Twitter Bot

Final Presentation

Sina Behdadkia
Tim Mennicken
Robert Rose

University of Applied Sciences Cologne

Tuesday the 4th of February, 2020

Table of Contents

Introduction

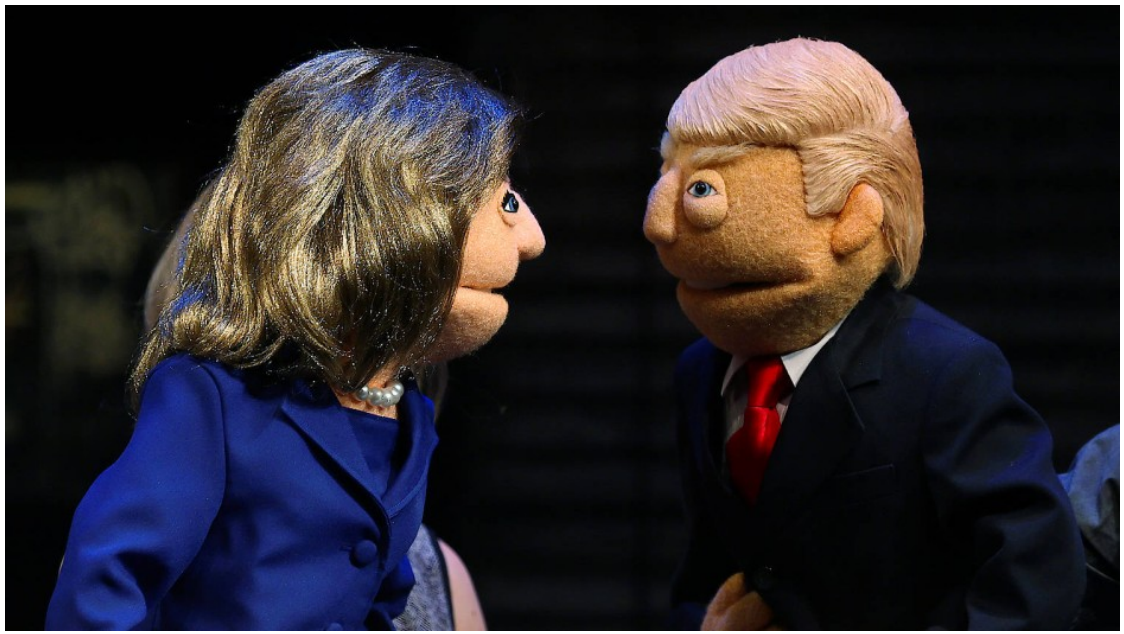
Goal

- Develop neural network, which is able to create Tweets
- Tweets should imitate a person as good as possible
- People to imitate:

Introduction

Goal

- Develop neural network, which is able to create Tweets
- Tweets should imitate a person as good as possible
- People to imitate:



Introduction

Goal

Wokflow divided into three parts:

- Create a dataset
  - Crawl Tweets
  - Preprocess gathered Tweets

Introduction

Goal

Wokflow divided into three parts:

- Create a dataset
  - Crawl Tweets
  - Preprocess gathered Tweets
- Design a network
  - Develop the architecture
  - Experiment with different setups

Introduction

Goal

Wokflow divided into three parts:

- Create a dataset
  - Crawl Tweets
  - Preprocess gathered Tweets
- Design a network
  - Develop the architecture
  - Experiment with different setups
- Generate Tweets
  - Post-process predictions
  - Present the results

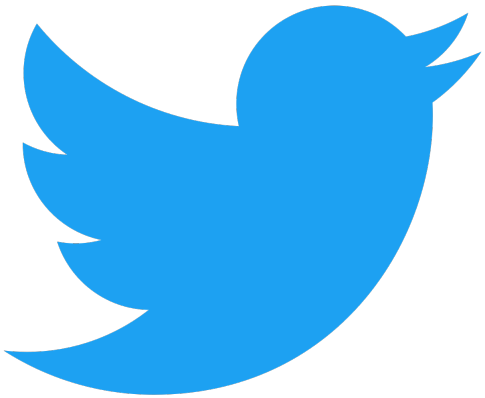
Introduction

Motivation

- Twitter is on of the big players in social media
- Microsoft tried to setup a chat bot to learn how young people communicate
  - Got token down after 24 hours
- 9-17 % of Twitter users are bots
- Risk of political manipulation is high

