


MIE490 Capstone Design
Columbus Blue Jackets

Final Design Specification

Deniz Nalbantoglu

deniz.nalbantoglu@mail.utoronto.ca

1004501470

Keaton Smith

keaton.smith@mail.utoronto.ca

1003100397

Yicheng Pan

yicheng.pan@mail.utoronto.ca

1002940213

Word Count: 7557

Date: Mar 26th, 2021

Contents

Executive Summary **1**

1 Project Background and Motivation **2**

2 Problem Description **2**

3 Scope Definition **3**

4 State-of-Art Review **4**

5 Requirement Specification **5**

5.1 Identifications of Stakeholders 5

5.1.1 Columbus Blue Jackets Scouts 5

5.1.2 Prospective and Current NHL Players 5

5.1.3 Columbus Blue Jackets Coaching Team 6

5.1.4 Columbus Blue Jackets Management Team 6

5.2 Functions 6

5.3 Objectives 7

5.4 Constraints 7

5.5 Service Environment 7

5.6 Design for Modularity 8

6 Solution Methods **8**

7 Idea Generation and Evaluation **8**

8 Sentimental Analysis Solution - VADER Score **10**

9 Process Model Architecture **11**

10 Model Specification **12**

10.1 Text Processing and Scraping 12

10.1.1 Static Scraping 13

10.1.2 Keyword Categorization 14

10.2 VADER Scoring Module 14

10.3 Output Analysis 15

10.3.1 Skill Adjustment 15

10.3.2 Proof of Concept: VADER Score vs. Written Phrases 17

10.4 Model Limitations 18

11 Implementation and Test Plan **19**

12 Environmental, Social, and Economic Analysis **20**

12.1 Social Impact 21

12.2 Economic Impact 21

12.3 Environmental Impact 22

13 Next Steps and Future Work **22**
13.1 Additional Textual Adjustments for Keyword Library 22
13.1.1 Stemming 22
13.1.2 Additional Text Screening for Implicit Sentiments 23
13.2 Functional Combination with BI Software 23
13.2.1 Tableau 23
13.2.2 PowerBI 23
13.3 Skill Adjustment Enhancements 23
13.4 Source and Date Adjustment 24
14 Conclusion **24**
References **25**
Appendix I - Keyword Library Developed by CBJ Capstone Team **27**

Executive Summary

Analytics and scouting reports provide a diverse set of insights for National Hockey League (NHL) player evaluation departments. analyses which need to be manually combined to create comprehensive player evaluations. The Columbus Blue Jackets (CBJ) requested an efficient methodology to assess the performance of players from a combined analytical and scouting report perspective. Thus, the project started with the motivation to find an effective way to find the optimal position of players based on their performance to find a combined approach. Thus, the primary function of the proposed design is to combine qualitative textual inputs from scouting with statistical performance data analysis to provide a comprehensive evaluation of players.

The proposed model mainly transforms textual inputs from scouting reports into more objective and quantifiable metrics using VADER Scores, which can capture the 'emotion' and sense of the text. The scores applied on scraped text from a scouting report source are based on four different main skill categories: shooting, skating, puck skill, and character. The model also proposes a flexible option of adjusting scores manually to reflect any optional team philosophy or perspective on the source. This allows the model to also capture the human element in the dynamic hockey game. However, the model has some implications as it is not strong in capturing any specific jargon used in the text. For instance, it may not be able to fully capture the hockey-specific language or any irony made in the scouting reports.

Since the design includes a new method for evaluating the performance of hockey players, it has a significant social impact on both CBJ and the hockey analytics industry. The CBJ would mainly benefit from the design by having a competitive advantage of using a novel and efficient method that would be able to transform textual scouting reports into quantifiable metrics. Besides, the rivals of the CBJ would be potentially interested in the design and this would result in competition. Furthermore, the operation of the design does not significantly affect any job roles but it could make their tasks more efficient.

To conclude, the final design provides a combined methodology to textual scouting reports with statistical tools to provide a metric for decision-making in player evaluations. The design would create a massive competitive advantage for Columbus and can be adapted to any competitive sports industry around the world.

1 Project Background and Motivation

Analytics have become a core component to player evaluation departments across the National Hockey League (NHL), being used as another method to gain a competitive edge. Using statistical models, teams can acquire a quantitative perspective of a player, including predicting a player's future salary and forecasting their future NHL production based on their production in different leagues [1; 2]. These models are used alongside scouting reports, which provide a subjective, qualitative perspective on players' attributes and skills, to create comprehensive player evaluations which guide various personnel decisions, including drafting prospects, signing free agents, and executing trades.

While analytics and scouting reports provide some overlapping insights, they are often treated as mutually exclusive analyses. This leaves the onus on NHL front office decision makers and analysts, including the Columbus Blue Jackets project client, to manually combine the analyses to create comprehensive player evaluations. Existing solutions are capable of independently generating insights from a quantitative and qualitative perspective, but have not proven to be capable of combining these perspectives into comprehensive, professional-quality evaluations. In this lies the motivation behind this project: to create a combined qualitative and quantitative perspective to improve player evaluation efficacy for the Columbus Blue Jackets.

2 Problem Description

The quantitative statistical analyses provide an objective, unbiased perspective of player performance for the data that it possesses, but statistics are incapable of capturing the full essence of player performance. As a result, a number of player qualities are either poorly evaluated by statistical analyses, such as skating and leadership acumen, or the data creates misrepresentations of a player's abilities, such as goal totals reflecting shooting ability.

The qualitative scouting reports are able to incorporate a human perspective into player evaluation, and are more readily equipped be able to evaluate the qualities which are not represented by statistics. However, they are subjected to differences in interpretation and bias, which can create an inconsistent and inaccurate perception of a player's abilities.

Given the contradictory set of strengths and weaknesses for each perspective, the potential for inaccuracy and misinterpretation of player performance is discernible without a combined perspective. To reduce this impact, Columbus requires an accurate approach to objectively define the subjective textual inputs of the scouting reports and the ability to combine them with their objective statistical analysis. This would improve the efficacy of player evaluations and provide them with a competitive edge in their player personnel decisions.

3 Scope Definition

The foundation of this project can be defined based on three key components as shown in Figure 1:

A. Quantitative analysis involving statistical inputs and models.

B. Qualitative analysis involving scouting reports.

C. A mechanism whereby the client is able to utilize the quantitative and qualitative analyses to be able to create combined player evaluation insights.

Based on this foundation, the team's industrial knowledge and educational experience, the project timeline, and the client's needs, the scope of this project has been defined as follows:

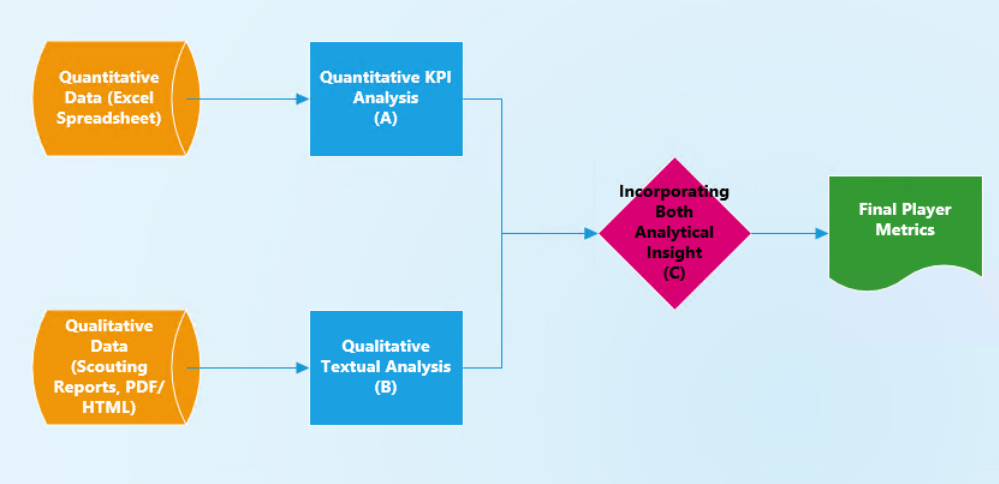

Figure 1: Simplified Flow Chart of the Design

1. Incorporates the format of the scouting reports and statistical data used by client.
2. Extracts and analyzes textual inputs from scouting reports to create a quantifiable sentiment with minimal bias.
3. Allows for the client to prioritize the date and source of scouting reports.
4. Allows for the client to prioritize skills based on the type of player being evaluated and team preference.
5. Designed with modularity so the client has the flexibility use each analysis independently or in different combinations through a centralized mechanism to create varying insights.

