kadınlar karşılaşmaları tek turludur ve karışık çiftlerde kalifikasyon yarışması yapılmaz	not_separated	0.117086	separated
İbn battuta kah lüks içinde kah güvensizlik içinde bir yaşam sürüyordu	not_separated	0.998467	not_separated
2006 yazında canonical küresel destek ve hizmetler için montreal'dede bir ofis açmıştır	separated	0.00018281	separated
karasal iklimin hakim olduğu elâzığ doğal bitki örtüsünün giderek yok olmasıyla birlikte bozkır step görünümündedir	not_separated	0.999993	not_separated
tümüyle fırat havzası içerisinde kalan il doğal sınırlarla kuşatılmış yüksek bir bölgedir	not_separated	0.999998	not_separated

(Figure 9)



Sentence	Ground Truth	Prediction	Predicted Label
1964 1969 yılları arasında trt ankara radyosu tiyatro bölümünde görev aldı aralarında İlyada don kişot goriot baba gibi başyapıtların da bulunduğu 18 yapıtı radyo için oyunlaştırdı 100'ü aşkın radyo oyun dizisini yönetti	not_separated	0.999999	not_separated
çünkü iso 9000 belgesi müşteriye kaliteyi anlatmak için alınmaktadır bu nedenlede seçilecek kuruluşun müşteri tarafından kabul görmesi gerekir	separated	4.48556e-06	separated
seferberliğin ilanından sonra kilise ve misyoner okullarının ermeni halka büyük devletlerin özelliklede rusya'nın yardıma geleceğini ve onları bağımsızlığa kavuşturacaklarına dair yaptıkları telkinler sonucunda birçok olaylar yaşandı	separated	2.5924e-05	separated
burada 2 bölümlü olarak düşünülen tesisin 200 m²'lik pide salonu bölümü sol tarafta yer almaktadır	not_separated	0.000143077	separated
tarım sektörü ise ekonomide sadece 4 gibi bir oranı oluşturur	not_separated	0.998734	not_separated
promosyon ise toplumda fiyat indirimi olarak algılanmaktadır asıl amaç tüketicinin gözünde markanın ve ürünün bilinirliğini ve ürünün değeri satışı ve imajı gibi kriterlerin değerini arttırmaktır	not_separated	0.997578	not_separated
dünyadaki en eski tavla 60 parçasıyla beraber güneydoğu İran'da bulunmuştur	not_separated	0.999999	not_separated
siyah beyazlı ekibin orta sahasının değişmez ismi olan ve çok kritik gollerede imza atan kaya köstepen beşiktaş'ta 4 lig şampiyonluğu yaşadı	separated	0.00188985	separated
bu sözcük ile şunlardan biride kastedilmiş olabilir	separated	8.47918e-06	separated
toplantıya katılan bazı kişiler trump'ın söylemlerini avrupa için tehlikeli bulduklarını ve yıkıcı sonuçlara gidebileceğini belirtirken kimileride abartılı veya bir fırsat olarak gördüklerini açıkladı	separated	0.000276234	separated
burada nowiki formula 2 nowiki planch sabiti ayrıca 1899 yılında zaten tanıtıldı planck'ın kuantum kuramı olarak da bilinir	not_separated	0.999821	not_separated
aztek İmparatorluğu'nun İspanyollar tarafından fethi sırasında 1521'deki tenochtitlan kuşatması sonrasında şehrin neredeyse tamamı yıkıma uğradı	not_separated	0.943618	not_separated
krişna çocukken tereyağı çalan gençkende flüt çalıp yaramazlık yapan bir tanrı olarak bilinse de yetişkin yıllarında daha ciddî tarafının ön plana çıktığı hikmetli bir filozof olarak tasvir edilmiştir	separated	0.0342758	separated
ülkede 1970'li yıllara kadar bulunmayan demiryolu ağı günümüzde 2008 yılı verilerine göre 649 km'ye ulaşmıştır	not_separated	0.999999	not_separated
dört tanede yeni şarkı bulunuyordu sweetheart i still believe when you believe ve do you know where you're going to theme from mahogany	separated	0.000225441	separated
enan aynı zamanda bazı berberi grupların da kendilerini arap bir ecdada dayamak için aldatıcı ve abartılı bir çabaya girdiklerini ayrıntılı şekilde belgeler	not_separated	0.999846	not_separated
giorgiode chirico	separated	9.43026e-05	separated
belirli bir piyasada belirli bir fiyat düzeyinde tüketicilerin almaya hazır oldukları mal miktarının üreticilerin o fiyattan satmaya istekli oldukları miktardan daha fazla olması sonucu ortaya talep fazlası çıkar	not_separated	0.999997	not_separated
sırt yüzgecide daha az dik ve üçgendir	separated	0.000125167	separated
oda sıcaklığının üstünde 3 katı halde bulunur	not_separated	0.999999	not_separated

(Figure 10)

7. APPENDIX

GOOGLE COLAB LINK FOR CODE