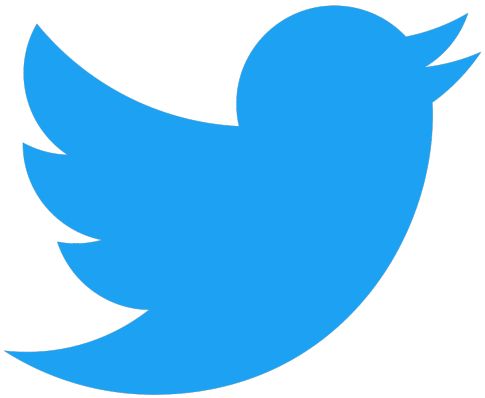
Data Acquisition

- Twitter developer platform
  - Twitter developer account needed
  - Tweepy: Python wrapper
  - Complete functionality

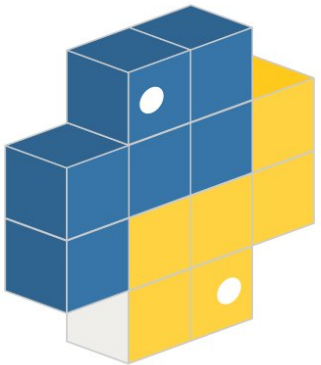


Data Acquisition

- Twitter developer platform
  - Twitter developer account needed
  - Tweepy: Python wrapper
  - Complete functionality



- GetOldTweets
  - Fetches Tweets from the website
  - Python package
  - Read only access



Data Acquisition

Twitter API

- Amount of Tweets limited
  - Request windows are separated in 15 minutes chunks
  - Specific amount of requests per window
  - Can be bypassed by cyclic requests paired with pauses

Data Acquisition

Twitter API

- Amount of Tweets limited
  - Request windows are separated in 15 minutes chunks
  - Specific amount of requests per window
  - Can be bypassed by cyclic requests paired with pauses
- Cannot go arbitrarily far to the past
  - Returns no Tweets older than roundabout a month
  - Cannot be bypassed

Data Acquisition

Twitter API

- Amount of Tweets limited
  - Request windows are separated in 15 minutes chunks
  - Specific amount of requests per window
  - Can be bypassed by cyclic requests paired with pauses
- Cannot go arbitrarily far to the past
  - Returns no Tweets older than roundabout a month
  - Cannot be bypassed
- Account got blacklisted
  - No further access to the Twitter API
  - Got unblocked on request

Data Acquisition

GetOldTweets

- No limitation
  - Arbitrary amount of Tweets
  - No restrictions with respect to publication date

Data Acquisition

GetOldTweets

- No limitation
  - Arbitrary amount of Tweets
  - No restrictions with respect to publication date
- Shortened functionality
  - Limited meta data of Tweets
  - No functionality for publishing Tweets

Data Acquisition

GetOldTweets

- No limitation
  - Arbitrary amount of Tweets
  - No restrictions with respect to publication date
- Shortened functionality
  - Limited meta data of Tweets
  - No functionality for publishing Tweets
- Better suited for getting large data sets

Pre-Processing

Particular Content

- Weblinks
- Picture & video links
- Punctuation symbols
- Special characters
- Retweets
- Hashtags
- Username references

Pre-Processing

Particular Content

Content we keep:

1. Selected punctuation symbols

   point, comma, exclamation mark, interrogation mark, colon and hash
2. Hashtags
3. References to usernames

Pre-Processing

Space characters

- Union of several space characters
- The Tokenizer splits the input text at space characters

Pre-Processing

Space characters

- Union of several space characters
- The Tokenizer splits the input text at space characters

Why we need to add spaces

house. ⇒ ["house."]
house . ⇒ ["house", "."]

