COSC 426 Final Project Proposal

Kayla Mistica, James Njoroge, Emmett Hintz

Introduction/ Motivation

- What is the big picture question you are trying to answer/ problem you are trying to solve? Why is this an interesting and worthwhile question to answer/ problem to work on?

If multiple football matches are happening at the same time, which match is the most interesting to watch? Solving this question is worthwhile because it has real-world applications and would be a valuable resource for football match viewers.

- What is the specific question you will pursue? Why? (Note: you will need to pick something that is feasible to answer in 3-4 weeks)

Is it possible for a language model to receive commentary on a football match and measure its interest? This information would then weigh into a viewer's decision when choosing between matches to watch.

- What are all the possible results/outcomes of your project? Do you think one (or few) of these results/outcomes are more likely than others? Why?

- The accuracy of the language model will not be very high due to the limited amount of data in the dataset. (most likely because having a lot of training data allows a model to be as accurate as possible, and we are limited in the amount of data available)
- The language model can accurate in predicting the true label of the match (likely)
- The model fails to accurately predict the sentiment of a match (likely as the language model may need more data to work with or might not be able to work with such a large context at all)

- How do your possible outcomes relate to your original question you are trying to answer/ problem you are trying to solve?

Each outcome will gauge the feasibility of using match commentary to classify how interesting a match is. This would thus provide a foundation for future work to utilize these techniques and apply it to broadcasting software in order to benefit football viewers when they are deciding which match to watch.

Background/ Literature Review

Find at least **three papers** related to your project. For each project write a paragraph or two summarizing:

Paper 1:

- The paper "Semantic Understanding of Professional Soccer Commentaries" aimed to develop a novel method for semantic parsing that learns correspondences between complex sentences in soccer commentaries and the rich sets of events occurring in the game. By utilizing weak supervision through rough temporal alignments between sentences and events, the authors employed a discriminative approach to model these correspondence patterns. Their method could form "macro-events" that align with single sentences, significantly outperforming previous state-of-the-art approaches. The ultimate goal was to advance towards automatic generation of soccer game commentaries by accurately aligning textual descriptions with in-game events.
- This work relates to our project in that it also leverages natural language processing techniques on soccer commentary data. However, while their focus was on semantic parsing to match commentary sentences with specific game events, our proposed research aims to predict the level of interest or excitement in a match based on the commentary. We plan to analyze linguistic indicators of excitement—including scorelines, significant events, and sentiment analysis—to develop a model that captures the subjective notion of an "interesting" match. In contrast to the paper, which seeks to understand and generate commentary, our project aims to assist fans in selecting the most engaging game to watch when multiple matches are played simultaneously, thus enhancing the football viewing experience from a different angle.

Paper 2:
https://ieeexplore.ieee.org/document/8561283/footnotes#footnotes-id-fn2
This paper aimed to analyze the sentiment of Twitter posts during a live football match.
This paper is similar to our project as they both involve the input of some text to assign it a label. In the case of this paper, the language model receives a tweet about the football match and outputs the sentiment of that tweet (i.e., positive, negative, or neutral). In the context of our paper, the language model will receive input commentary on a football match and label it as "High Stakes," "Exciting," "High Stakes & Exciting," or "Normal." This paper creates a football-specific sentiment dataset to evaluate the performance of different machine-learning algorithms and features in classifying football-related tweets. They conclude that performance was relatively consistent across the different ML algorithms and that performance could be improved if there was a balanced amount of data across each class (i.e., the number of negative tweets is relatively close to the number of positive tweets). What makes our project different is

that our input is the entire commentary of a football match. This means our context window is much larger than the average input in sentiment analysis experiments.

Paper 3: SOCCER: An Information-Sparse Discourse State Tracking Collection in the Sports Commentary Domain
https://aclanthology.org/2021.naacl-main.342/

This paper introduces a dataset designed to map in-game events to natural language soccer commentary. This dataset, known as SOCCER, includes over 2,200 matches and 135,000 commentary segments from top soccer leagues. The primary focus of this dataset and research is recognizing occurrences of game events within sparse commentary data, aiming to create a system capable of tracking “event-driven states” from the narrative provided by the commentators.

This SOCCER dataset is particularly difficult due to the low information density stemming from the nature of football commentary. Baseline experiments were done using GRU and GPT-2 models, highlighting the difficulty of this task by showing that while accuracy is high because of non-match event commentary, the model struggles with recall for actual game time events. This paper is an interesting study that is attempting something similar. However, unlike this paper, which focuses on aligning text with discrete events, our project will further leverage sentiment analysis and language indicators to classify games as “High Stakes,” “Exciting,” or “Normal” to guide viewer interest.

