University of Wolverhampton

Research Methods 7LN006

Thesis Proposal

Pregame Prognosis: A study in sentiment analysis of National Hockey League pregame interviews

Travis Browning

Update *13 December 2021*

Due to the availability of pregame interview data publicly available in text format, this experiment is currently offline. A similar experiment can be explored using player tweets as an alternative to pregame interview data at the following location: https://github.com/tsbrowning/NHLSentimentRegression

Abstract

Sentiment analysis is a widely distributed interest in the field of computational linguistics. Already widely utilized in many commercial applications such as social media monitoring and product analysis. Previous work including sentiment analysis within sports has already demonstrated that a speaker's use of language, collected outside of the actual gameplay when any other metric is collected, can be interpreted to amplify the results of traditional metric based prediction systems on an individual level.

Ice hockey statistics are less widely available than statistics kept for other sports. Goal events are rare. Media coverage teams for the National Hockey Association (NHL) are in the midst of adding new analytic statistics into the language of the everyday broadcast. However due to the fast and fluid nature of the game, metrics such as successful passing percentage as utilized in National Basketball Association statistics, or a rate of duels won used to define the performance of defenders in football, are often more difficult to define in hockey. For this reason, the amount of data available to metric based prediction models for hockey is smaller than it is for other sports such as baseball or basketball.

Metric based predictions of sport are utilized widely in the commercial betting industry. Despite the small dataset available, computational predictions of success on a team level of NHL games has been explored utilizing game summaries combined with statistical metrics. Although these predictions are available for predictions of team performance, it is not readily available for individual hockey players.

This study will aim to propose a model for analyzing the mood of a hockey player in a pregame interview by reviewing keywords within the players interview responses for an associated polarity score. The model will attempt to predict the players in game performance based not only on past performance statistics, but also based on the mood as defined by the answer set of the question and answer pairs in the interview. The scores of these predictions will be compared to a predictive model that utilizes past performance statistics on their own to assess how useful Natural Language Processing (NHL) can be in predictions of the statistical outputs of individual players during a game. With this aim in mind, our research questions are:

1) Can classification models utilize pregame interview text of NHL players to predict some level of variance in players' in game player performance at a game level?
2) Can text be combined with past performance metrics to produce better predictions?

This proposal is presented in three sections. In section 1 background literature will be explored, NLP processes will be explained, and a brief introduction into hockey statistics as well as statistical predictions will be given. Section 2 will discuss the process of gathering and processing data, as well as the details of data analysis. In section 3 the potential for the application of this research is discussed, as well as suggestions for further research in this area.

1 Introduction and Related Work

The intersection of sport prediction and NLP is not a novel field of research. Sentiment analysis of pregame reports have been demonstrated to be useful augmentations to traditional prediction systems attempting to estimate the outcome of NHL games (***Weissbock 2014***). Additionally, in prediction models of the performance of individual players, transcripts of pregame interviews (***Oved, 2020***) as well as individual tweets (***Xu 2015***) have been found to be helpful opportunities for sentiment analysis. These analyses as well have been shown to positively influence the predictive scores of models aiming to predict an individual player's statistical output during a game. Although NLP has been used to predict the performance of basketball players in an individual game, it does not appear to have been explored for hockey players.

1.1 NLP and Prediction of Human Behavior

Sentiment analysis (***Pang Et Al. 2002***) is well established in NLP. Widely utilized commercial applications include analyses of reviews, find patterns and trends, as well as automation of media and social media monitoring. The process involves processing numerous texts on a similar subject for training. Analyzation of these texts reveals not only the sentiment value of the text, but also how the author felt about the subject that they are writing about.

When applied to NHL coach interviews, sentiment analysis was found to have told the story of the series in the 2016 Stanley Cup Finals. Coach interviews were analyzed using NRC Word-Emotion Assocation Lexicon, which categorized keywords into categories of sentiment. Dramatic swings into sentiments such as 'joy', 'trust', and 'fear' are reflected in events of the series. Highlighted in the example below, are values of sentiments for the interviews. Following game 6 there was a clear shift in sentiments 'positive' and 'joy' from the winning coach. As well as a drop in frequency for similar emotions, the losing coach demonstrated a spike in 'anger' and 'disgust' (Figure 1).

