NEWS HEADLINE GENERATION WITH NLP

Hüma Bilgin
Computer Engineering Department
Yıldız Technical University, 34022 Istanbul, Turkey
l1118087@std.yildiz.edu.tr

***Özetçe* —Proje kapsamında NLP ile text generation işlemi yapılacaktır. Kullanılacak veri İngilizce haber başlıklarından oluşmaktadır. Bu haber başlıkları ile LSTM modeli eğitilerek haber başlığı üretilecektir.**

Anahtar Kelimeler—Doğal dil işleme, metin üretimi, haber başlığı üretimi

***Abstract*—Within the scope of the project, text generation will be done with NLP. The data to be used consists of news headlines in English. With these news headlines, the LSTM model will be trained and news headlines will be produced.**

Keywords—Natural language processing, text generation, news headline generation

I. Introduction

Natural language processing, which is one of the rapidly developing fields of our age, is constantly expanding its scope with new models and technologies. Natural Language Processing technology, which is used in many fields from language translation to voice recognition, from author detection to text classification, and which gives much better results as technologies for big data use develop, will be used in the field of Text Generation in this project. Within the scope of the project, it is aimed to produce synthetic news headlines. In classification problems, the unbalanced of the data used significantly affects the success of the classification. However, in the data sets that should be used, the classes often do not contain equal numbers of data. Various methods are used to synchronize this data. Reducing the data of classes with high data is not a preferred method. To solve this problem, performing synthetic data generation both prevents the data set from shrinking and makes the data balanced. The news headlines produced within the scope of the project are aimed to balance the unbalanced data set.

II. SYSTEM ANALYSIS

A. Neural Networks

The first thing to mention when starting to analyze the system is Neural Networks. Artificial Neural Networks, in other words, is a machine learning technique created to impose the human brain's abilities such as creating, generating and discovering new information on the computer. These systems are composed of Neurons as seen in Figure 4.1. and these Neurons are connected to each other by Synapses. The data flow proceeds from Input Neurons to Output Neurons. Each Synapse contains a coefficient. These coefficients determine the importance of the connections between neurons. The Addition Function calculates the net input of the charge by adding these multiplied coefficients. Finally, the Activation Function takes the weighted sum of all the inputs in the previous layer and generates the output value and sends it to the next layer (for example, ReLU or Sigmoid Functions). [1] The activation function used in the project is Softmax. This function is used for multiple classification problems. It makes a probabilistic interpretation by generating values between [0,1].

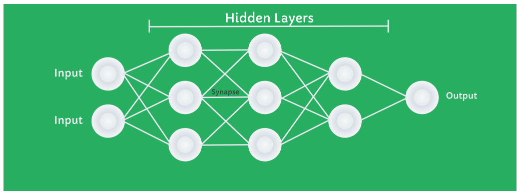


Figure 1 Neural Networks

B. Recurrent Neural Networks

Recurrent Neural Networks (RNN), on the other hand, are not dependent on the current input, unlike NN. In addition to the inputs at time t, the inputs at time t-1 also affect the output. Looking at this explanation, we can say that Recurrent networks have a memory. The reason why this structure is needed is that the input order makes sense in time-varying data such as writing and speaking.

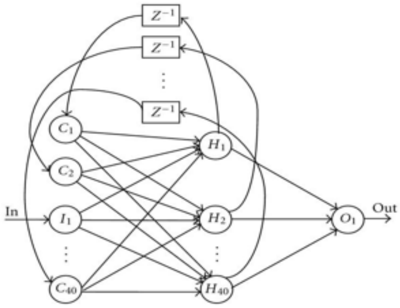


Figure 2 Recurrent Neural Networks

C. Long Short Term Memory

Although the LSTM structure is also an RNN structure, instead of a single NN layer as in RNN, it has four layers, which can be called gates, interconnected in a special way. Information flows from outside to this structure. It just doesn't use existing inputs. Through these gates, it is determined that the information coming from the outside will be subjected to the process or whether it will be received or not. The activation function passes or stops the information according to its weight, just like in RNN. These processes take place during the learning of the network.

III. SYSTEM DESIGN

Under the System Analysis Title, the model to be selected and the reason for choosing this model were explained. In this chapter, how this model will be implemented will be discussed. As it is known, there is a big difference between the languages that the computer can make sense of and the languages used in daily life. In order to carry out the Natural Language Processing steps, the languages we speak in daily life need to be translated into computer language. Translating words into computer language is called Word Embedding. This process can be performed in many different ways. Words can be interpreted for a computer by applying a variety of processes, from One Hot Encoding approach to 100-dimensional Word2Vec display. The method to be used in this project was chosen as OneHot Encoding.