Planned methods

Overall approach

What **steps** will you need to take to go from a dataset to a results table/ figure? How do these steps map onto the different modes in NLPScholar?

1. Data Loading

We begin by importing the necessary libraries for data manipulation and analysis. We then load the match details dataset from a CSV file into a DataFrame for processing. Then we inspect and verify the structure of the dataset to ensure all required information is present and correctly loaded.

2. Data Preparation

We define comprehensive lists of team name variations to account for different representations in the data (e.g., "Manchester United", "Man United"). Then we create a mapping dictionary that associates each variation with a standardized team label (e.g., "Team_1", "Team_2").

3. Data Masking

- We apply masking to replace team names in the “Home”, “Away”, “events”, and “summary” columns with the standardized team labels. This standardization ensures consistency and aids in anonymizing the data. We reckon this will help us be more objective and reduce bias towards certain teams.

- Then we check and verify the masked text for any unmasked team names to ensure that all instances have been successfully replaced. We display and review any rows where masking failed to correct issues in the mapping or masking function manually.

4. Feature Engineering

We define keyword lists that indicate high-stakes matches (e.g., "title race", "relegation battle") and exciting matches (e.g., "red card", "penalty"). We create a labeling function that scans the masked text for these keywords and assigns appropriate labels based on their presence. This is so that we can introduce the concept of “classification” and help us differentiate. Using online sources and our own football knowledge, we label games that are interesting.

5. Data Labeling

We apply the labeling function to categorize each match into labels such as "High Stakes", "Exciting", "High Stakes & Exciting", or "Normal". We add these labels as a new column in the DataFrame.

6. Data Visualization

We generate visualizations to understand the distribution of match labels. This is explorational and to help us understand/balance the dataset better.

7. Data Exporting

We prepare the data for export by selecting relevant columns, including masked team names, match details, masked text, and assigned labels. We then save the processed DataFrame to a TSV file.

Dataset

- What datasets will you use for each of the steps?

We will be using this dataset that we found on Kaggle.

- What are the preprocessing steps required? Will you use the entire dataset?

As stated in the overall approach, preprocessing the data will involve masking certain aspects of the dataset (home and away team) to eliminate any bias that might emerge when fine-tuning the language model. Each line of data will be assigned one of four different labels to categorize whether the game was "High Stakes," "Exciting," "High Stakes & Exciting," or "Normal." We will apply this to the entire dataset.

- How large is the final dataset you want to use?

The final dataset contains information from 300+ Premier League matches from the 23/24 season. The CSV file is about 6 MB in size.

Model

- What model will you use?

We will be using distilgpt2.

- What type of model is it on NLPScholar?

Text classification

- Will you have to finetune a model? If so, how long will finetuning it take?

We will finetine the model on 80% of the data and leave the rest for validation and testing. The finetuning will not take too long as our dataset is not that large. However, the size of the input text is larger than average.

Evaluation metrics

To evaluate our model's performance in distinguishing "interesting" vs. "not interesting" matches, we will leverage NLPScholar's automated output capabilities to generate our metrics, including accuracy, precision, recall, and F1 score. NLPScholar provides outputs in TSV format that capture detailed prediction data, such as model predictions, true labels, and specific conditions or contexts. We can leverage and directly calculate these metrics, offering a quantitative assessment of the model's effectiveness in capturing match interest.

The evaluation process will involve loading NLPScholar's prediction files and using `analyze` mode compute metrics, or even using visualizations if we can. Using the true and predicted labels within these TSV files, we will calculate accuracy, which shows the proportion of correct predictions, as well as precision, recall, and F1 score to assess balanced performance. Confusion matrices will also be generated to provide a breakdown of true positives, true negatives, false positives, and false negatives, allowing us to pinpoint areas of strength and improvement. These metrics and the confusion matrix can be easily summarized in tables or visualized as figures, providing a holistic view of the model's predictive ability.

In addition to quantitative metrics, we will analyze specific cases where the model misclassified matches, using these insights to guide future adjustments. For instance, we'll examine whether certain misclassifications stem from overfitting to technical terms or misinterpretation of sentiment within the commentary. This process ensures that we gain both quantitative and qualitative feedback from NLPScholar's outputs, enabling us to gain insights about our work and next steps.

Minimal working example

Github link