The final result should be allow for the client to create a comprehensive player evaluation which incorporates the statistical and textual analysis result for each player further reflect upon the a player's overall performance and capabilities.

4 State-of-Art Review

Countless projects which incorporate sports analytics to provide more complete player evaluations are available for public perusal on the internet. The vast majority of these projects focus upon the utilization of publicly available statistics to make various predictions and determinations about a player's performance and value. For example, predicting a player's future salary and forecasting their future production based on their production in different leagues [1; 2].

Conducting research into projects which utilize analytics in combination with qualitative means of player evaluation is far less fruitful. Few projects investigate the potential of transforming subjective scouting reports into a form which allows you to objectively evaluate the sentiment of a player, and combine those quantifiable sentiments with other quantifiable means of player evaluation. Fortunately, while few in number, similar projects investigating these possibilities are publicly available.

"Text Mining of Scouting Reports as a Novel Data Source for Improving NHL Draft Analytics", by Timo Seppa, Michael E. Schuckers, and Mike Rovito, was the inspiration behind the client's project proposal [3]. As shown in Figure 2, the project explored earlier attempts of combining scouting report information with traditional analytics. The project used this information to develop the framework of a tool which utilizes sentiment analysis and text mining of text-based scouting reports to quantify a scout's perception of a player based on a number of attribute categories. The research in this paper yields promising results and has demonstrated a proven approach for text mining and sentiment analysis of hockey scouting reports, but lacks the practical execution and only explores a specific set of statistical techniques and data.

Figure 2: Phrase extraction from NHL scouting reports to determine sentiment of various player attributes. While all 10 phrases use a variation of the term "compete hard", the project has been able to differentiate positive sentiment (highlighted in green) and negative sentiment (highlighted in red) within each phrase.

"The Race to Make the NFL Draft an Exact Science", by Kevin Clark, attempts to solve a similar problem in the context of American Football [4]. National Football League (NFL) teams scout teams from the collegiate ranks to find future talent, but face a similar set of challenges as the NHL when it comes to player evaluation. Some player attributes, such as scheme fit or football IQ, are poorly quantifiable and

require scouts to subjectively evaluate the efficacy of drafting a particular player. Additionally, while there are traditional stats such as sacks or interceptions, the raw numbers often do not create a comprehensive representation of players' current or future abilities. External factors, such as quality of teammates or quality of opponents, may skew these numbers, leaving scouts to subjectively confirm the legitimacy of the players' stats. While this project does not directly explore the quantification of subjective scouting reports, it discusses how NFL scouts can directly incorporate analytics to add another dimension to their player evaluations.

5 Requirement Specification

The requirements for this project have been aligned to successfully resolve the problem specified in the 'Problem Description' section and will guide the subsequent design phase.

5.1 Identifications of Stakeholders

The stakeholders have different interests which could influence of the direction of the design process. It is important to account for these interests to create an effective and useful design solution. Stakeholders who are not directly impacted by the implementation of the design, such as the University, or are already considered in the design process, such as the client and the team, are not listed.

5.1.1 Columbus Blue Jackets Scouts

Interests:

1. Accurate representations of their scouting reports in player evaluation results.
2. Minimal work required to prepare their reports to be used by the design.
3. Minimal changes to their existing process to accommodate the design.

Impact on Design:

1. Reduce unnecessary disruptions and work in scouting process after design implementation.
2. Ensure scouts' evaluations are accurately reflected.

5.1.2 Prospective and Current NHL Players

Interests:

1. Fair representation of their attributes and abilities.
2. Positive and negative changes in perception of their profile and their resultant change in receiving NHL opportunities.

Impact on Design:

1. Account for biases that may persist from both the scouting and analytical inputs.
2. Ensure there are no unethical and discriminatory components of the design.

5.1.3 Columbus Blue Jackets Coaching Team

Interests:

1. Comprehensive knowledge of players on both Columbus and opposing teams to effectively plan game strategies and practice sessions.
2. Composition of roster and which players are available for the coaching team to use in a game.

Impact on Design:

1. Ensure overall player evaluations are accurate and representative of players' abilities.
2. Ensure the results are easy to understand and are useful when making coaching decisions.

5.1.4 Columbus Blue Jackets Management Team

Interests:

1. Comprehensive knowledge of players around the world to identify undervalued players, gain a competitive edge in signing and trade negotiations, and improve roster construction.

Impact on Design:

1. Ensure overall player evaluations are accurate and representative of players' abilities.
2. Ensure the results are easy to understand and are useful when making management decisions.

5.2 Functions

The functional basis of this design is to combine qualitative and quantitative analysis results to create a comprehensive evaluation of players. The design must satisfy all of the primary functions to be considered an acceptable design. The secondary functions enable or result from the primary functions.

Primary Functions:

1. Provide statistical analysis results based on quantitative player performance data.
2. Provide textual analysis results from qualitative scouting reports.
3. Provide comprehensive player analysis results based on qualitative and quantitative inputs.

Secondary Functions:

1. Able to scrape statistical data from Excel Spreadsheets (.xls, .xlsx).
2. Able to scrape textual information from scouting reports (HTML, pdf).

3. Able to run from a Python or Anaconda Command Prompt.

4. Showing or temporarily storing player's historical data.

5.3 Objectives

The objectives of this project are designed to create more appealing and effective solutions for the client, but are not necessarily mandatory to provide a quality design for the client. Objective prioritization is listed in descending order in Table 1, with the highest priority objectives listed at the top of the table.

Table 1: Objectives for Columbus Capstone Project Deliverable

Objective	**Goal**	**Metrics**
Fast to Compute	Minimizing computational time without compromising the quality of analysis	Compute a thorough metrics for a player within 10 seconds
Easy to Use and Train	To make the design intuitive to clients and his associates.	Able to train a new user to use this design within an one hour session.
Inexpensive	To make sure the design will have no additional cost to clients	Should inflict 0 CAD$ of extra cost

5.4 Constraints

The constraints are the fundamental rules along with quantifiable minimum requirements that must be fulfilled by the design in order to be considered acceptable.

1. The design must not cause a hardware downtime of more than 1 hour from initiation to completion.

2. Input preparation and formatting of scouting reports and quantitative inputs must not require more than 1 hour of manual labour.

3. The solution must not compromise the security of any of the scouting reports or other proprietary information that is inputted into the design.

4. The implementation and utilization of the design must not require any additional costs to be incurred.

5. The design must not induce significant changes in the scouting process or otherwise burden the scouting staff to be functional.

5.5 Service Environment

The design will operate in a solely virtual service environment, and must be able to effectively function within following set of requirements to be considered viable. However, it is possible that exceeding these requirements will lead to better performance[5; 6].

1. Microsoft Windows 7, Microsoft Windows 10, or Mac OS X 10.11.

2. Python Version 3.6.0 or above to accommodate Python packages and sub-packages[5].

3. x86 64-bit CPU (Intel/AMD architecture), 4 GB RAM and 5 GB free disk space[6].

5.6 Design for Modularity

Modularity would allow for the Columbus Blue Jackets to utilize any combination of quantitative and qualitative inputs to generate unique sets of insights into players. For example, if the user would like to only analyze certain defensive factors of players, they can choose to only input the stats and qualitative inputs relevant to those factors and gain useful insights. Additionally, the design should also allow the user to choose to only analyze scouting reports or statistical analyses independently.

6 Solution Methods

The main solution method would involve the combination of quantitative and qualitative analysis to assess a player. The quantitative analysis would be based on the analysis of the statistical inputs and models while the qualitative analysis would be based on the textual inputs that come from scouting reports. The main tool that both approaches will be used is Python; an object-oriented programming language that is commonly used for data analytics projects due to its simplicity and capability of incorporating machine learning and data scraping into its process.