A. OneHot Encoding

OneHot Encoding represents words and characters by encoding in binary. First of all, OneHot Encoding method will be explained with a simpler classification example.

Table 1 Car Prices

Car	Category Value	Price
Audi A4	1	200.000
Honda Civic	2	350.000
Ford Focus	3	280.000
Audi A4	1	460.000

Table-1 indicates some car brands and their prices. Now these values will be displayed with OneHot Encoding.

Table 2 Car Prices Display with One Hot Encoding

Audi A4	Honda Civic	Ford Focus	Price
1	0	0	200.000
0	1	0	350.000
0	0	0	280.000
1	0	1	460.000

As seen in Table-2, each variable is represented by a binary number of output quantity size. In this project, output is the number of characters in the trained data.

B. Data Generation

The dataset to be used for the project was created from a dataset available on kaggle. This data set includes news texts, titles of news texts, categories of news and some features of texts. Only news headlines and categories were extracted from the data set. Among 51 categories, the ’POLITICS’ category was used in model training, and 6 categories were used in the creation of the dataset to be used when measuring synthetic data performance. 4 automations were written to create datasets read from json file. The first of these is for the creation of the data set used in model training. 32,739 news headlines on ’POLITICS’ were shot. The second automation was written to generate balanced data. From the json file, 500 data per topic were drawn from the news headlines on ’POLITICS’, ’ENTERTAINMENT’, ’BUSINESS’, ’SPORTS’, ’TRAVEL’, ’WELLNESS’, and each data was divided into 6 different folders in a different txt file. The reason for this process is to comply with the format requested by the Text2arff application programmed by Yıldız Technical University. The third automation creates an unbalanced dataset by emptying 350 of the POLITICS themed data in the balanced dataset. The fourth automation complements this unbalanced data using data generated by natural language processing. Of this filled data, 350 of the POLITICS themed data consists of the data produced within the scope of the project, while 150 of them consist of the data drawn from the kaggle data set.

IV. Application

The implementation of the project consists of two stages. The first step is to display the predictions that emerged after the training of the created model. First, a weight file is created as a result of training the model. A corpus is obtained by combining news headlines one after another. Random parts are taken from this corpus and new titles are produced as a continuation of these parts called sequences.

The second stage is to measure the performance of these news headlines as synthetic data. This measurement is done by classifying three different data sets. First, a balanced data set consisting of six classes is created and this data set is classified via Weka. By reducing the data of a subject in this data set, the data set is unbalanced and the classification is done again. Finally, the unbalanced data set is filled with the generated news headlines and reclassified. Naive Bayes Multinominal method was used while making classifications.

V. Experimental Results

The experimental results of the project depend on the success of the news headlines used as synthetic data.

A. Produced News Headlines

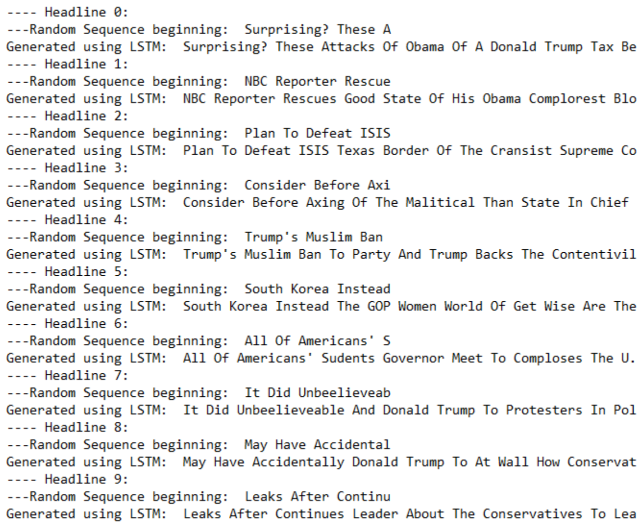
In the figure 3, news headlines on POLITICS produced with 60 epochs are seen.

B. Classification With Balanced Data Set

Figure 4 contains the classification results using Weka on the balanced data set. As can be seen from the figure, a 53.5667% success rate was obtained as a result of the classification.

