Ertugrul Kasikci 200104004097

CSE484 HW4

Introduction

This project focuses on developing a neural network-based model to classify sentences in Turkish for the correct usage of the "de/da" and "ki" suffixes. The aim is to determine whether these suffixes are used as separate words or attached to the preceding word in each sentence.

Methodology

Data Collection

Data was sourced from the 'wiki_00' file in the Turkish Wikipedia Dump, chosen for its diversity in sentence structures and language usage, providing a comprehensive training dataset.

Labeling and Dataset Balancing

In this project, sentences were labeled as having separated or unseparated "de/da" and "ki" suffixes using regular expressions. Due to an initial imbalance between the two categories, the dataset is balanced by down sampling the more numerous unseparated sentences. This step ensured an equal representation of both separated and unseparated suffixes, crucial for training the neural network effectively and avoiding bias towards the more prevalent class.

Data Preprocessing

Sentence Tokenization: Implemented using NLTK's sent_tokenize to split the text into sentences.

Text Cleaning: Each sentence was converted to lowercase and stripped of punctuation to reduce complexity and emphasize structural language aspects.

Feature Extraction

Word2Vec: Utilized for sentence vectorization, chosen over FastText for its efficiency and effectiveness in capturing semantic word relationships essential for understanding context-dependent suffix usage.

Model Details

Architecture

The neural network is constructed as a Sequential model, which consists of the following layers arranged linearly:

- **First Dense Layer:** Comprises 128 neurons and utilizes the ReLU (Rectified Linear Unit) activation function. This layer is designed to introduce non-linearity, allowing the model to learn complex patterns.
- **Dropout Layer:** Implements a 50% dropout rate for regularization. This layer randomly sets a portion of the input units to zero, helping to prevent overfitting and ensuring that the model does not rely too heavily on any single feature.
- **Second Dense Layer:** Contains 64 neurons and also uses the ReLU activation function. This layer further processes the features extracted by the previous layers, refining the information flow towards the output layer.
- **Output Layer:** A single neuron with a Sigmoid activation function. This final layer outputs a probability indicating the likelihood of a sentence belonging to one of the two classes (separated or unseparated suffix usage). The Sigmoid function is ideal for binary classification as it maps the output to a value between 0 and 1, representing the probability of a particular class.

Compilation and Training

- **Optimizer:** Adam.
- **Loss Function:** Binary cross-entropy.
- **Training:** Model trained on an 80/20 train-test split, 10 epochs, batch size of 32.

Performance Evaluation

The model's performance was assessed using several key metrics, reflecting its accuracy and reliability in classifying Turkish suffixes "de/da" and "ki":

- **Accuracy (92.86%):** Indicates the overall proportion of correctly classified instances, showcasing the model's effectiveness in general classification.

- **Precision (0.88):** Reflects the accuracy of positive predictions, implying that 88% of the model's predictions of separated suffixes are correct.
- **Recall (0.99):** Measures the model's ability to identify all relevant instances of separated suffixes, indicating it successfully detects 99% of true cases.
- **F1 Score (0.93):** The harmonic mean of precision and recall, demonstrating a balanced performance between capturing true positives and avoiding false positives.

These results suggest a high level of accuracy and reliability of the model in distinguishing between separated and unseparated suffixes in Turkish sentences, with a particularly strong ability to identify true cases of separated suffixes (high recall) while maintaining a good level of precision.

Test Run Results

The model was tested with 20 sentences to evaluate its real-world performance. The sentences and the corresponding model predictions are as follows:

1. Sentence: Bugün parkta bir yürüyüş yaptık. Predicted Label: 0
2. Sentence: Kitap masanın üstünde duruyordu. Predicted Label: 0
3. Sentence: Pencereden bakan çocuk da mutlu görünüyordu. Predicted Label: 1
4. Sentence: Yarınki toplantıya katılamayacağım. Predicted Label: 0
5. Sentence: Bu iş tam da bana göre. Predicted Label: 1
6. Sentence: Kediler de insanlar gibi duygusal olabilir. Predicted Label: 1
7. Sentence: Olay yerindeki deliller incelendi. Predicted Label: 0
8. Sentence: Anladığım kadarıyla ders çok zormuş ki düşük not almışlar. Predicted Label: 1
9. Sentence: Köpeğim dün gece de çok havladı. Predicted Label: 1
10. Sentence: Bu konudaki düşüncelerini merak ediyorum. Predicted Label: 0
11. Sentence: Evdeki hesap çarşıya uymaz. Predicted Label: 0
12. Sentence: Yeni aldığın ayakkabılar çok mu rahat ki? Predicted Label: 1
13. Sentence: Arkadaşımla dün sinemada vakit geçirdik. Predicted Label: 0
14. Sentence: Tatilde deniz kenarında bir evde kaldık. Predicted Label: 0
15. Sentence: Okuldaki öğretmenler çok iyiydi. Predicted Label: 0
16. Sentence: Dün akşamki yemeğin tadı hâlâ damağımda. Predicted Label: 0
17. Sentence: Yazın ortasında da kar yağdı. Predicted Label: 1
18. Sentence: Bu akşamki konseri kaçırmak istemiyorum. Predicted Label: 0
19. Sentence: Dünkü maçta çok heyecanlandım. Predicted Label: 0
20. Sentence: Kitaplıktaki kitaplar da tozlanmış. Predicted Label: 1