Assignment 4 — Supervised Learning and Neural Networks for Text Analysis

1. In this question, you will develop models for predicting the party of speaker from their speech. The website https://data.stanford.edu/congress_text collected the Congressional Record and parsed the data by speaker. The data also contains the party of the speaker. You will use this to build a number of models for predicting party from speech. Use you use the following code to load the data for the 113th Congress in r:

```
data1a <- read.delim("https://www.michaelperess.com/classdata/
 113_SpeakerMap.txt",sep="|")
data1b <- read.delim("https://www.michaelperess.com/classdata/
 speeches_113.txt",sep="|",quote="")
text1 <- data1b$speech
text1 <- iconv(text,from="",to="UTF-8",sub="byte")
party1 <- data1a$party[fmatch(data1b$speech_id,data1b$speech_id)]
n1 <- length(party1)
```

You can use the following code to load the data in *python*:

```
data1a = pd.read_csv(r"https://www.michaelperess.com/classdata/
 113_SpeakerMap.txt" ,sep="|",encoding="latin1")
data1b = pd.read_csv(r"https://www.michaelperess.com/classdata/
 speeches_113.txt",sep="|",encoding="latin1",on_bad_lines="warn")
 # there is a single bad line due to bad OCR
res = pd.merge(data1a,data1b,on="speech_id") # merge by key
text1 = res["speech"] # get text
party1 = res["party"] # get party
n1 = len(party1)
```

Each data point is a speech by a member of congress. To prepare the data, eliminate all speeches that are less than 100 characters and all speeches by independent members of congress. Form the dependent variable as 1 if the speaker is Republican and 0 if the speaker is a Democrat.

(a) Apply Naive Bayes to the data and describe the performance of the model on a test set. Report which words predict a Democratic speaker and which words predict a Republican speaker. Comment on the results.

(b) Apply the Lasso to the data and describe the performance of the model on a test set. Report which words predict a Democratic speaker and which words predict a Republican speaker. Comment on the results.

(c) Apply a basic Neural Network to predict the party of the speaker and describe the performance of the model on a test set.

(d) Apply an LSTM or a Transformer Model to predict the party of the speaker and describe the performance of the model on a test set.

2. In this question, you will expand upon the previous analysis and predict the party of a twitter poster based on patterns of speech found in the Congressional Record. This question is designed to begin exploring the utility of "fake" supervised learning, where coded data from one context is applied to another context. It is still artifical since the party of twitter users is known, but a similar technique is used by articles such as Gentzkow and Shapiro (2010) and Martin and Yurukoglu (2017) to measure the ideology or newspapers and television news shows using supervised learning without having to explicitly code training data. The following data contains tweets of members of Congress and their party. Use you use the following code to load the data in r:

```
download.file("https://www.michaelperess.com/classdata/twitter.zip",
 dest="twitter.zip",mode="wb")
unzip("twitter.zip")
data2 <- read.delim("twitter.dat",stringsAsFactors=FALSE)
text2  <- data2$Text
party2 <- data2$Party
n2 <- length(party2)
```

You can then combine the Congressional Record and Twitter data using:

```
text <- c(text1,text2)
party <- c(party1,party2)
ntrain <- n1
ntest <- n2
```

You can use the following code to load the data in *python*:

```
data2 = LoadDelimFromWebZip("https://www.michaelperess.com/classdata
 /twitter.zip")
text2 = data2["Text"]
party2 = data2["Party"]
n2 = len(party2)
```

and the following code to combine the Congressional Record and Twitter data:

```
text = pd.Series(list(text1) + list(text2))
party = pd.Series(list(party1) + list(party2))
ntrain = n1
ntest = n2
```

As before, prepare the data by eliminating all speeches that are less than 100 characters and all speeches by independent members of Congress. Form the dependent variable as 1 if the tweeter is Republican and 0 if the tweeter is a Democrat.

- Use one of the models from question 1 to predict the party of twitter users from their tweets, and check the results against their actual party. Comment on the results.

References

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78:35–71.

Martin, Gergory J. and Ali Yurukoglu. 2017. "Bias in Cable News: Persuasion and Polarization." *American Economic Review* 107:2565–2599.