Abbie Ives

06/12/25

CS 617

**Evaluating NHL Goalie Scouting Reports Using Natural Language Processing and Machine Learning**

**Abstract**

This project applies Natural Language Processing (NLP), machine learning, and clustering techniques to analyze pre-draft scouting reports of NHL goaltenders. Using a dataset of reports on goalies drafted between 2010–2020, the pipeline investigates whether the language used by NHL scouts correlates with long-term career success. Despite sophisticated modeling, results show that these reports do not meaningfully differentiate successful and unsuccessful goaltenders. This provides strong evidence that current scouting processes fail to capture the qualities that truly matter for goaltender success—indicating the need for a fundamental shift in how goalie talent is evaluated.

**1. Introduction**

From my experience as a former NCAA Division I and professional goaltender, and later as a coach, I have consistently observed a systemic gap in how goalies are scouted. It is well understood in the hockey world that NHL scouts struggle to accurately project goaltending success. I believe this failure stems from a fundamental misunderstanding of the position by decision-makers.

This project tests the predictive power of pre-draft scouting reports using NLP and machine learning. If these reports truly reflected a deep understanding of the position, we would expect the language to vary between successful and unsuccessful goaltenders. However, I hypothesize that this is not the case—and that the language used in these reports does not meaningfully correlate with future NHL success. The goal of this project is to provide empirical support for this claim and call into question the validity of the current goalie scouting process.

**2. Data Collection**

The dataset includes:

* Pre-draft scouting reports for NHL goalie prospects
* Metadata: Draft Year, NHL Games Played, NHL Save Percentage
* Source: *The Hockey Writers*, which quotes actual NHL scouts in pre-draft profiles to ensure consistency in language and tone

**3. Data Preprocessing**

**3.1 Success Labeling**

Goaltenders were categorized into four success groups based on career NHL games played:

* Elite: ≥ 500 games
* Successful: ≥ 100 games (pre-2015), ≥ 50 games (post-2015)
* Average: Between these thresholds
* Unsuccessful: ≤ 20 games

For model training, a binary Success flag was also created:  
1 = Elite/Successful, 0 = Average/Unsuccessful

**3.2 Text Cleaning**

The text of scouting reports was cleaned by:

* Converting to lowercase
* Removing punctuation, numbers, and common stopwords
* Removing domain-specific filler words such as “goalie,” “puck,” “saves,” and names of well-known goalies

**4. Text Vectorization & Classification**

**4.1 TF-IDF Vectorization**

Scouting report text was transformed into numerical vectors using TF-IDF (Term Frequency–Inverse Document Frequency), allowing the model to identify which words are most distinctive in the context of all reports.

**4.2 Classification Model**

A Logistic Regression model was trained on 70% of the data and tested on the remaining 30%. Despite tuning, the model only achieved 39% accuracy, with low precision, recall, and F1-scores—indicating poor predictive power.

**Key Insight:** If scouting reports truly reflected meaningful differences, we would expect at least moderate classification accuracy. The fact that even a basic binary classification task fails suggests the language in these reports is not predictive of future success.

**5. Lexical Analysis**

The most frequent words in each success group (Elite, Successful, Average, Unsuccessful) were analyzed. The results revealed:

* Highly similar language across all groups
* Common terms included “quick,” “calm,” “strong,” “big,” and “athletic”—regardless of actual career outcome

**Recurring Language Observation:** Words like “athletic” and “quick” are essentially prerequisites for being drafted into any sports league and therefore offer no meaningful differentiation. These are *baseline traits*, not predictors of long-term success.

**6. Sentence Embeddings & Clustering**

**6.1 Semantic Embedding**

Reports were embedded into dense vector representations using SentenceTransformers (all-MiniLM-L6-v2) to capture underlying semantic meaning.

**6.2 K-Means Clustering**

Using K-Means (k=3), reports were grouped based on similarity in language. However:

* Each cluster contained a mix of both successful and unsuccessful goalies
* No cluster consistently captured a particular success group

**Clustering Insight:** The semantic similarity of reports, regardless of career outcome, strongly reinforces the idea that scouting language is too generic or shallow to identify future NHL goaltending talent. This further demonstrates a lack of understanding goaltending

**7. Visualization**

**7.1 Similarity Heatmap**

A cosine similarity matrix was plotted to visualize how similar reports are to each other. Large, indistinguishable clusters further support the lack of language differentiation.

**7.2 Word Trends Over Time**

Key words such as “athletic,” “reflexes,” and “positioning” were tracked across draft years. Their flat or erratic usage patterns suggest no consistent evolution in how goalies are evaluated.

**Expert Insight:** In my experience, qualities like *patience*, *mental resilience*, and *positional discipline* are far more predictive of success—yet terms like “patience” appeared only once in the entire dataset.

**8. Conclusion**

The results provide clear evidence that NHL pre-draft goalie scouting reports do not meaningfully predict future success:

* The same descriptors are used for both successful and unsuccessful goalies
* The classification model fails to distinguish outcomes based on report text
* Clustering shows no language-based separation between groups

**Theory Validated:** If scouting reports were effective, we would expect some linguistic signals that differentiate successful goalies. Their absence supports my theory that scouting language is failing to capture the traits that matter most—not that such traits don’t exist.

**Major Insight Gained:** I believe there are meaningful qualities that separate future NHL goalies from others. The issue is that the current scouting process fails to identify or articulate them.

**Call to Action:** Goalie scouting needs to be rethought entirely—with deeper technical understanding, better language, and data-driven evaluation.

**Appendix:** Technologies Used

* **Libraries:** pandas, nltk, scikit-learn, sentence-transformers, matplotlib, seaborn
* **Models:** TF-IDF vectorizer, Logistic Regression classifier, Sentence Embedding (MiniLM), KMeans clustering
* **Visualizations:** Similarity heatmaps, word usage trends, clustering samples