Pre-Processing

Termination Symbol

- Model does not know when to stop
- Stopping generation at an arbitrary number looks choppy
- Termination symbol is introduced

Post-Processing

- Remove unfitting space characters
  - E.g. before punctuation symbols

Post-Processing

- Remove unfitting space characters
  - E.g. before punctuation symbols
- Make words uppercase
  - E.g. at the beginning of a sentence

Post-Processing

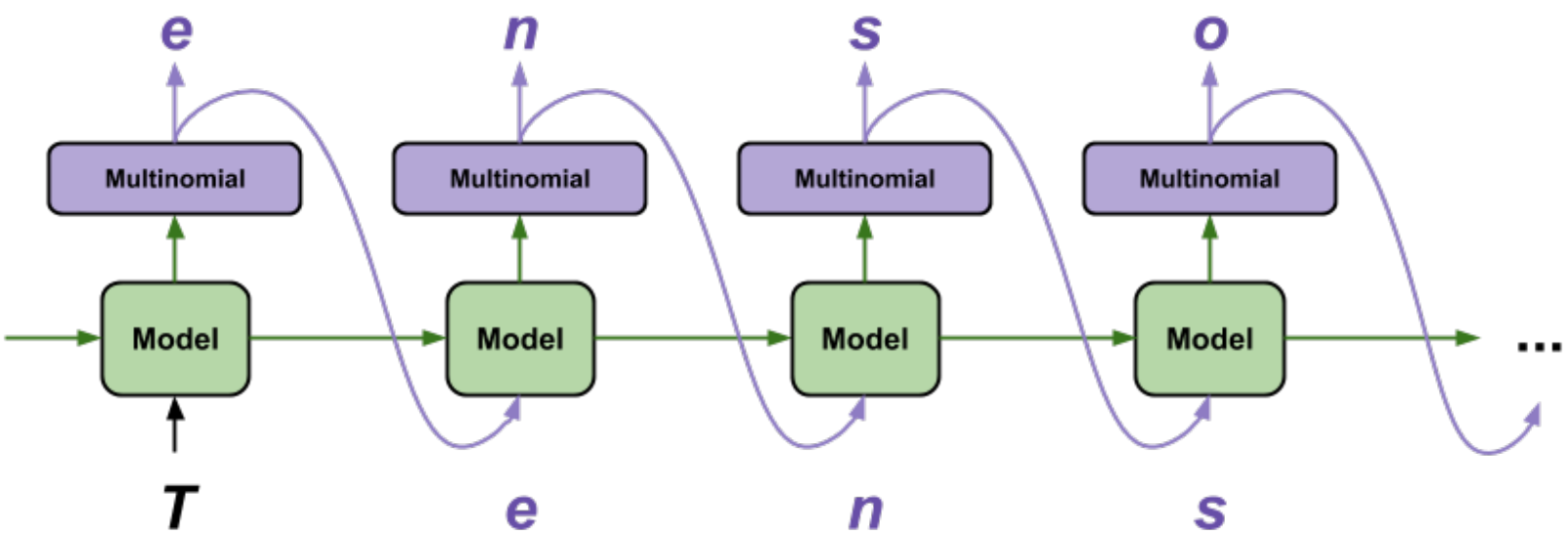
- Remove unfitting space characters
  - E.g. before punctuation symbols
- Make words uppercase
  - E.g. at the beginning of a sentence
- Restrict generation
  - Lower limit
  - Upper limit

Experiments

- Character level model
- Word level model

Experiments

Character level model



Experiments

Character level model

- Seed:

  “i just realized that if you listen to ca”

- Generated text:

  “nt beloels like scobous dveb ! vote selffiending up. #fitn graham of his tonight dominater wsa an ands. comfuntstaheos,”

