Twitter Language Models about Ukraine

Timon Tukei

Introduction

The largest armed conflict in Europe since WW2 was set off when Vladimir Putin launched an unjustified invasion into Ukraine on February 24, 2022.

This of course took me and many others by surprise and this is what had led me to choose this as my topic of research for data for my language models.

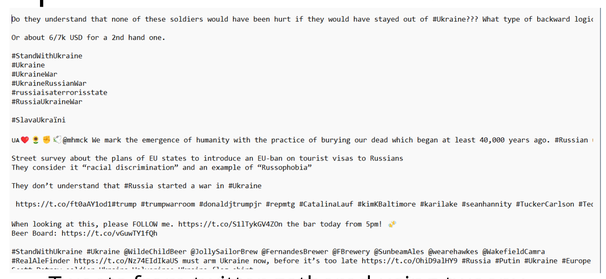
Ukraine Quick Facts

> Ukraine is the world's largest producer of sunflower seeds, and across every part of Ukraine, you'll see stunning fields of these yellow blooms – so many that they'd cover the entire of Slovenia if you uprooted them!

> Ukraine is the second largest country in Europe in terms of total area– spanning 603.55 square km (or 233 square miles – twice the size of Italy

Data

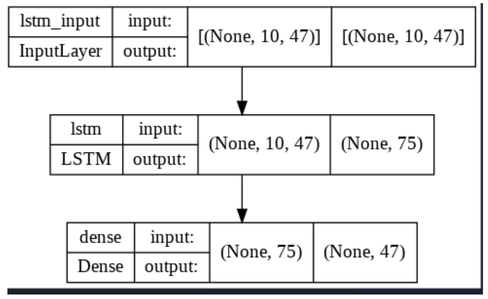
My data consisted of a variety of tweets collected from twitter via Tweepy. Tweepy is the twitter API for python and allows for querying tweets with specific search criteria.



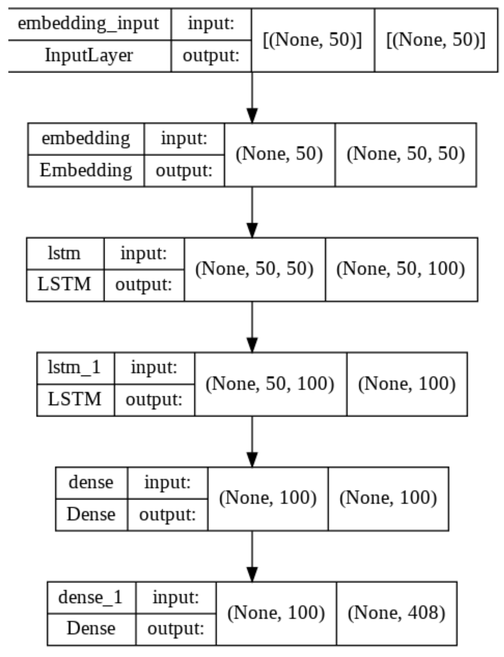
Tweets from twitter gathered using tweepy.

Implementation

I began with a character based neural language model to generate text using tweets as data for training the models. I then developed word-based models that used one-word, two-word, and line-based framings. Finally, I completed by developing a neural word-based language model. The following images display the different models that were in use:



Character base language model



Word base language model

Results

I had generated 23 tweets using each language model and the curated twitter dataset that I collected using Tweepy. Below is some of the tweets that were generated

temporarily people south country involvement another explosion Russian assets frozen used pay latest The music better see fight war going would given chance express sorrow think set flying experience aviation

Malva seen wheeled chassis Russia pope dialogue us humanly impossible hope said papal envoy blamed military warehouse northern forced evacuation people

Russian energy counter Russia influence help staff situation help support day every night see new territory temporarily people south country involvement another explosion Russian assets frozen used pay latest

The music better see fight war going would given chance express sorrow think set flying experience aviation Malva seen wheeled chassis Russia pope dialogue us humanly impossible

Character based language model generated tweets

Conclusion

Granted that the set of data used to train both the character-based model and word-based model is limited to 100 tweets there was some unexpected output although most was relatively well. The accuracy per epoch for the word base neural language model was the worst by far but it still manage to generate text!