# Male-to-Female Soccer Comparison Generator using Doc2Vec Embedding

Neython Lec Streitz and Jordan Jung

## I. Introduction

Historically, women's soccer was banned worldwide by FIFA, the international soccer association, until 1752. The ban was in place specifically because men's soccer organizations, such as the Football Association of England (the F.A.), were threatened by competition from an emerging popularity in women's soccer. As a result of the ban, women's soccer has always had to operate in the shadow of men's soccer. In this way, part of the reason for the wage gap between men's and women's soccer is because of intentional interference by men's soccer organizations, not just because 'people like watching men's soccer more.' With that said, one solution to the wage gap is reparations in the form of promotions and monetary compensation.

The goal of our project is to create a model that takes open source data from STATSBOMB, international leaders in soccer data analytics, and generates soccer player similarities across men's and women's competitions based on play-styles. To represent the play-styles of each player, our model tokenizes their in-game actions and averages the actions across games using Genism's Doc2Vec model. The end goal then, is to represent each player as an embedded vector, allowing us to easily find the cosine similarity between players.In this way, we can take a popular men's soccer player and find a women's soccer player that is similar to them in the kind of actions they perform during a game. Ultimately, our model could be a gateway for people who enjoy men's soccer to get into women's soccer, promoting the sport and women's soccer competitions.

## II. Related Works

The research for this study is comprised of five statistics articles that focus on the cataloging of professional soccer as well as the proliferating learning systems used for commentary and strategy analysis. "Generating and visualizing a soccer knowledge base," written by Buitelaar et al., features a generated knowledge base extracted from the FIFA and UEFA websites. Using this knowledge base, hyperlinks were automatically created for teams and players, giving short descriptions in much the same way that an online textbook might have highlighted words linked to a glossary. Similarly, "PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences" by van der Lee et al., takes data from Dutch soccer games and generates corpus-based game summaries with an emphasis on tone and other small linguistic features geared towards the person's relation to the game (home fan, away fan, neutral observer). Work by Tanaka et al., showcase an automatic commentary system for soccer matches using live feeds. Likewise, the article "Generating Live Soccer-Match Commentary from Play Data" by Taniguchi et al., also describes a commentary system that takes event data and turns it into live commentary.

While these four papers explore the combination of natural language processing (NLP) and soccer, they do not use their respective models to explore similarities between players. The main takeway from these works then, is that the intersection between NLP and soccer has been explored relatively well within the literature. In regards to our project, we narrowed our search on works that employ the use of vectorization techniques on sports.

Specifically, "A study on the analysis of soccer games using distributed representation of actions and players" by Zhong et al., uses Word2Vec processing on event data in order to analyze the strategies of RoboCup matches. Using its interpretation of teams and players, their learning model generates strategies to combat other teams. Their model is quite related to our own work, yet only features passes and dribbles and does not represent action locations in three dimensions. In this way, we hope to create a more granular model for our own project. Nevertheless, their use of Word2Vec and clustering of players are techniques we are planning on implementing.

## III. Methodology

The StatsBomb open source repository contains all the detailed event data required for this project. This repository includes around 900 different matches between 2003-2020 with over 4,000 men's and women's players. Each event document includes team and competition metadata and labeled event data of player actions. In total, there are about 3 million actions in the dataset, described by their action type, location on the field, player, time, and optional information like success or body part used. Unfortunately, due to RAM constraints, our project only uses 200 matches, instead of the full 900. Even still, we process nearly 1,000,000 actions. These 200 matches were simply taken from the most recently uploaded events to the data set, and not done randomly from the entire set.

In total, the steps of our project include, loading the data, preprocessing the events and player metadata, tokenizing actions, building a player corpus, training the Doc2Vec model, and then finding the cosine similarity scores of all the players compared to each other.

Step one of our project was data processing. Due to the size and granularity of our data set, this was not a trivial matter. Event data and player metadata was loaded from

the STATSBOMB data set, then was enriched by adding extra information. For the event data, this included type of shots (curve, finesse, outside-of-the-foot,...), types of passes (chip, through-ball, to-feet,...), result of the action (goal, pass complete, pass intercepted,...), and location on the field. For the player metadata, we added team names, jersey numbers, player positions, and more.

The next step in the process was to tokenize each action and create a corpus of actions and players. Event tokenization was as simple as representing each action by its name, location on the field, and extra, enriched information. Corpus creation for the actions was done by creating 5 action long 'sentences', in order to chronologize the actions and get a better sense of what actions typically happen consecutively. From this process, we returned a vocabulary, indexed vocabulary, action corpus, action to index dictionary, and index to action dictionary. The process was similar for the player corpus, and included action tokenization and then representing each player by the sum of their actions. In this way, each player was a 'document' made up of action 'sentences'.

With the action and player corpus created, the final step in our model was the training of Word2Vec using the action corpus and the training of Doc2Vec using the player corpus. While the creation of embedded actions was technically unnecessary for the final similarity model, it gave us a sanity check so we knew that actions were being properly represented and compared. The real bulk of the work came in the training of the Doc2Vec model with the player corpus, and then the averaging of player matches to create one representative vector. These average player embeddings were used to find the cosine similarity of each player compared to all other players. The top 10 players with the highest cosine scores were deemed the most similar players, and outputted by our model.

We utilized the UMAP python library in order to plot the player embeddings, which were over 100 dimensions long, on only two dimensions. Players were sorted by position, and in this way, the UMAP graph shows clusters of players based on their actions, color-coded by position.

The other way in which the accuracy of our model was tested was by selecting three popular men's soccer players, Lionel Messi, Luis Suarez, and to find their most similar female counterparts. After generating the similar players, we found news articles for each female player and swapped the names, pronouns, and other identifying features to the most similar male players. The reason we went from female to male player is so that rating the accuracy of the news articles would be easier for soccer fans more familiar with male players. Doing the reverse would require intimate knowledge of women's soccer, which most people, even soccer enthusiasts, do not possess (the primary reason for this project, in the first place). We asked three soccer enthusiasts to read the articles and determine whether the descriptions of the players made sense and matched the way they would describe the play-styles themselves. Thus, we can determine how close to reality the player similarities our model generated are.

## IV. Results and Analysis

To begin, the embedding of actions was largely successful. After running the tokenized actions through Word2Vec, we generated 5 random actions and their 5 most similar positive and negative actions. For instance, for ¡pass¿(1/6, 1/4) its most similar positive match was ¡pass¿(1/6, 0/4). Clearly, a pass is going to be most similar to the same pass just in the adjacent quadrant. The obviousness of these similarities meant that no further experimentation was done solely on the actions themselves.

In terms of the player embeddings, after training our Doc2Vec model and generating the multi-match average embeddings for each player, we plotted each player using the UMAP python library. Players were plotted by color according to their position. As expected, players in the same position

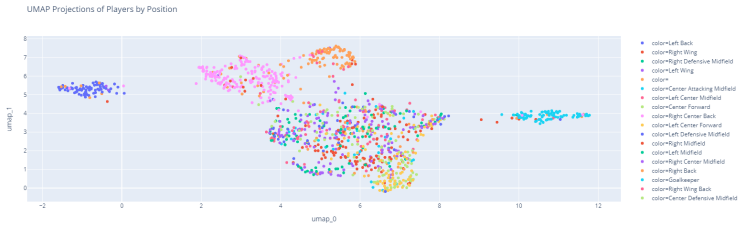


Fig. 1. UMAP plot of average player embeddings

were tightly clustered, and thus had similar embeddings. The grouping on the far left in blue are all 'left backs', whereas the grouping in the far right are all 'goalkeepers.' These are both positions that typically perform very similar actions even across teams. Positions like midfielders were more spread out, as midfielders have more freedom to be 'different' in the sense that they can be more offensive or defensive and do not necessarily share the same role from team to team. As a result, the spread in the center of the plot represents most of the midfielders embedded by our model. Ultimately, the graph reinforces the credibility of our model as it clusters players a soccer fan would consider to be similar.

As for the results of the cosine similarity, we'd like to highlight one especially powerful comparison. The soccer player chosen was Luis Suarez, a prolific goal scorer and forward for FC Barcelona and the Uruguayan national team. The top result of his cosine similarity generated Vivianne Miedema. Miedema is a fantastic forward for her club and country. In fact, Miedema is the all-time leading goal scorer in the F.A. Women's Super League. In a 2020 article called, "Arsenal's Vivianne Miedema sets WSL record in win over Manchester United" by The Guardian, Miedema's manager is quoted as saying, ""I don't worry about Viv not scoring against any opponent. When she hasn't scored against any opponent, it's just a matter of time before she will." Swapping "Viv's" name and pronouns with that of Suarez's proved to be massively successful. All three of our analyzers believed the article to be true and thought themselves that the sentiment was true. Interestingly, none of our analyzers had heard of

Miedema, and true to the goal of our project, expressed interest after learning of the comparison. Similarly, we chose Lionel Messi as another of our players to generate comparisons for. As one of the greatest players to ever play, Messi is a perfect candidate for this project as his reach lays even beyond the soccer community. Therefore, comparisons between Messi and other players might prove to be incredibly meaningful. One of the top results for Messi was Australian forward Samantha Kerr. In another article by The Guardian on Kerr, we swapped the names of Kerr and Messi, specifically for a passage about a cut-inside goal scored by Kerr. This fit perfectly with Messi's reputation and again, all three analyzers though the article to be true. And again, all three analyzers had never heard of Samantha Kerr. Thus, we believe that our model was excellent in generating comparisons between male and female players. Furthermore, by using articles to test our model, we saw the usefulness of the comparisons in action, generating interest from our analyzers towards the two women's soccer players. Likely, the biggest downside to our model is the lack of transparency for why comparisons are being made. The cosine score is a reflection of the actions taken by the players, yet, those actions become numbers in vectors, losing their soccer significance. Only after finding the articles and changing the names were we able to see why those comparisons were likely made. In the future, building a way to retain or return some soccer language as the model generates cosine similarity will vastly improve the transparency of our model.

## V. Threats to Validity

The threats to validity of our project split into two separate categories. The first being threats based on the complexity of representing players through their actions. The second being threats based on the actual methodology used for this project.

To expand further on the difficulty of representing players in terms of vectors, one gaping problem with our approach is the inability to represent the quality of an action. Two players might perform the same pass, making them comparable, yet in reality those passes might be of different quality. That is to say that our model cannot account for how well a player performs an action. In addition, there is no accounting for intangibles like flair, leadership, work ethic, just to name a few. One could derive the selflessness of a player through the number of passes completed, or a similar metric, but of course that is only one facet of what a selfless player might be. A players worth as a teammate also has to do with the general play-style of the team as a whole, meaning some actions might seem good in theory, but be ultimately detrimental to the team. Even though our model would greatly benefit from this increased granularity, the difficulty of doing so makes it inappropriate to incorporate into our model without further careful consideration.

The other threats to validity have to do with obstacles in the production of our project. Unfortunately, we were unable to utilize the complete data set provided by StatsBomb. As Google Colab has certain RAM limitations, the entire dataset frequently crashed the Colab session due to insufficient RAM. As a result, we were forced to use a sample of the dataset, meaning that some players only had a small number of games to derive events from. Of course, a player cannot be represented by the actions they took in only one or two games, so in this way, some players might have skewed representations based on the games that made it into our sample.

## VI. Conclusion

Ultimately, our player model generated accurate and powerful comparisons between male and female soccer players. Our team was happy with the results of the model and found it to be useful in the generation of new women's soccer players to follow. In the future, being able to use the entire dataset would mean stronger comparisons, more accurate player embeddings, and less chance of comparisons being made out of a lack of other players. Additionally, improving the way players are represented to somehow account for the quality of a player's actions as well as the intangible team effect players have would certainly lead to better comparisons. These tasks require much more work and computer power than was available to us at the time, and as such, we do not feel like our model is inadequate or unfinished. Beyond the promotion of women's soccer, we believe models such as ours have the potential to influence sports data analytics. Like the work on Robocup by Zhong et al., our model could be changed to generate strategies, possible team chemistry, or even skill evaluation. Because professional soccer has more data collection than Robocup, our model will likely be even more successful in those respects. With that said, sports analytics is a costly operation, especially if the goal is to be as granular as possible. Thus, our model would be inaccessible for high school and even most college soccer programs. Nonetheless, with the data available, our model does well to create comparisons between men's and women's soccer players. We hope that researchers continue to develop novel ways to promote and engage the public in women's soccer.

## VII. References

### References


[1] Buitelaar et al., "Generating and visualizing a soccer knowledge base." In Demonstrations. 2006.

[2] Tanaka et al., "MIKE: An automatic commentary system for soccer." In Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160), pp. 285-292, 1998.

[3] Y. Taniguchi, Y. Feng, H. Takamura, and M. Okumura, "Generating Live Soccer-Match Commentary from Play Data", AAAI, vol. 33, no. 01, pp. 7096-7103, Jul. 2019.

[4] van der Lee et al., "PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences." In Proceedings of the 10th International Conference on Natural Language Generation, pp. 95-104, 2017.

[5] Zhong et al.,"A study on the analysis of soccer games using distributed representation of actions and players." ICIC Express Letters vol. 13, no. 4, pp. 303-310, 2019.