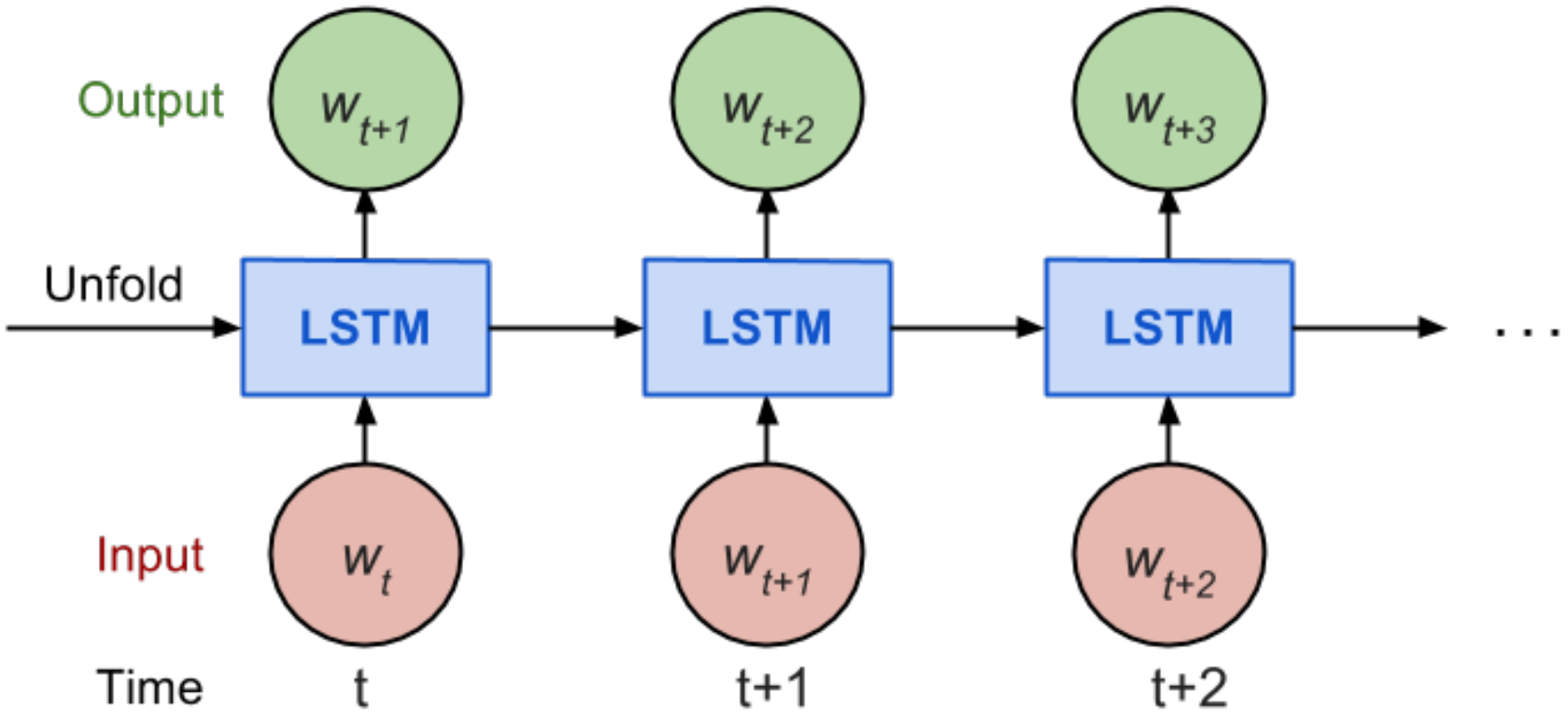
Experiments

Word level model

- Multiple undirectional LSTM
- Single bidirectional LSTM
- Multiple bidirectional LSTM

Experiments

Word level model – Multiple undirectional LSTM



Experiments

Word level model – Multiple undirectional LSTM

- Seed:

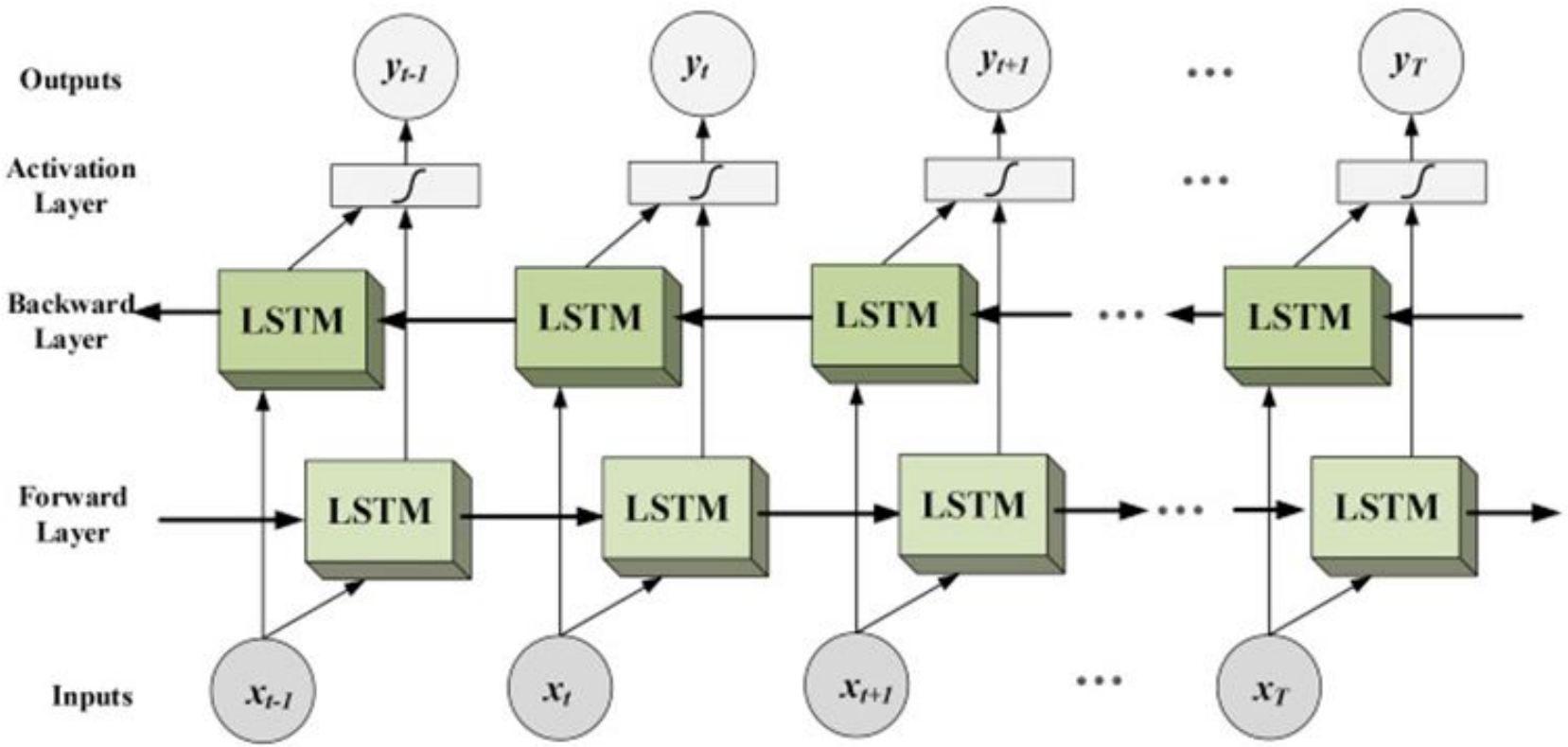
  “and how innocent she is , ask her to read peters insurance policy text , to her , just in case hillary loses . also , why were the lovers text messages scrubbed after he left mueller . where are they lisa ? ; the republican party has never been so”

- Generated text:

  “easy to gain. They are entrapping people with china. They have gone bonkers, and I havent seen millions of americans!”

Experiments

Word level model – Single bidirectional LSTM



Experiments

Word level model – Single bidirectional LSTM

- Seed:

  “co-founder of greenpeace: the whole climate crisis is not only fake news , its fake science . there is no climate crisis , theres weather and climate all around the world , and in fact carbon dioxide is the main building block of all life . @foxandfriends wow ! ; jewish”

- Generated text:

  “people are leaving the democratic party. The fires lost in favor of our seniors. Presidential world”

Experiments

Word level model – Multiple bidirectional LSTM

- Seed:

  “fraudulent speech knowingly delivered as a ruthless con , and the illegal meetings with a highly partisan whistleblower & lawyer . @60minutes forgot to report that we are helping the great farmers of the usa to the tune of 28 billion dollars , for the last two years , paid for”

- Generated text:

