COSC 426 Final Project Proposal - James Njoroge

[FROM HERE ON OUT SOCCER = FOOTBALL]

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?
    - Football is a subjective sport, you're more attracted to watching your own team. But for true football fans, we want to watch whatever good/interesting game is out there. Sometimes when there are multiple games being played at the same time, one has to watch multiple minutes of every game to get an idea of what they want to watch. Statistics of the game are an indicator, but one-sided games are boring and we can use language to get sentiment of the game.
    - One constant of every game is commentary, describing in detail what is going on in the game. So can we use the commentary data from a football match to judge if a game is interesting or not to some degree of subjective accuracy?
    - If we can do this for past matches, then it can be implemented in real time matches, where you can take a subset of commentary, and give real time analysis on how interesting a match is for people to have an easy pick.
- 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)
    - Can we use the commentary data from a football match to judge if a game is interesting or not? [Draft question. Final one would be confirmed with peers]
        - This question is significant because it explores the intersection of natural language processing and sports analytics. By analyzing commentary data, we can tap into the nuanced descriptions and emotions conveyed during a match, potentially providing a more accurate measure of excitement than traditional statistics alone.
- 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 possible outcomes for this project are:

        - Successful Prediction Model: we train a model that accurately predicts the interest level of football matches based on commentary data. The model effectively captures the excitement conveyed in the commentary, aligning closely with human assessments and objective measures like significant events and scorelines.

        - Partial Success with Limitations: we train a model that predicts interest levels with moderate accuracy. The model performs well for certain types of matches—such as high-scoring games or matches with numerous significant events—but struggles with others, like defensive games with fewer highlights.

        - Inconclusive Results: the model fails to predict interest levels accurately. The commentary data may not provide sufficient indicators of excitement, or the subjective nature of "interest" proves too complex to model effectively within the project's scope.

    - I believe that achieving partial success with limitations is the most likely. Given the subjective nature of what makes a match interesting and the complexity of natural language, the model may capture general trends but might not account for all nuances. The time constraints and the project's scope also make a fully accurate model less probable.

- How do your possible outcomes relate to your original question you are trying to answer/problem you are trying to solve?
    - Outcome 1 would confirm that commentary data is a reliable source for assessing match excitement, validating the project's premise and offering a tool that could enhance the viewing experience for fans.
    - Outcome 2 suggests that while commentary data has predictive value, additional factors or more sophisticated models might be necessary for higher accuracy. This outcome still contributes valuable insights and could guide future research.
    - Outcome 3 indicates that commentary alone may not suffice to gauge interest accurately, highlighting the need to incorporate other data sources like live fan reactions, social media sentiment, or more detailed statistical analyses.

Background/ Literature review

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

- Paper 1:

  - The goal of the paper "Innovative Approaches in Sports Science—Lexicon-Based Sentiment Analysis as a Tool to Analyze Sports-Related Twitter Communication" was to evaluate the feasibility of using lexicon-based sentiment analysis tools to assess football-related Twitter communication. By manually annotating 10,000 tweets about 10 football matches and applying several sentiment analysis tools, the study aimed to validate the accuracy of these tools compared to human judgments. The conclusions indicated that while individual tweets or small sets were not classified with high accuracy, lexicon-based tools could correctly determine the general sentiment of large tweet sets—with a predominant sentiment polarity—with over 95% accuracy.
  - This paper relates to our project as both utilize sentiment analysis/classification within the context of football to extract subjective perceptions of excitement. However, our proposed research differs by focusing on match commentary data instead of Twitter communication to predict the level of interest in games. We aim to identify linguistic indicators of excitement within the commentary, incorporating scorelines, significant events, and expert evaluations to develop a model that defines an "interesting" match. Unlike the previous work, which assesses public sentiment on social media, our study seeks to assist fans in real-time decision-making when multiple matches are occurring simultaneously, enhancing their viewing experience through NLP applied directly to match commentary.
- Paper 2:
  - The paper titled "A big data analysis of Twitter data during Premier League matches: Do tweets contain information valuable for in-play forecasting of goals in football?" aimed to determine whether Twitter data could enhance the accuracy of in-play goal forecasting in football matches. By analyzing nearly two million tweets from over 400 Premier League games, the study employed sentiment analysis and random forest models to assess if real-time Twitter sentiment could improve predictions compared to pre-game betting odds. The conclusions indicated that in-play information, including Twitter data, did not significantly enhance forecasting accuracy. This suggests that extracting predictive value from unstructured textual data like tweets is highly challenging, and that pre-game information remains more reliable for forecasting purposes.
  - While this paper and our proposed research both explore the use of textual data and sentiment analysis in football, they differ in focus and application. The previous work concentrated on forecasting the occurrence of goals during matches using Twitter data, primarily for the benefit of coaches, analysts, and broadcasters. In contrast, our project aims to predict the level of interest or excitement in a game by analyzing match commentary. By identifying linguistic indicators of excitement and incorporating elements like score lines and

significant events, we seek to develop a model that helps fans choose the most engaging match to watch when multiple games are on simultaneously. Therefore, while both studies delve into the intersection of sports analytics and computational linguistics, our research shifts the focus from in-play goal forecasting to enhancing the football viewing experience for fans through real-time analysis of commentary data.

- Paper 3:
  - 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.

Planned methods

Overall approach

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

Dataset

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

    - We will be using this dataset and splitting it into train, validation, and test sets appropriately (the split has not been decided yet): https://www.kaggle.com/datasets/pranavkarnani/english-premier-league-match-commentary

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

    - Yes. We will use the entire dataset. Find the detailed pre-processing steps here (sorry for the inconvenience, they're just a bit verbose for here): https://github.com/James-Njoroge/nlp-final/blob/main/pre-processing.md

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

    - 303 rows

Model

- What model will you use?

    - distilgpt2 (For now. Can pivot depending on advice.)

- What type of model is it on NLPScholar?

    - text_classification_model

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

    - Yes we will. We don't expect it to take too long since the dataset is rather small as compared to others. On the other hand, it is a dataset that does contain verbose sentences.

Evaluation metrics

- How will you evaluate your model's performance (give an example that references model outputs)?

    - To evaluate our model's performance, we will use standard classification metrics: **accuracy, precision, recall (sensitivity), and F1 score.** We will mainly focus on accuracy which measures the proportion of all correct predictions. Also, the F1 score will be important for us as it combines precision and recall into a single metric for balanced performance evaluation. In our context, these metrics will clarify the model's ability to classify matches as "interesting" or "not interesting," providing insight into how well it distinguishes between these classes.

    - For instance, imagine we test the model on 60 matches with the following results: 25 true positives (correctly predicted "interesting" matches), 20 true negatives (correctly predicted "not interesting" matches), 10 false positives (incorrectly predicted "interesting" matches), and 5 false negatives (incorrectly predicted "not interesting" matches). Based on this, the calculated metrics would be: Accuracy at 75%, Precision at 71.4%, Recall at 83.3%, and F1 Score at 76.9%. These metrics would offer a quantitative assessment of performance, and we will also review confusion matrices and specific cases of incorrect predictions to gain further qualitative insights.

- What counts as success? What would negative results look like (give a concrete example)?

    - Success for our model is defined by achieving an accuracy and F1 score above a set threshold (e.g., 80%) and closely aligning with our and expert assessments of match interest. For example, if the model correctly identifies high-stakes matches with dynamic commentary (e.g., frequent goals or close calls) as "interesting," it shows the model captures excitement effectively. Conversely, negative results would involve poor evaluation scores, such as low accuracy (near 50%) or high misclassification rates, which may indicate overfitting or misinterpretation of commentary. An example of failure would be if the model mistakenly classifies technical jargon or certain phrases as "interesting," suggesting it hasn’t learned to interpret match excitement accurately.

- What conclusions (if any) can you draw from negative results?

    - Negative results could reveal several critical insights regarding model limitations and data issues. First, the model choice, such as using `distilgpt2 or other model` for classification, might be inappropriate since it is not generalized enough and suited for distinguishing nuanced interest levels in sports commentary. Other models designed for that purpose could potentially be more effective. The dataset size also poses a challenge, as the small sample (303 rows) may be insufficient for learning reliable patterns, leading to weak performance on new data. Additionally, it could mean that current feature extraction may not capture essential commentary details, with preprocessing that might overlook sentiment, key terms, or structural patterns that could better inform the model's decisions.

    - The subjective nature of what constitutes an "interesting" match further complicates modeling, as it introduces potential bias when labels are based on individual judgment. Using a broader data set, including match statistics (e.g., shots on goal, possession percentages), audience reactions, or social media sentiment, could add valuable context. Negative results would suggest re-evaluating the approach by experimenting with alternative models, refining the labeling process with multiple annotators, and gathering more comprehensive data (social media) to better generalize and reduce biases. These results underscore the importance of iterative refinement, data enhancement, and model experimentation, guiding future improvements for a robust predictive model.

Minimal working example

Create a **public github repository (shared with your group)** with the following:

- Sample tsv files that you will input into NLPScholar for each of the different steps (formatted exactly as is required).
  - https://github.com/James-Njoroge/nlp-final/blob/main/lab7_minimal_example.tsv
- README documents that outline in detail the sequence of steps required to go from the dataset to the sample tsv files.
  - https://github.com/James-Njoroge/nlp-final/blob/main/pre-processing.md
- Config files for each of the steps.
  - https://github.com/James-Njoroge/nlp-final/blob/main/train_config.yaml
  - https://github.com/James-Njoroge/nlp-final/blob/main/eval_config.yaml
- Turing job submission files for each of the steps.
  - https://github.com/James-Njoroge/nlp-final/blob/main/nlp_job_scheduler%20-%20final%20project.sh
- README document(s) outlining how you will go from the output of NLPScholar to your evaluation metrics / table/ figures.
  - https://github.com/James-Njoroge/nlp-final/blob/main/metrics-analysis.md