CSS Bootcamp
Neural Analysis of Text Data
Day 1: Transformer Models

Xingyuan Zhao

Sept. 12 2022

1 Sentiment Analysis

Find the most positive and negative sentence in the paragraph below.

"Eighteen years have gone by, and still I can bring back every detail of that day in the meadow. Washed clean of summer's dust by days of gentle rain, the mountains wore a deep, brilliant green. The October breeze set white fronds of head-tall grasses swaying. One long streak of cloud hung pasted across a dome of frozen blue. It almost hurt to look at that faroff sky. A puff of wind swept across the meadow and through her hair before it slipped into the woods to rustle branches and send back snatches of distant barking-a hazy sound that seemed to reach us from the doorway to another world. We heard no other sounds. We met no other people. We saw only two bright, red birds leap startled from the center of the meadow and dart into the woods. As we ambled along, Naoko spoke to me of wells. Memory is a funny thing. When I was in the scene, I hardly paid it any mind. I never stopped to think of it as something that would make a lasting impression, certainly never imagined that eighteen years later I would recall it in such detail. I didn't give a damn about the scenery that day. I was thinking about myself. I was thinking about the beautiful girl walking next to me. I was thinking about the two of us together, and then about myself again. It was the age, that time of life when every sight, every feeling, every thought came back, like a boomerang, to me. And worse, I was in love. Love with complications. Scenery was the last thing on my mind. Now, though, that meadow scene is the first thing that comes back to me. The smell of the grass, the faint chill of the wind, the line of the hills, the barking of a dog: these are the first things, and they come with absolute clarity. I feel as if I can reach out and trace them with a fingertip. And yet, as clear as the scene may be, no one is in it. No one. Naoko is not there, and neither am I. Where could we have disappeared to? How could such a thing have happened? Everything that seemed so important back then-Naoko, and the self I was then, and the world I had then: where could they have all gone? It's true, I can't even bring back Naoko's face-not right away, at least. All I'm left holding is a background, sheer scenery, with no people up front."

2 Toxic Speech Detection

You've already seen how the model can classify a sentence as positive or negative using those two labels — but it can also classify the text using any other set of labels you like. In this exercise, you will try to find whether a tweet is toxic or not. *Please note that the*

data contains offensive or sensitive content, including profanity and racial slurs.

Abusive language on online platforms has become a major concern in the past few years. One of the most effective strategies for tackling this problem is to use computational methods to identify offense, aggression, and hate speech in user-generated content (e.g. posts, comments, microblogs, etc.).

2.1 Find whether the tweets are offensive or not(hint: using zero-shot-classification pipeline) on testing examples

Today, let's first try to use the pre-trained transformer model to find out offensive tweets without any fine-tuning on this specific task.

Let's start from these testing examples[1].

The labels of offensiveness are shown in brackets.

1. @USER Liberals are all Kookoo !!! (Offensive)
2. @USER @USER Go home you're drunk!!! @USER #MAGA #Trump2020 (Offensive)
3. Amazon is investigating Chinese employees who are selling internal data to third-party sellers looking for an edge in the competitive marketplace. URL #Amazon #MAGA #KAG #CHINA #TCOT (Not-Offensive)

2.2 Find whether the tweets are offensive or not and evaluate the model

Now let's move to larger-scale dataset. We provide data drawn from SemEval 2019 task on offensive language detection. The files consists of tweets annotated for offensiveness.

The first column (text) contains the text of a tweet, the second column (label) contains an offensiveness label[2]:

1. (NOT) Not Offensive - This post does not contain offense or profanity.
2. (OFF) Offensive - This post contains offensive language or a targeted (veiled or direct) offense.

The current task is to first using pre-trained transformer to classify each sentence as offensive or not-offensive, then evaluate the model using two criteria: (1) F1 Score over hate speech detection("NOT" is considered the positive label) and (2) False Positive Rate (FPR), how often the model mis-classified non-toxic speech as toxic.

[1] From SemEval 2019

[2] The file "offenseval-annotation.txt" provides additional details on the annotation scheme FYI