  “the democrats . the democrats are a very good of the democrats . the democrats are a very good of the democrats . the democrats are a very good of”

Experiments

Suggested Model

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 50, 50)	503500
lstm_1 (LSTM)	(None, 50, 128)	91648
dropout_1 (Dropout)	(None, 50, 128)	0
bidirectional_1 (Bidirection	(None, 256)	263168
dense_1 (Dense)	(None, 100)	25700
dense_2 (Dense)	(None, 10070)	1017070

Total params: 1,901,086
Trainable params: 1,901,086
Non-trainable params: 0

Conclusion

Results

- ✓ Char-level tweet creation
- ✓ Word-level creation
- ✓ Readable results

- – Still easy to distinguish
- – Clinton model sometimes won't predict readable results

Conclusion

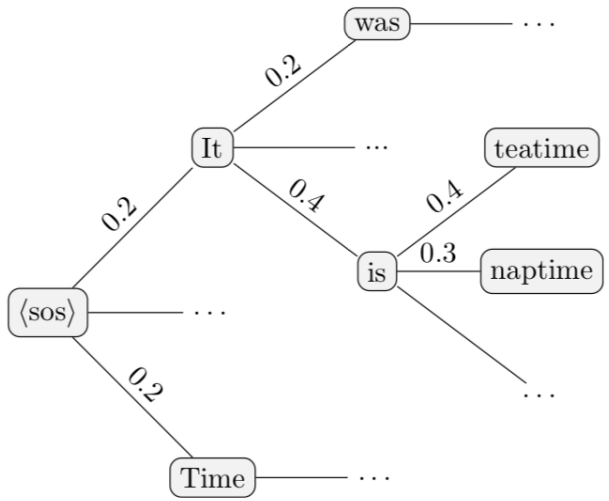
Possible Improvements

- Gather more data
- Deeper network
- Use optimization tools to find good hyperparameter combinations
- Implement beam search
- Use Attention techniques

Conclusion

Possible Improvements

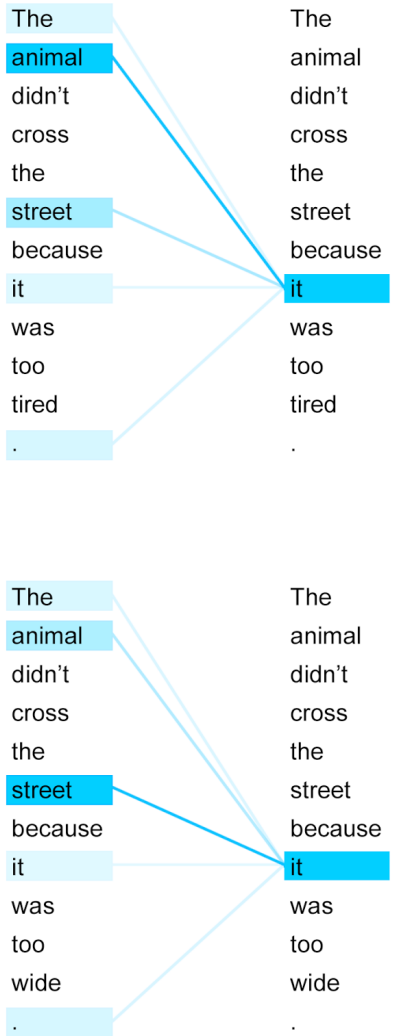
- Gather more data
- Deeper network
- Use optimization tools to find good hyperparameter combinations
- Implement beam search
- Use Attention techniques



Conclusion

Possible Improvements

- Gather more data
- Deeper network
- Use optimization tools to find good hyperparameter combinations
- Implement beam search
- Use Attention techniques



References

- Twitter Developers – https://developer.twitter.com/en.html
- GetOldTweets – https://github.com/Jefferson-Henrique/GetOldTweets-python
- Transformer – https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html
- Beam Search – https://towardsdatascience.com/the-arti-canon-neural-text-generation-2wwa8f032c2a68
- Tay (German source) – https://www.faz.net/aktuell/wirtschaft/netzwirtschaft/microsofts-bot-tay-wird-durch-nutzer-zum-nazi-und-sexist-14 html