# Language Detection via a Neural Network

Maddie Bauer

Bellevue University

# Table of Contents

# The Problem

- In today's world, organizations deal with a mass amount & wide variety of data
  - Phone calls & messages
  - Emails
  - Online Reviews
  - Social Media
  - Mobile Applications
  - Etc.
- One of the core skills in extracting information from **text data** is called Natural Language Processing (NLP)
- Language detection is often an important first step within any NLP problem as it's necessary to know the language of text data in order to interpret it accurately
- The goal of this presentation is to predict the natural language of a word, sentence or other piece of text by creating a neural network

Source: https://www.analyticsvidhya.com/blog/2017/10/essential-nlp-guide-data-scientists-top-10-nlp-tasks/

# The Problem Continued...

- Language Detection is important because it is necessary to know the language of the given text before any other actions, such as translating, sentiment analysis or autocompleting, can occur.

- Examples:

  - Spell-Checking feature must first determine what language the text is being written in before it can provide you with the correct spelling of a word

  - Applying the right subtitles or closed captioning to a presentation or video

  - Routing Emails to the right geographically located customer service department for a large international company

# The Dataset

- From Tatoeba.org - A free online database with example words and sentences geared towards foreign language learners
- 9,714,915 pieces of text data in 328 different languages
- I took a subset of this dataset and used only 5 languages
  - English, Spanish, German, French and Italian
  - 50,000 text entries for each
  - 250,000 total in resulting dataset

| | lang | text |
|---|---|---|
| **4264763** | eng | A concept is an idea that can be represented b... |
| **9099072** | eng | Please dont tell me youre serious |
| **8334374** | eng | Tom was one of those present |
| **3627522** | eng | Is the richest country in the European Union r... |
| **1898133** | eng | They want to have a meeting with you |
| **8145212** | eng | Algeria has no autonomous provinces |
| **7681743** | eng | This stance justifies several points of view |
| **4910945** | eng | Try harder to find the answer there must be one |
| **7320987** | eng | Sami called me six days later |
| **3448569** | eng | The back of Toms right hand was injured by a s... |

# Method Used

- Split Dataset
  - Training (70%), Testing (10%) & Validation (20%) Sets

- Feature Engineering
  - Trigrams
    - Example: “hello there” trigrams include: hel, ell, llo, lo_, o_t, _th, her, ere, re_ and so on.
  - Encoded each language to a numeric value

- Build Neural Network
  - Keras Library in Python

# What is Keras?

- High-level API written in Python for deep learning purposes.
- Keras runs on top of the machine learning platform, TensorFlow, and was created to enable fast experimentation and implementation of neural networks.
- Great for beginners, like me!
- Widely used today by academic researchers to engineers at large companies


Image Sources
Netflix
Google
Uber
Yelp
Square

# What is a Neural Network?

- Mimic and reflect the behavior of the human brain.
- They are created and implemented to recognize patterns and solve common problems in the fields of AI, ML and DL
- Training a neural network revolves around the following items:
  - *Layers*
  - *Input Data* and their corresponding *targets*
  - *Loss function*
  - *Optimizer*

# My Model

- 4 layers
  - 3 with ReLU
  - 1 with Softmax
- Loss
  - categorical cross entropy
- Optimizer
  - RMSProp