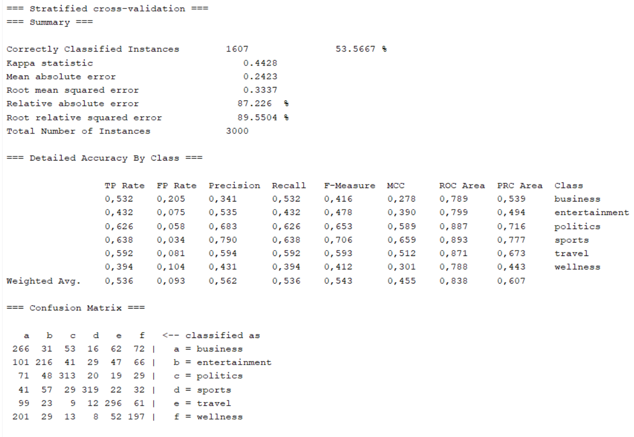
C. Classification with an Unbalanced Dataset

Figure 5 shows the results of the classification made with the unbalanced data set. This time the classification performance decreased to 46.3667%.



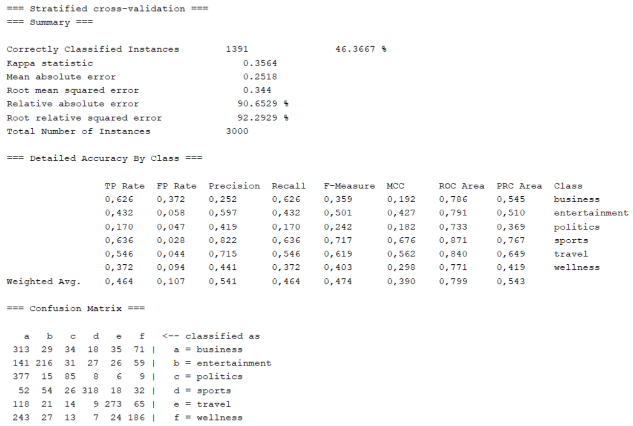
```
---- Headline 0:
---Random Sequence beginning:  Surprising? These A
Generated using LSTM:  Surprising? These Attacks Of Obama Of A Donald Trump Tax Be
---- Headline 1:
---Random Sequence beginning:  NBC Reporter Rescue
Generated using LSTM:  NBC Reporter Rescues Good State Of His Obama Complorest Blo
---- Headline 2:
---Random Sequence beginning:  Plan To Defeat ISIS
Generated using LSTM:  Plan To Defeat ISIS Texas Border Of The Cransist Supreme Co
---- Headline 3:
---Random Sequence beginning:  Consider Before Axi
Generated using LSTM:  Consider Before Axing Of The Malitical Than State In Chief
---- Headline 4:
---Random Sequence beginning:  Trump's Muslim Ban
Generated using LSTM:  Trump's Muslim Ban To Party And Trump Backs The Contentivil
---- Headline 5:
---Random Sequence beginning:  South Korea Instead
Generated using LSTM:  South Korea Instead The GOP Women World Of Get Wise Are The
---- Headline 6:
---Random Sequence beginning:  All Of Americans' S
Generated using LSTM:  All Of Americans' Sudents Governor Meet To Comploses The U.
---- Headline 7:
---Random Sequence beginning:  It Did Unbeelieveab
Generated using LSTM:  It Did Unbeelieveable And Donald Trump To Protesters In Pol
---- Headline 8:
---Random Sequence beginning:  May Have Accidental
Generated using LSTM:  May Have Accidentally Donald Trump To At Wall How Conservat
---- Headline 9:
---Random Sequence beginning:  Leaks After Continu
Generated using LSTM:  Leaks After Continues Leader About The Conservatives To Lea
```

Figure 3 Produced News Headlines



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1607               53.5667 %
Kappa statistic                          0.4428
Mean absolute error                      0.2423
Root mean squared error                  0.3337
Relative absolute error                 87.226  %
Root relative squared error             89.5504 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,532    0,205    0,341      0,532    0,416      0,278    0,789     0,539     business
                 0,432    0,075    0,535      0,432    0,478      0,390    0,799     0,494     entertainment
                 0,626    0,058    0,683      0,626    0,653      0,589    0,887     0,716     politics
                 0,638    0,034    0,790      0,638    0,706      0,659    0,893     0,777     sports
                 0,592    0,081    0,594      0,592    0,593      0,512    0,871     0,673     travel
                 0,394    0,104    0,431      0,394    0,412      0,301    0,788     0,443     wellness
Weighted Avg.    0,536    0,093    0,562      0,536    0,543      0,455    0,838     0,607

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 266  31  53  16  62  72 |   a = business
 101 216  41  29  47  66 |   b = entertainment
  71  48 313  20  19  29 |   c = politics
  41  57  29 319  22  32 |   d = sports
  99  23   9  12 296  61 |   e = travel
 201  29  13   8  52 197 |   f = wellness
```

Figure 4 Classification With Balanced Data Set

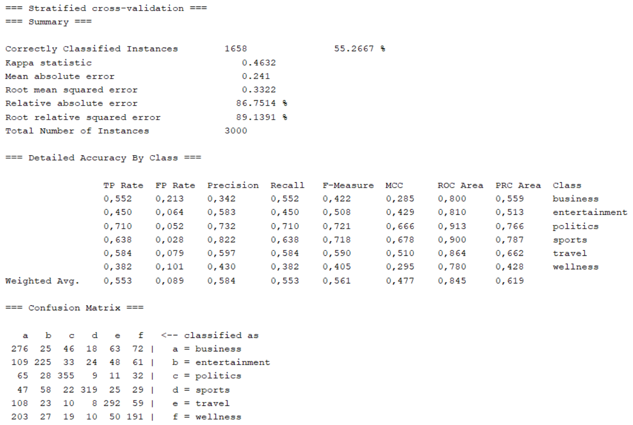


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1391               46.3667 %
Kappa statistic                          0.3564
Mean absolute error                      0.2518
Root mean squared error                  0.344
Relative absolute error                 90.6529 %
Root relative squared error             92.2929 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,626    0,372    0,252      0,626    0,359      0,192    0,786     0,545     business
                 0,432    0,058    0,597      0,432    0,501      0,427    0,791     0,510     entertainment
                 0,170    0,047    0,419      0,170    0,242      0,182    0,733     0,369     politics
                 0,636    0,028    0,822      0,636    0,717      0,676    0,871     0,767     sports
                 0,546    0,044    0,715      0,546    0,619      0,562    0,840     0,649     travel
                 0,372    0,094    0,441      0,372    0,403      0,298    0,771     0,419     wellness
Weighted Avg.    0,464    0,107    0,541      0,464    0,474      0,390    0,799     0,543

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 313  29  34  18  35  71 |   a = business
 141 216  31  27  26  59 |   b = entertainment
 377  15  85   8   6   9 |   c = politics
  52  54  26 318  18  32 |   d = sports
 118  21  14   9 273  65 |   e = travel
 243  27  13   7  24 186 |   f = wellness
```

Figure 5 Classification with an Unbalanced Dataset

D. Classification by Filled Dataset

Figure 6 contains the results of the classification made with the filled data set. The performance has increased to 55.2667%, exceeding the accuracy value provided in balanced data.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1658               55.2667 %
Kappa statistic                          0.4632
Mean absolute error                      0.241
Root mean squared error                  0.3322
Relative absolute error                 86.7514 %
Root relative squared error             89.1391 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,552    0,213    0,342      0,552    0,422      0,285    0,800     0,559     business
                 0,450    0,064    0,583      0,450    0,508      0,429    0,810     0,513     entertainment
                 0,710    0,052    0,732      0,710    0,721      0,666    0,913     0,766     politics
                 0,638    0,028    0,822      0,638    0,718      0,678    0,900     0,787     sports
                 0,584    0,079    0,597      0,584    0,590      0,510    0,864     0,662     travel
                 0,382    0,101    0,430      0,382    0,405      0,295    0,780     0,428     wellness
Weighted Avg.    0,553    0,089    0,584      0,553    0,561      0,477    0,845     0,619

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 276  25  46  18  63  72 |   a = business
 109 225  33  24  48  61 |   b = entertainment
  65  28 355   9  11  32 |   c = politics
  47  58  22 319  25  29 |   d = sports
 108  23  10   8 292  59 |   e = travel
 203  27  19  10  50 191 |   f = wellness
```

Figure 6 Classification by Filled Dataset

1) Epoch 10:

VI. Performance Analysis

The performance of synthetic news headline production has changed according to the encoding method used, the values of the variables used while creating the LSTM model, and the size of the data set. For the measurement of this performance, classification processes were carried out with the help of Weka on 3 different data sets. First, classification was made in a balanced data set, and the success rate was measured as 53.5667%. When the data set is destabilized by reducing the data of a class from this data set, the success of the classification was measured as 46.3667%. When the reduced data set was filled with the news headlines and a balanced data set was obtained again, the classification success increased to 55.2667%.

It is seen that the synthetic data produced based on these data increases the success by 9%. This value will vary depending on values such as the size of the data sets and the classification method used.

DataSet	**Success**
Balanced	53.5667%
Unbalanced	46.3667%
Filled	55.2667%

VII. Conclusion

As a result of the evaluation of the classification ratios, the model created within the scope of the project was able to produce successful synthetic data that could correct the imbalance in classification problems. The produced model can be used not only to produce news headlines, but also to produce various synthetic data. In addition, the data set can be selected in both English and Turkish.

References

[1] S. Aylak. (1999) Neural network nedir? nasıl Çalışır. [Online]. Available: