# Project 3 - NLP

Student Name: Shai Cahana

Submission Date: 10/11/2024

**Part 1 - Text Classification**

**1.1**

|  |
| --- |
|  |

**1.2** **Explain why precision and recall metrics are better‬**

This dataset is imbalance in the favor of neutral emotions. In this case, using can be misleading, because a model could achieve accuracy by always predicting “neutral” .

Instead, measures how accurate our positive predictions are, while shows the coverage of actual positive cases from a category.

Together, these metrics provide a more meaningful evaluation of the model's performance across all emotions, not just the dominant neutral class.

**1.3**

For the unbalanced training set:

A screenshot of a computer

Description automatically generated

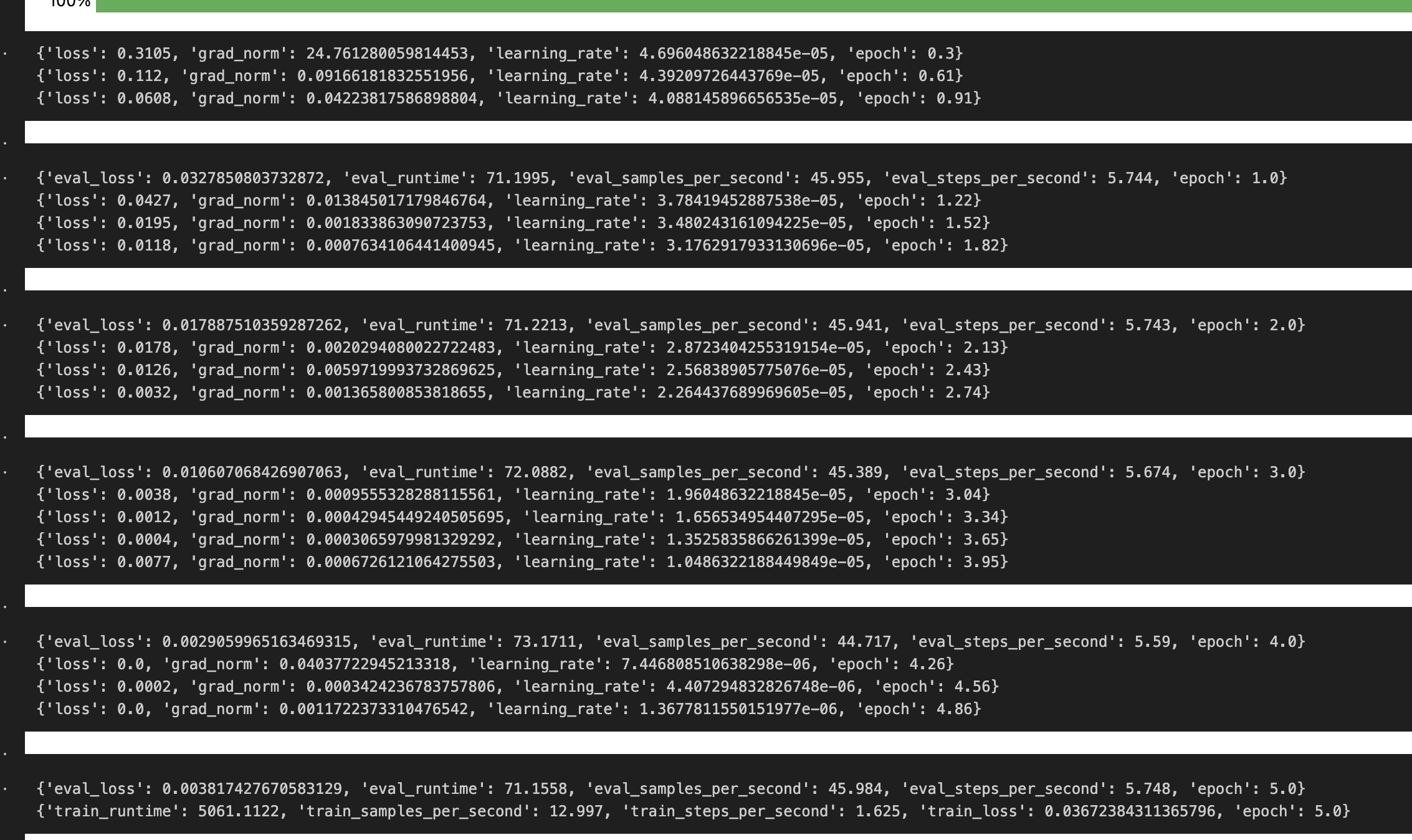
A computer screen shot of a program

Description automatically generated

Extra - After balancing the dataset:

A computer screen shot of a program

Description automatically generated



Balancing the dataset slightly improved the precision and recall, this could be due to the fact that I oversampled the data for the non-neutral emotions.