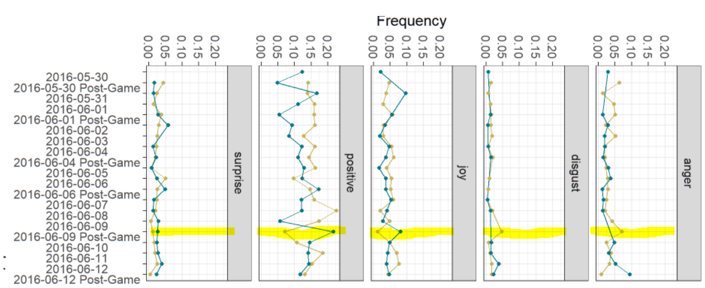

Figure 1

One research objective of Oved Et Al. was to identify an optimal model of language representation for predicting the in-game actions of basketball players based off mood analyses of pregame interviews. Three models were considered due to their performance in text classification: Convolutional Neural Networks (CNN) (***Kim 2014***), Bidirectional Long Short-Term Memory (BiLSTM) (***Liu Et Al. 2016***), and Bidirectional Encoder Representations from Transformers (BERT) (***Delvin Et Al. 2019***).

The CNN model excels at analyzing a text on a word level. BiLSTM views a text on at sentence-level. For every sentence, a mean score is calculated based off of the scores of every token within the sentence. BERT models accel at analyzing dependent statements simultaneously. The best performing text based model included primarily the architecture of BERT, as well as an attention mechanism adopted from BiLSTM. This model, ultimately was amended into a metric based Long Short-Term Memory (LTSM) system to produce the highest scoring predictions.

Through tweets, rather than interview transcripts, Xu Et Al. also attempted to predict the on-court performance of NBA players through NLP. Tweets were selected as a medium due to the insight that they provided into players thoughts and opinions. The medium did provide the potential for self-selection bias, compared to the interview format, in that through tweets players are entirely in control of the product. Similar to prediction models containing interview transcripts analyzed for sentiment, analyzed tweets were found to improve results when estimating a player's statistical output.

1.2 Prediction and Analysis in Ice Hockey

Ice hockey data is gathered on a play by play level for events which are recorded in the box score. The following is a summary of metrics used to evaluate skaters found in a box score:

Goal (G)*: A goal is scored when the puck crosses the goal line. Credit to the goal is assigned to the last attacking player to have touched the puck. At the end of the game, the winning team has scored more goals. Goals can be categorized and tallied as* ***Even strength goals (EV), Powerplay Goals (PPG)****,* ***Shorthanded Goals (SHG)****. Every goal will fit into one of these categories. As well, goals can be recorded as* ***Game Winning Goals (GWG)*** *if they are the goal that decide the outcome of the game.*

Assist (A)**:** *Up to two assists are rewarded for teammates of a goal scorer, if they have passed the puck in sequence leading to the goal.*

Point (P)*: The sum of goals and assists a player has been awarded.*

***Goal Differential (+/*-): *Goal differential is calculated for individual players in hockey, rather than being calculated for entire teams as it is in other sport. A player's +/- is defined as an integer, beginning at 0 for all players. If the player is on the ice for a goal scored their +/- will increase by 1, and if they are on the ice for a goal conceded it will decrease for the same amount.*

Penalty minutes (PIM): *For most infractions a two minute penalty will received. While a player is penalized, they are banished to the penalty box, unable to reenter the ice until the penalty timer has expired. This leaves the team of the penalized player with a numerical disadvantage. More serious infractions such as fighting receive longer major penalties.*

Shot on Goal: *Any event in which the puck is heading to the net, destined to either result in a goal or a save from the goalie. Missed shots and blocked shots do not count as shots on goal.* ***Shooting percentage (S%)*** *is calculated as a ratio of goals over shots on goal.*

Hit: *Defensive action in which one player attempts to bump an opposing player off of possession of the puck. Hits do happen on open ice, but most often the occur along the boards on the boundary of the ice rink.*

Block: *Defensive action in which a player obscures a goal bound shot (shots at a professional level are often recorded traveling over 100 mph (160 km/h)) with their stick or their body. This is typically done by sliding along the ice in between the shooter and the goal.*

Faceoff win percentage (FO%): *In hockey, play is started by a faceoff. A referee drops the puck between two opposing players on one of five locations on the ice. The players compete to win possession for their team.*

Time on Ice (TOI)*: Players do rolling substitutions in ice hockey, they can enter and leave the ice from the bench during live play. As a result, players' involvement cannot be tracked with substitution events as it is in other sports. The time in between entering and leaving the ice is a* ***shift****.* Shifts normally last between 30 and 80 seconds at a professional level, separated by 3 to 5 minute periods of rest on the bench ***[17].***

Powerplay TOI*: Time on ice during powerplay situations. During a penalty situation, the team with the numerical advantage is on the 'powerplay'.*

Shorthanded TOI*: Time on ice during penalty kill situations. During a penalty situation, the team with the numerical disadvantage is on the 'penalty kill' or playing 'shorthanded'.*

While compiling data to predict performance on a team level, ***Weissbock [8]*** included advanced statistical features such as the Fenwick Close Percentage, a ratio of time a team is in possession of the puck compared to its opposition, and a representation of luck calculated by the sum of successful and shooting percentage. The Fenwick percentage was chosen due to having the highest r^2 correlation to points in the standings, not to be confused with player points, compared to G, +/-, Hits, and others (Charron, 2013, as cited by Weissbock, 2014). Luck is defined as the results of gameplay that fall outside of the boundaries and variance in a player's performance. Luck regresses to a value of 100% over the course of time. A value over 100% is good luck, and a value below is bad luck. Team level statistics were fed into a tool called WEKA for classification tasks. WEKA contains Neural Network, Naïve Bayes, J48 and other algorithms. These algorithms were combined with traditional, advanced, and combined statistics were evaluated for effectiveness when predicting a result. Weissbock found Neural Networks combined with mixed data to be the most effective in predicting an outcome on the macro scale, and Neural Networks combined with traditional data to be most effective at predicting on a micro scale.

Rather than defining statistical boundaries, Oved Et Al. decided to use a player's statistical performance over the last three games to model statistical inputs into his classifier. This representation over three games is could vary greatly from 3/82 of a players statistics over an entire regular season, or even 3/7 of a playoff series. For this reason

2 Methodology

2.1 Research Hypothesis

A classification model can utilize player pre-game interview text to predict some variance of a player's performance on a game-by-game level. This text could then be combined with past performance metrics to produce better predictions.

2.2 Data Collection and Processing

A game of hockey is played on a rink between two teams. Under normal circumstances each team has five skaters and one goalie, controlling a rubber puck with sticks. Skaters are players who share the primary objective of scoring goals on the opposing team. The primary objective of the goalie is to block the opposing team from scoring goals. A game consists of three 20 minute periods. Teams play 82 games in an average regular season to compete for a chance to win the Stanley Cup, leading to 1230 games played in the entire league for the Regular Season. To win a round of the Stanley Cup playoffs you must beat your opponent four times in seven games. The Stanley Cup winner will have won four rounds of playoffs, or 16 games.

As a contractual obligation with their teams and the NHL, players are occasionally asked to attend interviews before or after a game. A major validity concern going into this experiment will be the quantity of available pregame interviews. Interviews are a series of open-ended question and answer pairs, between a single member or multiple members of the media. Athletes view interviews as a secondary function of their job ***[11]***, as a result interview transcripts can seem worthless in terms of lexical density.

Transcribed Pregame interviews of NHL players are available through ASAP sports from 1997 onwards. These interviews are presented in a Q/A format, already compatible with the BERT model. Within this archive there was, for most years, at least one playoff series containing a series of player interviews. For many seasons three series of interviews are available, in the form of two conference finals and one Cup final.

2.3 Model design and methods of data analysis

The goal, for each interview process will be to produce a sentiment score for the interviewee that can be interpreted as the player's mood. The AFINN[3] lexicon is used to analyze player sentiment. AFINN is a dictionary of 3,000 keywords. Keywords are assigned a polarity score between -5 and +5. Texts are first filtered through stopword lists. Next, they are summarized by adding the polarity of the keywords in text. To address bias of statement length, the sum of these keywords is divided by the number of statement length. Pictured below is the end of an output of an interview with Mike Modano, before game 3 of the 1999 Stanley Cup finals, passed through AFINN for sentiment analysis in Figure 1. In Figure 2 the values of individual content words within questions and answers are displayed

```
Q/A pair 9
Q) During the Playoffs Satan and Hasek were publicly criticized because they took themselves out of lineups and they could
n't help their team some fans and media didn't like that. Do you worry if you take yourself out of the lineup there will be
a stigma on you saying he couldn't go?
A) Well, you know, that feeling is between me and my teammates, I think; whether the fans like it or not, I don't really car
e. It is my decision and the players will understand that 100%. If I am able to go and contribute and play the way I can, as
long as I am skating well and my legs are fine, then I think everything else will kind of fall in place and the chances may
come, but, you know, what the consequences are of not playing doesn't really bother me.
The question has a mood score of: -0.0196078431372549
The answer has a mood score of: 0.0989010989010989

---------------------

Q/A pair 10
Q) Mike, did you feel any discomfort when you were out there this morning and, if so, where did it affect you the most?
A) Well, actually just -- I stretched a lot of ligaments in my wrist and my fingers at the same time while hurting my wrist
so there is a lot of discomfort in the whole area. But like I said, I could do a lot more than I thought I could do this mor
ning, but we will see tonight with the doctor and see what the magic of the needles can do.
The question has a mood score of: -0.08695652173913043
The answer has a mood score of: -0.028169014084507043

---------------------

Average mood score of questions: -0.02421305530741525
Average mood score of answers: 0.08491169556073383
Average mood score of interview: 0.03034932012665929
```

Figure 2

```
Q/A pair 9
Q:
[('During', 0.0), ('Playoffs', 0.0), ('Satan', 0.0), ('Hasek', 0.0), ('publicly', 0.0), ('criticized', -2.0), ('took', 0.
0), ('lineups', 0.0), ('could', 0.0), ("n't", 0.0), ('help', 2.0), ('team', 0.0), ('fans', 0.0), ('media', 0.0), ("n't",
0.0), ('like', 2.0), ('.', 0.0), ('Do', 0.0), ('worry', -3.0), ('take', 0.0), ('lineup', 0.0), ('stigma', 0.0), ('sayin
g', 0.0), ('could', 0.0), ("n't", 0.0), ('go', 0.0), ('?', 0.0)]
A:
[('Well', 0.0), (',', 0.0), ('know', 0.0), (',', 0.0), ('feeling', 1.0), ('teammates', 0.0), (',', 0.0), ('I', 0.0), ('th
ink', 0.0), (';', 0.0), ('whether', 0.0), ('fans', 0.0), ('like', 2.0), (',', 0.0), ('I', 0.0), ("n't", 0.0), ('really',
0.0), ('care', 2.0), ('.', 0.0), ('It', 0.0), ('decision', 0.0), ('players', 0.0), ('understand', 0.0), ('100', 0.0),
('%', 0.0), ('.', 0.0), ('If', 0.0), ('I', 0.0), ('able', 0.0), ('go', 0.0), ('contribute', 0.0), ('play', 0.0), ('way',
0.0), ('I', 0.0), (',', 0.0), ('long', 0.0), ('I', 0.0), ('skating', 0.0), ('well', 0.0), ('legs', 0.0), ('fine', 2.0),
(',', 0.0), ('I', 0.0), ('think', 0.0), ('everything', 0.0), ('else', 0.0), ('kind', 2.0), ('fall', 0.0), ('place', 0.0),
('chances', 2.0), ('may', 0.0), ('come', 0.0), (',', 0.0), (',', 0.0), ('know', 0.0), (',', 0.0), ('consequences', 0.0),
('playing', 0.0), ("n't", 0.0), ('really', 0.0), ('bother', -2.0), ('.', 0.0)]
The question has a mood score of: -0.0196078431372549
The answer has a mood score of: 0.0989010989010989
```

Figure 3

Utilizing the BERT model, a series of Question/Answer pairs will be fed into the prediction model. Oved found BERT the most effective method at addressing the bias that the player only being in control of the answer would provide. Question/Answer pairs are represented as a vector that can be averaged to produce a numerical output for every pair. These vectors are averaged over the course of an interview transcript to generate a mood value for both participants to produce annotated values on word embeddings similar to those produced above by AFINN.

Question-Answer pair tokens are fed into BERT. The tag [CLS]will be used to mark the beginning of a pair. [SEP] will be used to mark the end of a statement in the pair. BERT will output an *h* score, denoting mood, for every segment of the pair. These *h* scores will be averaged to produce another *h* score for the entire pair. For every pair in the interview, an *h* score is inputted into the attention mechanism layer to produce an *h* score for the entire interview. Here, our two tested models will differ. In the BERT-A-TM model, this interview *h* score will be concatenated with the players metric output from the past three games and fed into a Deep Neural Network linear classifier to produce a prediction on the occurrence of an event. In the BERT-A-T model we will feed the interview *h* score into the classifier independently.

Ultimately our goal is to recognize patterns in a player's game by game statistical patterns, correlated to the players mood as defined by BERT. This understanding will lead to higher performing classifiers.

2.4 Results Validation

We expect to validate the findings of Oved Et Al. We expect a text based analysis, combined with a metric analysis, when fed into a classifier, will produce the strongest prediction scores for in game actions.

A potential for bias exists, as a small group of media professionals are the only interviewers for our examples. The goal of these interviewers is to generate attention for their publication, and one of the skills nurtured by interviewers is bringing about an emotional reaction from interviewees with their questions. The objective of the interviewees is to promote themselves and their teams as a brand, before answering the interviewers questions. Team provided media training, as well as access to other interviews throughout the league, can lead to interview answers feeling like a product of route memorization of scripted cliché, rather than a genuine conversation. In official NHL interviews, players have been recorded reflecting on the inauthentic and route nature of interview responses ***[10].***

Factors outside of the player's mood will influence the statistical metrics. A player's skill level and athleticism will both play as large an influence on their metrics as their cognitive function over the course of a season. These factors tend to remain constant for players over the course of a season, outside of events such as injury. On a game level, factors such as the ability of varying opposition, familiarity of ice conditions and the presence crowd will have an influence on a player's output and can lead to the creation of a sampling bias.

As the models in our experiment are taken from Oved Et Al., our models will be expected to perform similarly with interviews and metrics of hockey players, as it has previously with basketball players. To validate, 20% of our interviews will be used in tandem with corresponding performance metrics as a test set. A 5-fold cross validation procedure for each metric label is implemented, within each fold randomly sampling from a training set containing 64% of the interviews and a development set containing 16% of interviews. If our model performs similarly, the ratio of positive and negative examples in each subset is expected to be identical to the ratio in the data set.

3 Expected Contributions and Further Research

This experiment seeks to reproduce and validate a subsection of outcomes of Oved 2020 for a new subject matter. Findings of this experiment can be useful in team level personnel decisions, as well as in the sport betting industry.

Large quantities of transcribed interviews with NHL coaching staff are available. It is worth exploring the effect that a coach's mood has on player performance on a game level. This study could be expanded to include goalie statistics, as goalies are often interviewed. Any study of hockey culture in North America is of course incomplete without the inclusion of data from the hockey hotbed of Quebec. It is common for Francophone players to be interviewed in French in multiple Canadian markets. To address this hole in data, first the transcripts of French language interviews would need to be obtained and then a parallel French language mood dictionary such as CamemBERTwould need to be utilized to replace BERT. As well, these models are worth extending into different team sports to test effectiveness.