The quantitative aspect of the solution method will be based on raw statistics and data which includes the statistical inputs for relevant players. Microsoft Excel will be used to format this data into CSV files so that it can be easily imported into Python. Python's statistical libraries, such as Numpy, SciPy, and Pandas will be used for to perform the statistical analysis on the data, while visualization libraries such as Seaborn and Matplotlib will be used to visualize the results. External tools, such as Tableau, may also be considered for visualization depending on how the results can be formatted.

The qualitative aspect of the solution method will be based on using the BeautifulSoup text mining library in Python to scrape the textual data, and then to use sentiment analysis modules in Python to determine how a scout evaluated each skill for each particular player. Python statistical libraries, along with Excel spreadsheets of adjustment factors based on report date and source, will be used to create the final quantifiable result of scouting reports.

7 Idea Generation and Evaluation

With the modular nature of this project, morphological analysis combined with brainstorming was the ideal tool for idea generation. First, the overall problem was broken down into each of its characteristics: Text Scraping (HTML and PDF) and String Processing, Sentiment Analysis, Sentiment Analysis Adjustment, and the final Qualitative and Quantitative Combination. For the purpose of this Design Review, only first

three characteristics are in focus as they are responsible for the quantification of the textual scouting reports. Various tools and methodologies were brainstormed to satisfy each problem characteristic with solutions that satisfied the project functions and constraints outlined in Section 1 summarized below:

- **HTML Text Scraping**: BeautifulSoup in Python and 'rvest' in R
- **PDF Text Scraping**: Py2PDF in Python and 'tabulizer' in R.
- **Text Processing**: 'nltk' package in Python and string functions in Python.
- **Sentiment Analysis**: VADER Scores, Word2Vec, and GloVe.
- **Sentiment Analysis Adjustment**: Manual Adjustments, Neural Networks, and Machine Learning.

The chosen solution for each problem characteristic, outlined in the Table 1, Table 2, and Table 3 evaluation matrices, was determined based on its abilities to meet the objectives of the project outlined in Section 1. All of the chosen solutions were found to be compatible and can be implemented with Python.

Table 2: Evaluation Matrix for HTML Text Scraper, PDF Scraper, and Text Processing

	Accuracy/Bias	Computing Time	Intuitive	Inexpensive
BeautifulSoup	4	4	4	5
rvest	4	3	3	5
Py2PDF	4	4	4	5
tabulizer	4	3	3	5
nltk	4	4	3	5
String Functions	4	2	2	5

BeautifulSoup, Py2PDF, and the 'nltk' packages chosen for the HTML Text Scraper, PDF Text Scraper, and Text Processing respectively, as each were unanimously evaluated to best satisfy all objectives.

Table 3: Idea Evaluation for Sentiment Analysis

	Accuracy/Bias	Computing Time	Intuitive	Inexpensive
VADER	4	4	4	5
Word2Vec	2	2	3	5
GloVe	2	2	3	5

VADER Scores were chosen to complete the Sentiment Analysis portion of the project, as it was unanimously evaluated to best evaluate all objectives.

Table 4: Idea Evaluation for Sentiment Analysis Adjustment

	Accuracy/Bias	Computing Time	Intuitive	Inexpensive
Manual	2	5	5	5
Machine Learning	4	3	3	5
Neural Networks	4	3	3	5

Although manual manipulation is less accurate and more subjected to user bias, it does not require Machine Learning or Neural Network knowledge to implement and modify and allows the user to control the outputs

based on their needs and beliefs.

While the manual sentiment analysis adjustments have yet to be implemented, the VADER Scores are able to be score approximately 90% of the text scraped from the HTML and PDF Scrapers. Additionally, the 'nltk' package is effectively able to identify and categorize sentences based on their skill category before they are scored, increasing the amount of text and viable scouting reports that can be processed.

8 Sentimental Analysis Solution - VADER Score

In order to satisfy the functions and constraint mentioned above, the team have brainstormed several sentimental analysis methods for this project and had finally narrowed down to two methods, Word2Vec and VADER (Valence Aware Dictionary for Sentiment Reasoning) model.

In the Word2Vec method developed by Tomas Mikolov, the embedded words will be transformed into vectors with numbers, which preserve the syntactic regularity (similar word have similar vectors) and the semantic regularity between the word paired through vector algebra. For instance, the word vector between the word "boy" and "prince" will be similar with the vector between "girl" and "princess". These vectors will then be used to train the machine learning sentimental analytic model while preserving the relationship between words. The nature of Word2Vec model is a neural network with input, projection and output layers for word vectors [7]. Word2Vec will require a large amount to train, more complex text embedding and model building. Even then, the correct extraction of the sentiment is not guaranteed and it's dependant on the scouting report wording environment. Due to the lack of precedent work, the team was unable to find a labeled testing set to validate the model, which is another reason why the team had proceeded with VADER score as the sentimental analysis solution.

In comparison, VADER score is much more progressive and accommodating to the modern day semantics. Originally designed for social media analysis, VADER is a lexicon- and rule-based sentiment analysis tool that can quantify the sentiment of words, sentences, abbreviations, slang, as well as punctuation marks. It's known for its efficiency, as the model is able to deal with bulk of texts and requires no training.[8] The dictionary within VADER was humanly labeled, and is updated timely where the develops keeps adding the sentiments for new slang or abbreviations to enrich the dictionary.

VADER is a model that considers the polarity (positive/negative) and the intensity of emotion within a group of texts, and further quantify them into a compound score. The scale of VADER score is -4 to 4, where -4 is the most negative, +4 being the most positive and 0 being perfectly neutral. To quantify the sentiment of the input texts, VADER makes use of five heuristics to incorporate the impact of each sub-text on the perceived intensity of sentiment within sentences[9] as shown in Figure 3. The compound score will be used to assess the scouting report sentiment in respect to player attributes. Please see below as some of the common sentiment adjustment (the main five heuristics) of VADER score.

1. Punctuation, namely the exclamation point "!". This punctuation would act as a sentiment enhancer in a sentence. It would increase the sentiment intensity without modifying the sentence structure.

2. Capitalization, specifically the usage of ALL-CAP words within a sentence. Similar to the exclamation point, this is another form of sentimental enhancement, and it would also increase the sentiment intensity without modifying the sentence structure.

3. Other Verbal Degree Modifiers, also known as degree adverbs (words like very, slightly, extremely, etc.). By adding those words in front of adjectives or verbs, it would impact sentiment intensity by either increasing or decreasing the sentimental intensity.

4. Polarity shift due to Conjunctions. In a plain sentence, there are no emphasis on intensity on a certain section of a sentence. However, if a sentence contains words like "but" and "however", it would demonstrate a shift in sentiment polarity, with the sentiment of the text after the conjunction dominating the sentiment of the sentence.

5. Catching Multiple Polarity Negation, VADER is able to catch nearly 90% of cases where negation changes the polarity of the text, by examining the continuous sequence preceding a sentimental lexical feature within a sentence. The sentence "His skating isn't really that bad." would imply a positive sentiment where the polarity of the sentence has changed two times.

```
Tim Stützle is a good skater.
{'compound': 0.4404, 'neg': 0.0, 'neu': 0.58, 'pos': 0.42}
Tim Stützle is a GOOD skater.
{'compound': 0.5622, 'neg': 0.0, 'neu': 0.524, 'pos': 0.476}
Tim Stützle is a good skater!
{'compound': 0.4926, 'neg': 0.0, 'neu': 0.556, 'pos': 0.444}
Tim Stützle is a very good skater.
{'compound': 0.4927, 'neg': 0.0, 'neu': 0.61, 'pos': 0.39}
Tim Stützle isn't a bad skater but his shooting is hard to watch.
{'compound': 0.3612, 'neg': 0.105, 'neu': 0.677, 'pos': 0.218}
```

Figure 3: Different Adjustments in VADER score Analysis

In this project, the team will be using the compound VADER score to quantify the sentiment in different sections of scouting report.

9 Process Model Architecture

This model transforms textual scouting reports into quantifiable metrics using VADER Scores, as outlined in Section 8. The insights from the scouting reports can then be directly incorporated into other quantitative, analytical models or used independently to quantitative compare players based on how the scouts view each

of them. Scores are broken down into 4 key skill categories: Skating, Shooting, Puck Skill, and Character (which includes defensive ability, physicality, and overall intangible leadership and personality traits). Figure 4 outlines the architecture of the entire process from text mining a set of scouting reports that the user provides at the start to outputting a final adjusted VADER Score at the end.

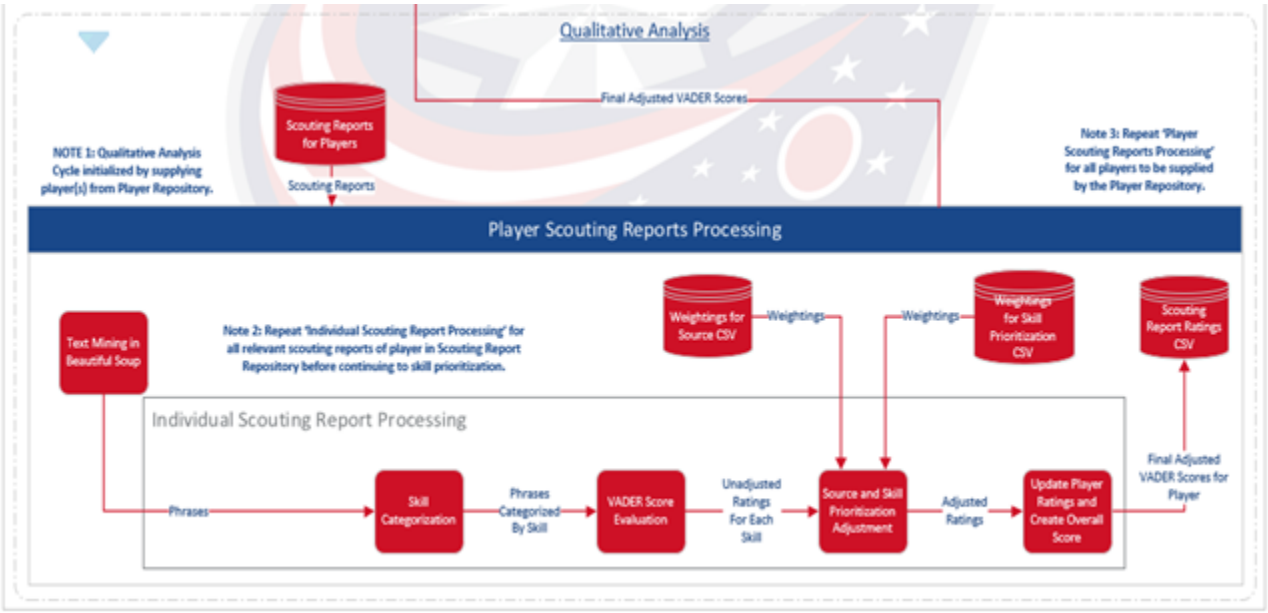

Figure 4: Scouting Report Sentiment Analysis Architecture

10 Model Specification

This section is dedicated to present a detailed walk-through of the sentimental analysis model. The three main steps of this model are: **Text processing and scraping, VADER scoring module, and output analysis**.

10.1 Text Processing and Scraping

The team have explored a variety of different scouting report sources. After consulting with the clients, the team had determined to accommodate two different sources in this project, PDF and Webpages. Due to the different writing styles of scouting reports, the team have categorized the report into 4 different sections that dedicated to unique player attributes: **Skating, Character, Puck Skill and Shooting Skill**. The goal of this process is to extract, clean up, and categorize the scouting report texts to make it ready of the subsequent VADER scoring module. Due to the different writing styles of different sources, some of the scouting reports have already divided up into sections with respect to player attributes where others have not. To accommodate this issue, the team have developed two different scraping method: **static scraping and keyword categorization**.

10.1.1 Static Scraping

One of the scouting report sources was Lastwordonsoprt.com, where the scouting reports were divided into different sections to discuss different player attribute. For this source, the team proceeded with the "Static Scraping" technique, where the sentences and paragraphs stay the same, and the scraper would simply extract all the report information and put them into player attribute categories. The scraping process follows the steps below and is illustrated in Figure 5:

1. The scraper first access the online annual scouting report repository for lastwordonsports.com,
2. From there, the scraper will be able to access all the player scouting report in that year through the links on the web page.
3. The scraper then access the HTML code for a certain player's scouting report, and extract all the HTML code that contains report information.
4. It will then clean up noises within the HTML code(ads, HTML codes, other useless information, etc.) and leave only the player and report information.
5. Since the report has already been divided into different player categories, the scraper will categorize text into the predetermined player attributes.

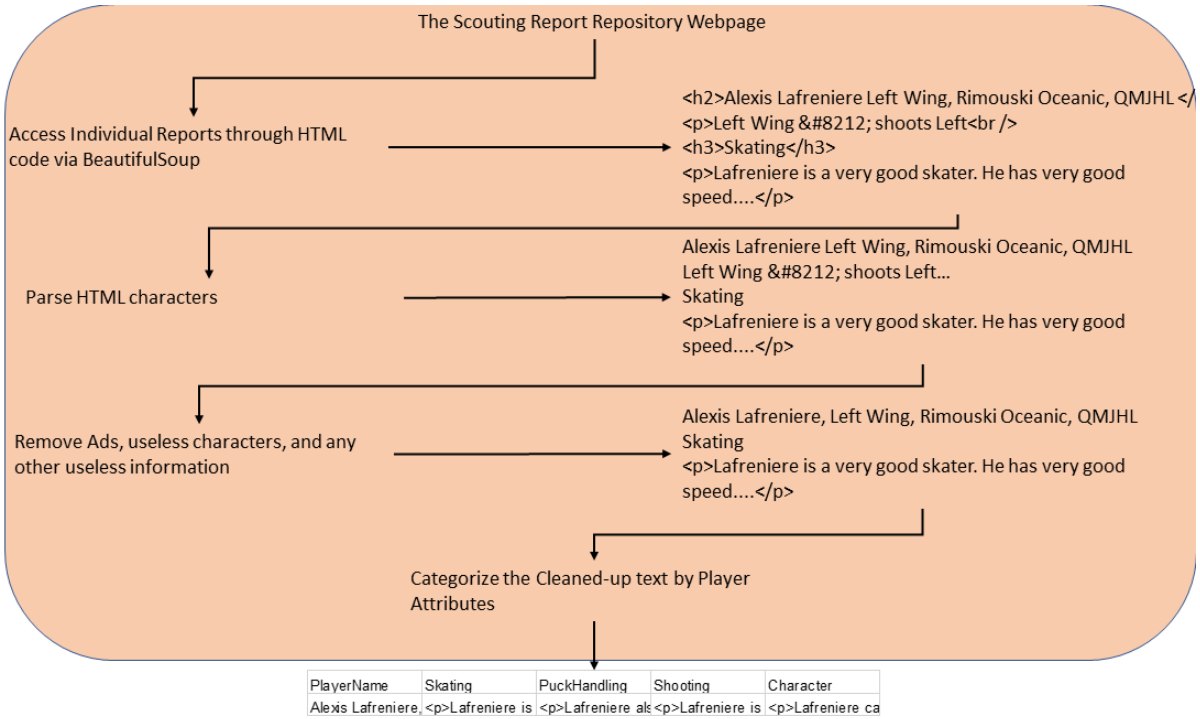

Figure 5: Static Scraping Process

10.1.2 Keyword Categorization

The keyword library categorizes sentences scraped from websites and PDF's based on predefined player attributes through the use of a dictionary with a set of noun and verb keywords. VADER Scores will then be able to be performed on all sentences corresponding to each attribute at the same time. Figure 6 outlines the steps of teh Keyword Categorization process.

1. The scraper first access the online annual scouting report repository for the source, in this case, it could be a large PDF that contains all the player scouting reports from that year.
2. The team then did some re-formatting of the PDF document, which was to make sure each page only contains one player scouting report to make it easy to parse in next steps.
3. The scraper then access the scouting report, and break it down by sentences. This transforms the scouting report in a list of sentences.
4. The list of sentences then go through the keyword screen, and will be categorized into the predetermined player attribute categories.

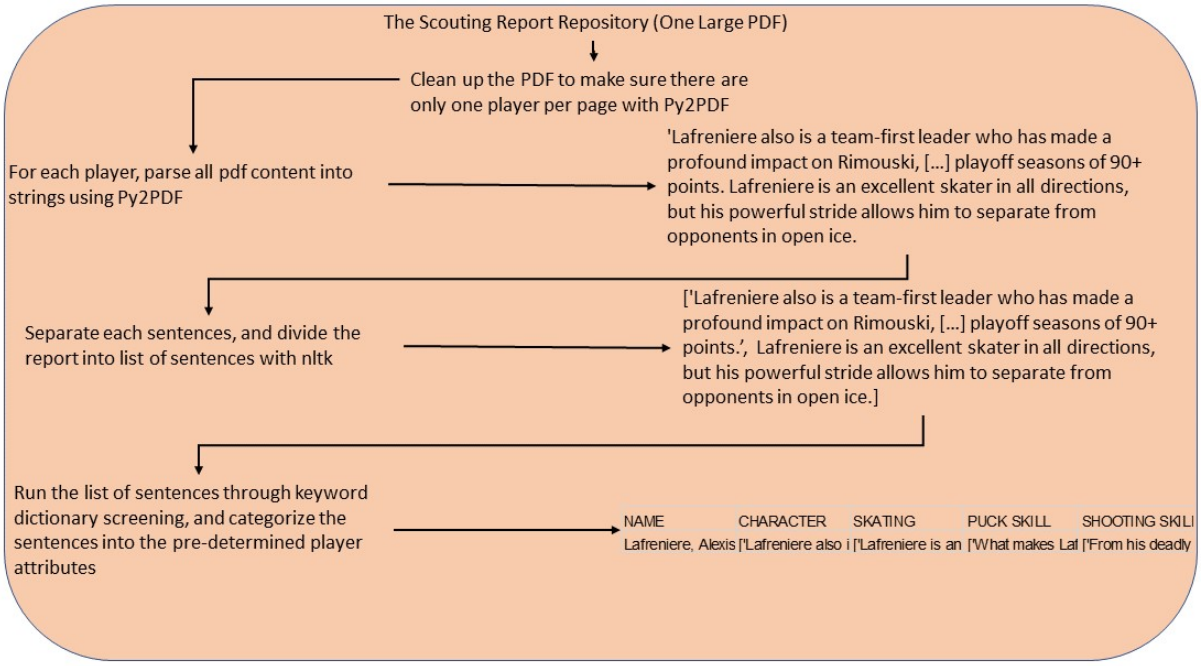

Figure 6: Keyword Screening Process

Both of the scraping process above will produce a .csv file that contains all the categorized player scouting report information for the next scoring step.

10.2 VADER Scoring Module

The VADER scoring module will take the .csv file form the previous step, read it into a Pandas data frame, and provide a compound polarity score with nltk.VADER module based on the content of categorized

scouting reports, as shown in Figure 7.

NAME	SOURCE	CHARACTER	SKATING	PUCK SKILL	SHOOTING
Alexis Lafreniere, Left Wing, Rimouski Oceanic...	LWOS 2020	<p>When Lafreniere is on the ice at the junior...	<p>Lafreniere is a very good skater. He has ve...	<p>Lafreniere also has great hands, allowing h...	<p>Lafreniere is also a natural goal scorer. H...
Quinton Byfield, Centre, Sudbury Wolves, OHL	LWOS 2020	<p>Byfield is also already well-developed in h...	<p>Byfield has everything that teams want in a...	<p>Byfield also has the hands to make moves wi...	<p>Byfield is also an excellent playmaker. He ...
Tim Stutzle, Centre/Left Wing, Adler Mannheim...	LWOS 2020	<p>Stuzle's defensive game is a bit of a work ...	<p>Stutzle is amongst the best skaters in the ...	<p>Stutzle adds to his impressive skating with...	<p>Stutzle sometimes has a tendency to over-pa...
Jamie Drysdale, Right Defence, Erie Otters, OHL	LWOS 2020	<p>Drysdale can also get back in his own end a...	<p>Drysdale is one of the best skaters in this...	<p>Drysdale can rush the puck up the ice, maki...	<p>Drysdale could use some work on his slap sh...
Lucas Raymond, Left Wing, Frolunda HC, SHL	LWOS 2020	<p>Raymond is willing to work in all three zon...	<p>Raymond is a fantastic skater. He has an ex...	<p>Raymond combines his excellent skating abil...	<p>Raymond also has a good wrist shot and a qu...

NAME	SOURCE	CHARACTER	SKATING	PUCK SKILL	SHOOTING	RANK	CHARACTERVADER	SKATINGVADER	PUCKSKILLVADER	SHOOTINGVADER
Alexis Lafreniere, Left Wing, Rimouski Oceanic...	LWOS 2020	<p>When Lafreniere is on the ice at the junior...	<p>Lafreniere is a very good skater. He has ve...	<p>Lafreniere also has great hands, allowing h...	<p>Lafreniere is also a natural goal scorer. H...	NaN	0.9442	0.9622	0.9701	0.9578
Quinton Byfield, Centre, Sudbury Wolves, OHL	LWOS 2020	<p>Byfield is also already well-developed in h...	<p>Byfield has everything that teams want in a...	<p>Byfield also has the hands to make moves wi...	<p>Byfield is also an excellent playmaker. He ...	NaN	0.9661	0.9660	0.8173	0.9755
Tim Stutzle, Centre/Left Wing, Adler Mannheim...	LWOS 2020	<p>Stuzle's defensive game is a bit of a work ...	<p>Stutzle is amongst the best skaters in the ...	<p>Stutzle adds to his impressive skating with...	<p>Stutzle sometimes has a tendency to over-pa...	NaN	0.9861	0.9001	0.9756	0.9020
Jamie Drysdale, Right Defence, Erie Otters, OHL	LWOS 2020	<p>Drysdale can also get back in his own end a...	<p>Drysdale is one of the best skaters in this...	<p>Drysdale can rush the puck up the ice, maki...	<p>Drysdale could use some work on his slap sh...	NaN	0.9464	0.9828	0.9741	0.4459
Lucas Raymond, Left Wing, Frolunda HC, SHL	LWOS 2020	<p>Raymond is willing to work in all three zon...	<p>Raymond is a fantastic skater. He has an ex...	<p>Raymond combines his excellent skating abil...	<p>Raymond also has a good wrist shot and a qu...	NaN	0.8739	0.9470	0.9868	-0.0258

Figure 7: VADER Scoring Illustration

The scoring result, along with the output in 8.1, will be write into a .csv file as well, which could easily be manipulated with Excel or any other business intelligence software (Tableau, Power BI, etc.).

10.3 Output Analysis

10.3.1 Skill Adjustment

The Skill Adjustment will allow for the user to manually adjust the resultant VADER Score based on the importance of each player attribute relative to their perceived player role. Player attributes are categorized into skating, shooting, puck skill, and character in accordance to current Columbus Blue Jackets categorization techniques. Player roles are defined as Power, Sniper, Playmaker, Two-Way, and Energy for forwards and Offensive, Defensive and Two-Way for defensemen, as this is the current standard for NHL player role categorization. Player types are listed in each row, while each column corresponds to one of the evaluated skills. If all values in the table are equivalent, then no adjustments are made. If a number is set to '0', then the corresponding skill will not be incorporated into the overall evaluation for those types of players. Negative and blank entries will trigger an error. An example of different weightings is shown in Table 5 and Table 6.

Table 5: Forward Adjustments

	Shooting	Skating	Puck Skill	Character
Power	2	1	2	2
Sniper	3	2	2	1
Playmaker	2	2	3	1
Two-way	1	2	2	2
Energy	1	2	1	3

Table 6: Defenseman Adjustments

	Shooting	Skating	Puck Skill	Character
Offensive	3	3	3	1
Two-way	2	2	2	3
Defensive	1	2	2	3

All adjustments are completed in terms of ratios. Across the row are adjustments for each skill in relation to the player type (see Example 1). Down a column are adjustments for each player type in relation to the skill (see Example 2). The user must modify the numbers to create ratios that reflect their belief of the relative importance of each player type/skill combination.

Example 1: If the 'Skating' skill is twice as important for a 'Sniper' forward than it is for a 'Power' forward, then the number in the cell corresponding to the 'Sniper' + Skating' combination should be twice as large as the number entered for the 'Power' + 'Skating' combination.
Example 2: If 'Character' is twice as important for an 'Energy' forward than 'Puck Skill', then the number in the cell corresponding to the 'Energy' + 'Character' combination should be twice as large as the number entered for the 'Power' + 'Skating' combination.

Once the adjustment scores are incorporated into the Python code, the user will receive a user prompt asking which player type they would like to evaluate **all** players under, as seen in Figure 8.

```
Please enter 'F' for forward or 'D' for defense: F
Please enter one of the following forward player types - 'TwF' for two-way forward, 'GRN' for energy player, 'PwF' for power forward, 'SNP' for sniper, 'PLY' for playmaker: PwF
```

Figure 8: User Prompt for Player Type

The VADER scores for all players will then be adjusted relative to the factors put in the skill adjustment Excel file and finally calculate a final average VADER score across all 4 skill categories as seen in Figure 9.

NAME	SOURCE	CHARACTER	SKATING	PUCK SKILL	SHOOTING	RANK	CHARACTERVADER	SKATINGVADER	PUCKSKILLVADER	SHOOTINGVADER	AVERAGEVADER
Raymond, Lucas	DraftAnalyst 2020.pdf	['Raymond has tremendous upper-body strength f...	['A slightly hunched skater with excellent str...	['He seems partial to the wrister, but his cri...	['Once heÕs free, Raymond keeps his head up an...	9.0	1.9886	0.9246	1.9850	1.9084	1.701650
Ronan Seeley, Left Defence, Everett Silvertips...	LHOS 2020	<p>Seeley's strong skating allows him to maint...	<p>Seeley is an outstanding skater and this he...	<p>Seeley combines his excellent skating with ...	<p>Seeley is much more of a playmaker than a f...	NaN	1.9052	0.9761	1.9716	1.9352	1.697025
Tristen Robins, Centre/Right Wing, Saskatoon B...	LHOS 2020	<p>Robins brings his strong work ethic to the ...	<p>A bit undersized, Robins makes up for it wi...	<p>Robins has a very good arsenal of shots. Hi...	<p>Robins can also play the role of a playmake...	NaN	1.8884	0.9705	1.9552	1.9498	1.690975
Johannesson, Anton	DraftAnalyst 2020.pdf	['Johannesson is an effortless skater who move...	['Johannesson is an effortless skater who move...	['This is when Johannesson is at his best:Ñ wi...	['He also performed in two games for Sweden at...	72.0	1.9732	0.9165	1.9662	1.8856	1.685375
Will Cuylle, Left Wing, Windsor Spitfires, OHL	LHOS 2020	<p>Cuylle is also effective in the defensive e...	<p>Cuylle is an effective skater, especially f...	<p>Cuylle is a power forward in the making. He...	<p>Cuylle could stand to work on his passing a...	NaN	1.8600	0.9584	1.9688	1.9442	1.682850
...	...	...	...	...	...	...	...	...	...	...	...
Lafreniere, Alex	DraftAnalyst 2020.pdf	['A strong dual-threat winger with an advanced...	['On the puck, Lafreniere is a good north-sout...	['He displays a commanding on-ice presence by ...	['He was the top scorer for Des Moines and als...	94.0	-1.7668	0.8767	1.8508	-1.8084	-0.211925
Savoie, Carter	DraftAnalyst 2020.pdf	['HeÕs definitely an inside player, although h...	['SavoieÕs quickness is more deceptive than it...	['SavoieÕs quickness is more deceptive than it...	['The leading goal scorer in the AJHL, Savoie...	99.0	1.2972	-0.7063	-1.7282	-0.4526	-0.395875
Miller, Mitchell	DraftAnalyst 2020.pdf	['A speedy playmaker from the blue line and po...	['A speedy playmaker from the blue line and po...	['A speedy playmaker from the blue line and po...	['Whether shooting the puck or distributing it...	88.0	-0.1678	0.7506	-1.6436	-0.6178	-0.419650
Didkovsky, Ivan	DraftAnalyst 2020.pdf	['A quick and physical two-way winger who look...	['DidkovskyÕs overall skating is average in t...	['He has very quick hands and a soft touch tha...	['Not only is Didkovsky a lethal goal scorer ...	98.0	-0.0404	-0.7096	-0.2560	-1.6962	-0.675550
Kilkorka, Karel	DraftAnalyst 2020.pdf	['KilkorkaÕs game on the puck personifies poss...	['He is very agile for a defender of any size,...	['Awareness is another area where KilkorkaÕs h...	['He can run the point on the power play thank...	64.0	-1.8529	0.8542	-1.8322	-1.8322	-1.165700

Figure 9: Final output after skill adjustment.

It is understood that bias exists within these manual adjustments when determining a final ranking, but since there are biases that persist in other areas of the team which is outside of the scope of this project and therefore cannot be eliminated, these adjustments provide the user with the opportunity to create congruence between the model and these biases.

10.3.2 Proof of Concept: VADER Score vs. Written Phrases

To demonstrate the effectiveness of the VADER scores on reflecting the sentiment of the scouting reports, two players selected at different poitns in the 2020 NHL Draft will be analyzed. Alexis Lafreniere was selected 1st overall by the New York Rangers. Sam Colangelo went 36th overall to the Anaheim Ducks. Both are forwards with some similarities but Lafreniere was expected to be taken significantly higher than Colangelo, and the reports reflect this sentiment. Ideally, the VADER scores also reflect this sentiment. For the purpose of this analysis, the source for both player's scouting reports is from The Draft Analyst.

Comments on Alexis Lafreniere

Character: 'Lafreniere also is a team-first leader who has made a profound impact on Rimouski, turning them from a 59-point doormat in 2016-17 to back-to-back playoff seasons of 90+ points.'

Skating: 'Lafreniere is an excellent skater in all directions, but his powerful stride allows him to separate from opponents in open ice.'

Puck Skill: 'Biggest Strength Playmaking: Lafreniere's vision and passing skills are at the forefront of his game'

Shooting: 'From his deadly wrist shot to his superior playmaking and vision, the St. Eustache native checks every block imaginable when it comes to possessing the puck.'

The comments are mostly positive for Alexis Lafreniere, as is often the case for top-ranked prospects. His VADER scores shown in Table 7 are among the highest for all players evaluated except for his shooting ability, which suffers from the colloquialism curse that is outlined further in the Model Limitations section in 10.4. In addition, we occasionally see cases with some top prospects where they may not rank quite as high as some lesser prospects because they have normalized their excellency and scouts are generally not

as excited to talk about players whom they are already quite familiar. However, this is not deemed as a limitation as this model is designed to reflect the sentiment of scouts; not to force them to be the most excited about top prospects.

Table 7: Vader Scores for Alexis Lafreniere

	Shooting	Skating	Puck Skill	Character
Scores	0.3818	0.8804	0.7783	0.9714

Comments on Sam Colangelo

Character: 'Opposing defenders of all sizes have difficulty handling him in front of the net or during board battles, and his quickness and sharp anticipation allows him to beat them to openings with frequency.'
Skating: 'Colangelo is a nimble skater for his size but on Monday he was facing equally agile defenders who seemed to know his playbook all too well.'
Puck Skill: 'He stayed to the outside for most of the match but on several occasions was delivering crisp cross-ice or diagonal passes that led to scoring chances.'
Shooting: 'A powerful winger with dual-threat capability, Colangelo was one of the top scorers in the USHL while playing alongside a host of skilled prospects on the Chicago Steel.'

The scout was more critical for Sam Colangelo than he was for Alexis Lafreniere, which is often the case for prospects who are not ranked among the absolute best in the class and if the scout is not "hyping" the prospect as someone who should be ranked higher than commonly perceived. His VADER scores shown in Table 7 are mid-tier, which is also where he mostly ranks as a prospect. The VADER scores have proven that they have accurately reflected the sentiment of the scout's perception of Sam Colangelo. More generally, this is proof that the scores can admirably reflect the difference in sentiments between different players as expected.

Table 8: Vader Scores for Sam Colangelo

	Shooting	Skating	Puck Skill	Character
Scores	0.8271	0.3291	0.3405	0.6127

10.4 Model Limitations

In some sections of the scouting reports, a small amount of the wording does not fit into the traditional semantics. Some of the wording will reflect a negative VADER score while the scout is trying to compliment a player's skills. Please see an specific example from the scouting report of Ivan Didkovsky below:

Not only is Didkovsky a lethal goal scorer blessed with a plus-plus shot, but he also is leaned on for tough defensive-zone situations, penalty killing, and assignments that seem to be designed to throw a top opponent

off his game.

The VADER score output of this sentence is:

'compound': -0.8481, 'neg': 0.244, 'neu': 0.66, 'pos': 0.096

In this sentence, the scout does not have any negative sentiment towards Ivan Didkovsky, instead he was complimenting him. The scout intended to say that Ivan has a good shooting skills, able to handle difficult situations and surprise his opponents. However, the VADER module provided a negative score, because the word "lethal" and the phrase "throw a top opponent off his game" would be a negative sentiment in VADER dictionary. This problem can be solved by changing some of the words to be "explicitly positive":

Not only is Didkovsky a outstanding goal scorer blessed with a plus-plus shot, but he also is good at dealing defensive-zone situations, penalty killing, and assignments that will best a top opponent.

The VADER score output of this sentence is:

'compound': 0.6908, 'neg': 0.207, 'neu': 0.45, 'pos': 0.344

This would require an additional layer of text analysis, such as switching words like "lethal" to "outstanding" and "throw a top opponent off his game" to "best a top opponent". Similar phrases like "making his opponent's life difficult" will also return a negative VADER score while the scouting report is actually giving a good comment on this player's skills. Fortunately wording like this is not common in this project, most of the positive sentiment is explicit and easy to be detected by VADER scoring module.

There are also some minor instances where a phrase talks about more than one skill at one time. If the sentiment about both is the same, then there are no issues as the model currently assigns that phrase to both categories to contribute to that VADER score. However, there are some limitations that arise if the sentiments are different between the two skills. For example, 'Colangelo is a nimble skater for his size but on Monday he was facing equally agile defenders who seemed to know his playbook all too well.' In this case, Colangelo is described as a "nimble skater", which is a positive sentiment towards Colangelo's skating ability. However, the portion which described 'equally agile defenders who seemed to know his playbook all too well' is a negative sentiment towards Colangelo's hockey IQ, which falls under the 'Character' category. In these particular cases, the phrase would be sorted under both categories, but instead of praising Colangelo's skating and penalizing his character, both categories would be assign a middling-value between the two sentiments for this phrase. To solve this limitation, the phrases can be split to be more granular (i.e. split the section about Colangelo being a nimble skater and being too predictable into two separate phrases) and therefore represent each sentiment independently.

11 Implementation and Test Plan

The implementation of the project requires for the user to have the appropriate service environment as outline in Section 5.5 and to have the appropriate packages and libraries found in Figure 10.

```
import urllib.request
import requests
from bs4 import BeautifulSoup
from urllib.request import Request, urlopen
import io
import math
import pandas as pd
import numpy as np
import re
import nltk
import glob
from nltk.sentiment import SentimentAnalyzer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
```

Figure 10: Required Libraries and Packages for the Sentiment Analysis Model.

The test plan requires for the user to ensure that they are able to combine the the final adjusted (or unadjusted) VADER scores with other quantitative models or traditional statistics as designed. As the format of the final scores is in a .csv file and are now numerical metrics, you import the average the average (or any of the individual skill) VADER scores as you would any other traditional statistic. Figure 11 outlines an example comparing goals and games for the projected top 3 OHL forward from the 2020 NHL Draft (Quinton Byfield, Cole Perfetti, and Marco Rossi) to their Draft Analyst Scouting Report VADER score for shooting.

	First Name	Last Name	Games	Goals	SHOOTINGVADER
0	Quinton	Byfield	45	32	0.4764
1	Cole	Perfetti	61	37	1.0846
2	Marco	Rossi	57	39	1.6864

Figure 11: Proof of concept for combination of traditional statistics with VADER scores to create new insights.

Without reading any scouting reports, the user can immediately compare the scout's sentiment about three high end players' shooting ability with their goal totals from the OHL and generate valuable insights: 1) Was this scout was overly harsh on Byfield's shooting ability considering he had put up similar goal totals as Rossi and Perfetti in 12+ fewer games; 2) Does this scouting think Byfield's skills do not match his production, and perhaps also believes that his goal scoring from the OHL will not translate to the NHL? This example is trivial and only scratches the surface in terms of the potential of what the scores can be used to accomplish, but does test the model's ability to be combined with traditional statistics as designed.

12 Environmental, Social, and Economic Analysis

The new solution includes a new method for assessing the performance of hockey players, which would have a direct impact on player evaluation and analytic departments in the NHL. The methodology brings a new

perspective to the business as it allows transforming qualitative scouting reports into quantifiable metrics. The proposed approach would result in a change in processes and tools used by the Columbus Blue Jackets, which directly affects the organization itself. Thus, this section provides an impact assessment to evaluate and identify the environmental, social, and economic impacts of the proposed solution.

12.1 Social Impact

The stakeholder objectives and needs are important to consider while assessing the social consequences of the process. Hockey is a complex and dynamic game where the human element is undeniable. A method capturing a scout's perception of a player's performance that would objectively be combined with statistical inputs would be useful for the organization.

The CBJ would benefit from the change as it would be able to use an innovative and efficient approach to evaluate player performance. Thus, it would give them a competitive advantage in the NHL league as they would use a new and efficient approach to get player insights. The rivals would be interested in the design and want to implement it as well. However, CBJ would have the chance to be the leader of that competition as they would be the first using the design. However, one also needs to consider the impact on the players themselves. This model is not designed to eliminate any potential biases that may exist in scouting reports, and its use may perpetuate them within player evaluation circles not unlike the scouting reports they are reflecting.

It is also important to consider the role of the analyst in player evaluation departments at NHL. Currently, the analytic teams and decision-makers have to manually combine the qualitative and quantitative inputs to get a comprehensive review of player performance. The new solution eliminates this task. As the new methodology for evaluating player performance starts to be used in the CBJ, the job roles performing the analytical tasks could slightly change. The analytic team now would have a better and more efficient method regarding decision-making. Thus, the job roles can slightly change as the new process would eliminate some processes and fasten the others. Besides, the analytic team might require additional training based on their knowledge and background. They could potentially need to learn how to implement the solution and how to interpret the results.

12.2 Economic Impact

The implementation of the design has no direct economic impact for the client as it does not require hiring new labor, materials, technology, or contract of services of any kind. Thus, the design does not require any additional cost for the client to use it. However, the client may choose to purchase some online scouting reports or additional scouting services to provide a more diverse set of inputs for the model. These costs are optional and not related to the operation of the design.

12.3 Environmental Impact

Since the operation of the design does not include any manufacturing, decommissioning, and disposal processes, none of the components of the proposed solution poses a severe environmental risk. The environmental impact of the project is considered to be insignificant as no positive or negative impact is expected.

13 Next Steps and Future Work

Although the team was able to accomplish the requirements outlined at the outset of this project, there was months of research and testing which was conducted but was not able to be included in the final product. However, the following section outlines some of these pursuits which can be important next steps for the enhancement of this model.

13.1 Additional Textual Adjustments for Keyword Library

Although the team has made a significant effort on text processing, such as developing multiple working web and PDF scrapers, as well as composing a detailed keyword library dedicated to scouting report analysis, there are still improvements in the area of text processing of this project. Please see below as some of text processing areas that could be improved.

13.1.1 Stemming

Currently, since the wording used in scouting report are rather consistent, the keyword library was statically coded by the project team. This method has been proved to be successful to identify direct hits of a certain word in a sentence and further categorize the it into a predetermined player attribute. However, with the expanding of the sources, it is very possible that there are some minor change in verb tense and usage of different form of words. To solve this issue, since there are a set of root words with identical meanings, it would be more sufficient to store the "root"(or the "stem") of the word into the keyword library [10]. Refer to the examples in Figure 12 below:

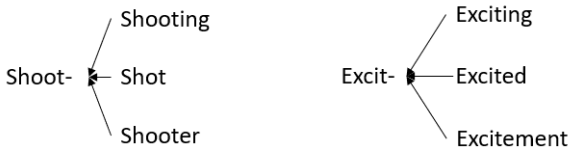

Figure 12: Example of stemming on Shoot- and Excit-

There are currently a lot of different stemming modules within the nltk library which could be used on the future steps of this project [11]. In this way, the keyword library will be come more inclusive and versatile.

13.1.2 Additional Text Screening for Implicit Sentiments

As mentioned in section 10.4, model limitations, VADER score is sometimes unable to identify the implicit sentiments within the scouting reports (please see section 10.4 for specific examples). To tackle with this problem, it is possible in the following steps to develop an additional step of text screening, to translate the sentiments in an explicit form. The process would be similar to the keyword library. Some of the commonly used negative words or phrases with positive sentiments in scouting reports such as "mean(shooting skills), lethal (shot), hard time(to opponent), etc." will be stored in that dictionary, and switch into words and phrases with explicit positive sentiments such as "skillful, outstanding, best his opponent in game, etc.". This step would require more detailed research in scouting report from more sources.

13.2 Functional Combination with BI Software

In some of the earlier meetings with client, it was mentioned that a possible future step of this project would be running the python scripts in a business intelligence software such as tableau. The team have kept that in mind and have ensured the final output of the model is a .csv file or Pandas Data frame, which are compatible with some of the popular BI software in the data analytic industry. Below are the instructions about how to run the python scripts in Tableau and Power BI.

13.2.1 Tableau

To include Python scripts in Tableau flow, a connection between Tableau and a TabPy server need to be configured. Then, Python scripts should be functional to apply supported functions to data from the Tableau flow using a pandas data frame. Tableau flows with Python scripts can be executed in Tableau Server as long as you have configured a connection to the TabPy server. It's not possible to run Python script in Tableau Online as of Mar. 26th, 2021 [12].

13.2.2 PowerBI

To run Python scripts in Power BI Desktop, Python need to be installed on the same desktop. Pandas and Matplotlib modules may need to be installed separately depending on the Python version. From there, Python scripts can be imported in the File menu by following the steps of "File - Options and settings - Options - Python scripting" and the script could be entered or imported from there [13].

13.3 Skill Adjustment Enhancements

At this current stage, the Skill Adjustment portion of the project must adjust all of the players with the same player type. For example, the user is able to evaluate all players as snipers, regardless of whether the player is a forward or defenseman or if the user wants to evaluate multiple player types at one time. Ideally, the user is able to compare each player with their own unique, desired player type to see how each player performs relative to one another based on how the team views their future player role.

13.4 Source and Date Adjustment

The Source and Date Adjustment portion of the project would allow for the user to adjust VADER scores based on their trust of the reports' sources and the dates in which the reports are written. If the user prefers one source over the other, the VADER score output from the preferred source would be weighed heavier when calculating an average VADER score across all sources. The same concept applies for the date, where more recent and relevant scouting reports may be weighed heavier than reports written months or years prior. Similar to the Skill Adjustment portion, these scores would be determined completely by the user.

14 Conclusion

The model developed in this project has provided the Columbus Blue Jackets with a method by which they can combine quantitative metrics with qualitative scouting reports in a comprehensive manner. The results are extremely promising and accurate, and show the potential future impact that Artificial Intelligence and Sentiment Analysis can have in shaping the future of player evaluation in not only Hockey, but many other sports industries around the world. The team would like to thank the Columbus Blue Jackets, and specifically their Hockey Analyst Zac Urback, for the opportunity to work on this project and that we believe this project can bring the competitive advantage Columbus ultimately needs to win the Stanley Cup!

References

[1] C. Nugent, "NHL salary data Prediction: cleaning and modeling." https://www.kaggle.com/camnugent/nhl-salary-data-prediction-cleaning-and-modeling, 2017. [Online; Accessed 16-September-2020].

[2] K. Wu, "Which League is Best?." https://hockey-graphs.com/2020/03/02/which-league-is-best/, 2020. [Online; Accessed 16-September-2020].

[3] T. Seppa, M. E. Schuckers, and M. Rovito, "Text Mining of Scouting Reports as a Novel Data Source for Improving NHL Draft Analytics." http://statsportsconsulting.com/main/wp-content/uploads/TextMiningScoutingNHLDraftAnalyticsFeb2017.pdf, 2017. [Online; Accessed 01-October-2020].

[4] K. Clark, "The Race to Make the NFL Draft an Exact Science." https://www.theringer.com/nfl/2019/4/22/18510099/nfl-draft-analytics-player-tracking, 2019. [Online; Accessed 01-October-2020].

[5] pandas development team, "Installation." https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html, 2020. [Online; Accessed 16-September-2020].

[6] J. March, "Enthought Python Minimum Hardware Requirements." https://support.enthought.com/hc/en-us/articles/204273874-Enthought-Python-Minimum-Hardware-Requirements, 2020. [Online; Accessed 16-September-2020].

[7] R. K. Toni Pano, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." https://www.mdpi.com/2504-2289/4/4/33/pdf, 2020. [Online; Accessed 22-Mar-2021].

[8] E. G. C. Hutto, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." https://ojs.aaai.org/index.php/ICWSM/article/view/14550, 2019. [Online; Accessed 22-Mar-2021].

[9] N. Swarnkar, "VADER Sentiment Analysis in Algorithmic Trading." https://blog.quantinsti.com/vader-sentiment/, 2020. [Online; Accessed 12-Jan-2021].

[10] . C. U. Press, "Stemming and lemmatization." https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html, 2009. [Online; Accessed 22-Mar-2021].

[11] N. Project, "nltk.stem package." https://www.nltk.org/api/nltk.stem.html, 2021. [Online; Accessed 22-Mar-2021].

[12] T. S. LLC., "Use Python scripts in your flow." https://help.tableau.com/current/prep/en-us/prep_scripts_TabPy.htm, 2021. [Online; Accessed 22-Mar-2021].

[13] M. P. BI, "Run Python scripts in Power BI Desktop." https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-python-scripts, 2020. [Online; Accessed 22-Mar-2021].

[14] thehockeyfanatic.com, "HOCKEY ANALYTICS." http://www.thehockeyfanatic.com/hockey-facts/hockey-analytics/, 2019. [Online; Accessed 23-October-2020].

Appendix I - Keyword Library Developed by CBJ Capstone Team

keywords["Character"] = ["improvement", "adjust", "hard work","dangerous","playmaker","smart","effort", "willingness", "fight", "battle", "offensive game", "defensive game","high-end skills","elite","high compete level", "selflessness", "playmaking", "superstar", "offense", "defense", "creativity", "decision making","defender", "two-way forward", "threat", "dominant", "timing", "vision"]

keywords["Skating"] = ["acceleration", "maneuvers", "stride", "balance", "gravity","core strength","slow", "fast", "lower body strength","speed","change","agility", "edgework", "first step", "top speed", " powerful stride", "foot speed", "long stride", "agile", "fluid", "mobility", "top-end speed"]

keywords["Puck Skill"] = ["transition","project","steal","fight","battle","support","tenacity","stickhandling"]

keywords["Shooting Skill"] = ["heavy","wrist","accuracy","release","shoot","point shoot","score","scorer","goal", "hand-eye coordination","shoot","quick hands","penalty killer","above average"]