**1.4**

**1.5**

**1.6**

**Part 2 - Text summarization**

**2.2**

**A graph of a graph

Description automatically generated**

**2.3.**

**A graph with blue squares

Description automatically generated**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Article** | **Summary** |
| **Highest** | **Index** 621 | (CNN) -- ZipRealty is a full-service residential real estate brokerage firm which uses the Internet, proprietary technology and employee real estate agents to provide home buyers and sellers with value-added online service. ZipRealty's Web site provides users with access to comprehensive home listings data. ZipRealty's Web site provides users with access to comprehensive local Multiple Listing Services home listings data, as well as other relevant market and neighborhood information. Its proprietary business management system and technology platform help to reduce costs, allowing the company to pass on savings to consumers. ZipRealty was launched on August 29, 1999. Three weeks later, on September 21, the company celebrated the acceptance of its first real estate offer. ZipRealty completed its initial public offering on November 10, 2004. It currently operates in 34 major metropolitan areas in 19 states and the District of Columbia with over 2,000 sales agents. | ZipRealty uses Internet, proprietary technology, employee real estate agents .  Web site provides users with home listings, market and neighborhood info .  ZipRealty completed its initial public offering on November 10, 2004 . |
| score: 0.320754 |
| **Lowest** | **Index** 70 | The company was founded in 1985 by seven communications industry veterans -- Franklin Antonio, Adelia Coffman, Andrew Cohen, Klein Gilhousen, Irwin Jacobs, Andrew Viterbi and Harvey White. One of Qualcomm's first products was OmniTRACS, introduced in 1988, which is currently the largest satellite-based commercial mobile system for the transportation industry. Today, Qualcomm's patent portfolio includes approximately 6,100 United States patents and patent applications for CDMA and related technologies. More than 130 telecommunications equipment manufacturers worldwide have licensed QUALCOMM's essential CDMA patents. Qualcomm is among the members of the S&P 500 Index, Fortune 500, and a winner of the U.S. Department of Labor's" Secretary of Labor's Opportunity Award." The company has been listed among Fortune's "100 Best Companies to Work For in America" for nine years in a row and the magazine's list of" Most Admired Companies." Qualcomm's Annual revenue for 2006 was $7.53 billion, with a net income of $2.47 billion. E-mail to a friend . | The company has become a huge name in communications in just 20 years .  Qualcomm has a portfolio of approximately 6,100 U.S. patents .  Fortune lists the company as one of the 100 best places to work in the U.S. |
| score: 0.0 |

**‬**

The T5 has token limit of 512 but some of our texts are longer, which might affect the result as well.

**2.4**

**The rouge 2 score of the first 10 entries are in the bolded paragraph**

|  | index | **generated\_rouge2** | ground\_truth\_rouge2 | is\_lower |
| --- | --- | --- | --- | --- |
| 0 | 0 | **0.050209** | 0.155642 | True |
| 1 | 1 | **0.035503** | 0.050420 | True |
| 2 | 2 | **0.012195** | 0.040230 | True |
| 3 | 3 | **0.041841** | 0.039683 | False |
| 4 | 4 | **0.019960** | 0.026565 | True |
| 5 | 5 | **0.011080** | 0.026525 | True |
| 6 | 6 | **0.015190** | 0.028571 | True |
| 7 | 7 | **0.061538** | 0.132075 | True |
| 8 | 8 | **0.038023** | 0.057971 | True |
| 9 | 9 | **0.023256** | 0.043011 | True |

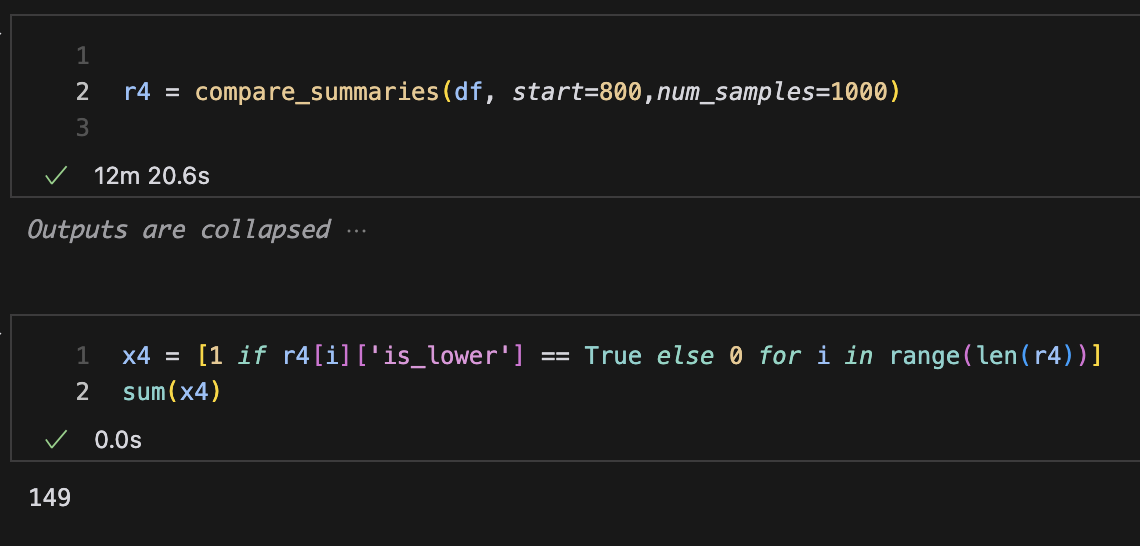
There are instances where the score is lower than ground truth. These results are marked as True.

**A screen shot of a computer program

Description automatically generated** **A screenshot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated**

****

In most cases, the score is lower than the ground truth! Meaning we get a better summry by T5 than the ground truth, according to the Rouge-N metric.

**A blue pie chart with a triangle

Description automatically generated**

**2.5**

**Part 3 - Information Retrieval**

**3.1**

**3.2**

**3.3**