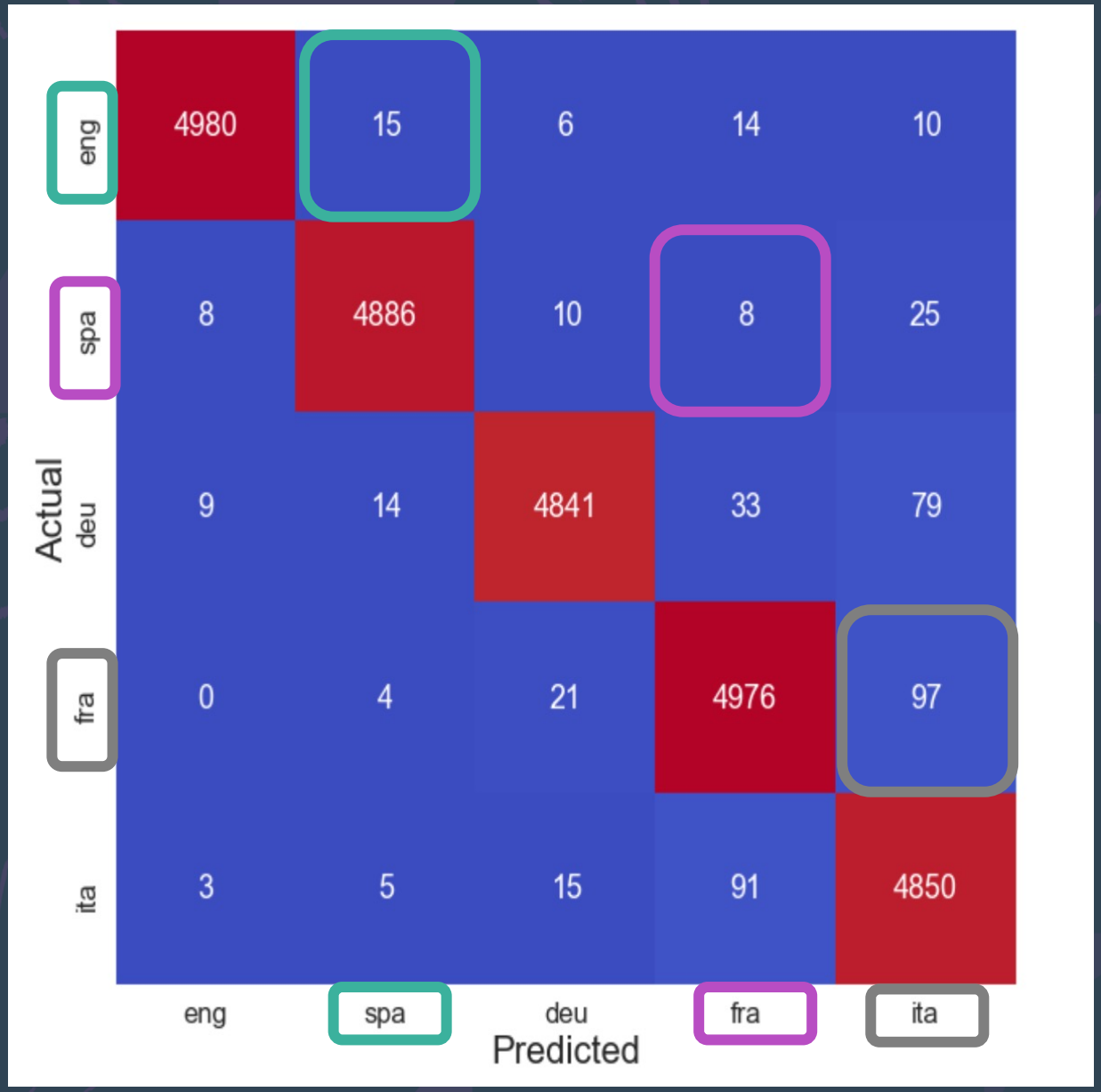
```
# Instantiate a model
# large datasets have lower epochs
# input dimension is number of unique trigrams +1
model = Sequential()
model.add(Dense(500, input_dim = 617, activation = 'relu'))
model.add(Dense(500, activation = 'relu'))
model.add(Dense(250, activation = 'relu'))
model.add(Dense(5, activation = 'softmax'))
model.compile(loss = 'categorical_crossentropy',
              optimizer = 'rmsprop',
              metrics = ['accuracy'])

# Train model
history = model.fit(x, y, epochs = 4, batch_size = 25, validation_data = (x1, y1))
```

# Results

- Training data accuracy: 98.36%
- Test set accuracy: 98.132%.
- Minor overfitting may have occurred with the training data but overall, this is a healthy accuracy number.

- Interpretation:
  - English was incorrectly predicted as Spanish in 15 cases
  - Spanish was incorrectly predicted as French in 8 cases
  - The worst scenario is that French was predicted as Italian 97 times.



# Conclusion

- The goal of this presentation was to create and train a neural network that detects the difference between the English, Spanish, German, French and Italian languages.
- Natural Language Processing projects need to identify which language their input text is in order to move forward and complete any remaining tasks.
- Examples:
  - Online Chatbots
  - Sentiment Analysis on reviews, comments, etc.

# Thank You!

# References


- About Keras. (n.d.). Retrieved July 15, 2021 from https://keras.io/about/
- Analytics Vidhya (October 26, 2017). The essential NLP guide for data scientists (with codes for top 10 common NLP tasks). Retrieved July 20, 2021 from https://www.analyticsvidhya.com/blog/2017/10/essential-nlp-guide-data-scientists-top-10-nlp-tasks/
- Bogdanov, V. (February 15, 2019). 8 thought-provoking cases of NLP and text mining use in business. Retrieved July 16, 2021 from https://becominghuman.ai/8-thought-provoking-cases-of-nlp-and-text-mining-use-in-business-60bd8031c5b5
- Chollet, F. (2018). *Deep Learning with Python*. Manning Publications Co.
- Edell, A. (April 25, 2018). *How I trained a language detection AI in 20 minutes with a 97% accuracy*. Retrieved July 15, 2021 from https://towardsdatascience.com/how-i-trained-a-language-detection-ai-in-20-minutes-with-a-97-accuracy-fdeca0fb7724
- Mujtaba, H. (August 29, 2020). What is rectified linear unit (ReLU)? Retrieved July 16, 2021 from https://www.mygreatlearning.com/blog/relu-activation-function/
- N-gram. (April 20, 2021). In Wikipedia. Retrieved July 15, 2021 from https://en.wikipedia.org/wiki/N-gram
- O'Sullivan, C. (August 25, 2020). *Deep neural network language identification*. Retrieved July 15, 2021 from https://towardsdatascience.com/deep-neural-network-language-identification-ae1c158f6a7d
- Tatoeba. (June 27, 2021). In Wikipedia. Retrieved July 15, 2021 from https://en.wikipedia.org/wiki/Tatoeba
- Simplilearn. (March 12, 2021). *What is Keras? The best introductory guide to keras.* Retrieved July 15, 2021 from https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras