

An Investigation into Biased Language in the American Sport Industry



Eoghan Ó Gallchóir (16339936)
School of Computer Science
National University of Ireland, Galway

Supervisors

Dr. Peter Paul Buitelaar, Dr. Mihael Arcan

In partial fulfillment of the requirements for the degree of

MSc in Artificial Intelligence

August 26, 2021

DECLARATION

I, Eoghan Ó Gallchóir, do hereby declare that this thesis entitled *An Investigation into Biased Language in the American Sport Industry*, is a bonafide record of research work done by me for the award of MSc in Artificial Intelligence from National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: Eoghan Ó Gallchóir

Abstract

To do after results have been determined

Keywords: Machine Learning, Natural Language Processing, Harvard General Inquirer, Sentiment Analysis, Text Classification

Contents

1	Introduction	7
1.1	Background	7
1.2	Context	9
1.3	Bias	10
1.3.1	Implicit Bias	10
1.3.2	Bias in Media	10
1.4	Thesis Structure	11
2	Literature Review	13
2.1	Sentiment Analysis	13
2.1.1	Basic Classification	13
2.1.2	Using an Inquirer Dictionary	14
2.2	Text Classification with Machine Learning Algorithms	15
2.3	Dataset Imbalance	16
2.4	Existing Work on Sports Media	16
3	Methodology	19
3.1	Data Gathering	19
3.2	Using the Harvard General Inquirer to produce data	21
3.3	Classification	21

3.3.1	Pre-processing	22
3.3.2	Data Imbalance	22
3.3.3	Algorithms	22
3.3.4	Evaluation	23
3.4	Sentiment Analysis	24
3.4.1	Obtaining all GI Tags	24
3.4.2	Tagging the dataset	25
4	Data	26
4.1	Classification Data	26
4.1.1	All Sentence Dataset	27
4.1.2	1-Sentence Dataset	29
4.1.3	2-Sentence Dataset	29
4.1.4	5-Sentence Dataset	30
4.2	Sentiment Analysis Data	30
4.2.1	The GI dictionary	30
4.2.2	GI Tagged Dataset	32
5	Experimental Settings	34
5.1	Classification Settings	34
5.1.1	Data Imbalance	35
5.1.2	Algorithms	35
5.1.3	Evaluation	36
5.2	Sentiment Analysis Settings	36
6	Results	38
6.1	Classification	38
6.1.1	1-Sentence Results	38
6.1.2	2-Sentence Results	39

CONTENTS

6.1.3	5-Sentence Results	40
6.1.4	All Sentence Results	41
6.1.5	Results Interpretation	42
6.2	Sentiment Analysis	44
6.2.1	Experiment Results	44
6.2.2	Interpretation	45
7	Conclusion	46
7.1	Thesis Evaluation	46
7.1.1	Contributions	46
7.1.2	Weaknesses	46
7.2	Future Work	47
	References	51
A	Code	52

List of Figures

1.1	Example of a players strengths	8
1.2	Example of a players weaknesses	8
1.3	Comparison of Racial Representation Among Enrolled Vs. Disciplined Females	9
3.1	NFL.com's 2021 draft table	20
4.1	Random sample of Master dataset rows	28
4.2	Random sample of 1-Sentence dataset rows	29
4.3	Random sample of 2-Sentence dataset rows	30
4.4	Random sample of 5-Sentence dataset rows	31
4.5	Random 10 tag sample of GI tag comparison dataset	32
4.6	Example of a sentence with multiple GI tags	33

List of Tables

6.1	1-Sentence Classification Accuracy	39
6.2	1-Sentence Classification F1-Scores	39
6.3	2-Sentence Classification Accuracy	39
6.4	2-Sentence Classification F1-Scores	40
6.5	5-Sentence Classification Accuracy	40
6.6	5-Sentence Classification F1-Scores	41
6.7	All-Sentence Classification Accuracy	41
6.8	All-Sentence Classification F1-Scores	41

Chapter 1

Introduction

The work carried out in this project involves classifying sentiment towards white and non-white American football players. This is done through machine learning models.

1.1 Background

When players first enter any of the 4 major American sports (baseball, basketball, football and hockey) they are drafted from colleges (or overseas) rather than signed or brought up from youth squads as commonly seen in European sports like soccer or rugby. The National Football League (NFL) draft consists of 7 rounds, with each team having 1 pick per round. With 32 teams in the league, that is a total of 224 selections to be made. Teams choose players in order based on performance, with worse performing teams receiving a higher pick in the draft. This is seen as a way to increase the competitiveness in the league, theoretically the worse the team is, the higher selection they have so they will get a better player. Also, the team with the 1st pick in the first round (no.1 overall), also receive the 1st pick in the second round (no.33 overall), and so on and on. It is

1.1 Background

however extremely difficult to evaluate college-level players, for example 45% of quarterbacks taken in the first round of the NFL draft were no longer on a team 5 years after they were drafted ¹. Currently there are 350 Division 1 universities in America ², this status essentially describes the standard their sports programmes are at, Division 1 being the highest. This is where most NFL players come from. Due to this size, it is unrealistic to expect that every NFL coach can assess every player of interest to him to draft onto his team, so teams rely on scouting reports, from scouts they either hire or trust, to help inform them in deciding on who to draft. These scouting reports describes a players physical and cognitive attributes in the form of a short paragraph, also describing his strengths and weaknesses. An example of these strength and weaknesses can be seen in Figure 1.1 and Figure 1.2. Most prospective players also attend a draft combine, which essentially is a standardized workout and assessment used to quantify their abilities. American sports media also produces similar player analysis ³, as well as game previews and post-game reports.

Strengths:

- Well-developed route running
- Advanced technique
- Tracks the ball well
- Late hands
- Body control
- Tracks the ball well
- Dangerous on 50-50 passes
- Adept at making catches over defensive backs
- Can make some highlight-reel catches
- Good size, build
- Gritty, competitive syle
- Run-after-the-catch skills
- Nose for the ned zone

Weaknesses:

- Lacks mismatch speed
- Lacks twitch
- Could struggle to separate from NFL defensive backs
- Too many dropped passes
- Needs to improve his hands

Figure 1.2: Example of a players weaknesses

Figure 1.1: Example of a players strengths

¹<https://football.pitcherlist.com/pessimists-guide-to-the-nfl-draft/>

²<https://www.ncaa.org/about/division-i-schools>

³<https://nfldraft.theringer.com/mock-draft>

1.2 Context

Studies done on the American sports industry and its media has been time and time again shown that racism exists in sports media. One study by Harrison (2000) showed that racial bias still exists in NFL commentary, with the problem persisting across several different television network coverages. The NFL is in a unique position for analysis as there is a huge discrepancy in both players of colour and people of colour not only in team coaching and executive positions ¹, but in reporting roles in mainstream media ². In the 2020/2021 season of the NFL, players of colour accounted for 69.4% of all players, while there were only 4 head coaches of colour in the league in that same season (12.5% of all head coaches). Similarly, it was found that 85% of sports editors were white, and 80.3% and 82.1% of columnists and reporters were white, respectively.

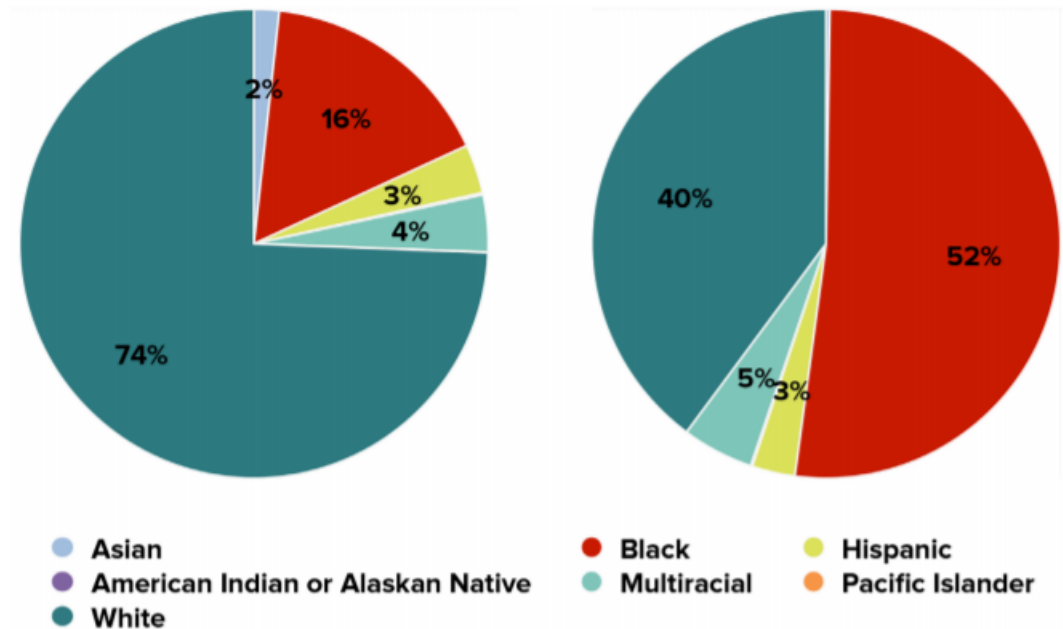


Figure 1.3: Comparison of Racial Representation Among Enrolled Vs. Disciplined Females

¹<https://bit.ly/3gxzsEu>

²<https://bit.ly/3mvzsbI>

1.3 Bias

1.3.1 Implicit Bias

Implicit bias has been described by Staats et al. (2017) as attitudes that subconsciously affect our actions and decisions. These attitudes can be either positive or negative, and are activated unintentionally. This is the opposite of explicit (or conscious) bias, where attitudes are intentionally affecting our actions. Given how implicit bias is present in every one of us, it can take many different forms in many different areas of life. For example in education, a study¹ examined the rates of disciplinary actions in the Ohio education system faced not only by students of different races but also compared those rates on a per gender basis. It was found that:

- black students received a disproportionate amount of discipline compared to their percentage of enrollment.
- black female students experienced the highest level of overrepresentation.

This clearly shows how prevalent and damaging bias can be. The study's results can be seen in Figure 1.3.

1.3.2 Bias in Media

Media has long played a role in shaping how some people view minority groups and race at large (Greenberg et al., 2002). Signorielli (2009) found the existence of both gender and racial stereotyping on television. Viewing women and minority through this lens provides those who watch with conventional ideological views about them (Signorielli, 2009). This study used code to gather network television

¹<http://kirwaninstitute.osu.edu/implicit-bias-training/resources/race-matters.pdf>

data over a period of nine years. The same coding schemes, and validity testing was used year-on-year. This study is relevant not only because of its subject matter but also due to its aforementioned methodology. The paper does however fall short in several categories.

- No popular paid channels like HBO were tracked during this time.
- The coders who implemented the research were only given a few weeks of training. Could lead to inconsistencies too as a different group oversaw each year.

With television being some peoples only interaction with ethnic groups outside their own, some authors have examined the effect that this stereotyping has on those who consume these programmes. As long as mainstream media continues to reproduce racial and ethnic stereotypes, these false characteristics will continue to exist, and viewers will see it as how people truly are (Castañeda, 2018).

1.4 Thesis Structure

This thesis is divided into seven chapters.

- Chapter 1 provides background information and puts the investigation into context.
- Chapter 2 reviews current work in this domain, as well as literature review of papers on media bias, classification algorithms, and sentiment analysis techniques.
- Chapter 3 is a detailed look at the methods used to both analyse player sentiment and classify player ethnicity.
- Chapter 4 describes the data created for and during the experiments.

- Chapter 5 details the experimental settings.
- Chapter 6 provides experiment results and interpretation.
- Chapter 7 gives a final review of the results in the context of the aim of this paper.

Chapter 2

Literature Review

This chapter covers existing literature that relate to the core topics of this thesis.

2.1 Sentiment Analysis

2.1.1 Basic Classification

Sentiment analysis is the process in which the attitude or emotion expressed by a text is classified. This is done through natural language processing techniques. The rise of social media applications, Twitter in particular, has only increased interest in sentiment analysis. As NFL scouting reports are subjective, opinion-based documents, it is important to understand current classification methods of textual reviews. Experiments conducted by Mouthami et al. (2013) showed that sentiment analysis of textual (in their case movie reviews) reviews could be classified at a document level. The research conducted does prove that large scale sentiment analysis of documents can produce concrete results due to its use of the Cornell movie-review corpora. However, documents were not classified beyond positive and negative sentiment. While this binary classification is somewhat useful, no specific information can be obtained from this, which is why Inquirer

dictionaries shall be utilised in experiments conducted in later chapters of this paper.

2.1.2 Using an Inquirer Dictionary

Inquirer dictionaries are essentially tools to map words to dictionary-supplied categories. The General Inquirer (GI) dictionary is the most widely used in semantic analysis research, containing over 182 categories in all (Stone et al., 1966), far surpassing simple positive or negative sentiment that we have seen before. Currently, this dictionary is a combination of the Harvard IV-4, the Lasswell dictionaries, as well as contributions based on the social cognition work of Semin and Fiedler. A study by Pollach (2006) uses this GI dictionary to perform computer-assisted semantic analysis on product reviews on consumer opinion web sites. Outside of the positive and negative tags, 11 categories were found relevant to their research. Pollach’s findings corroborated earlier experiments pertaining to word frequencies, causing them to conclude that these product reviews follow “implicit genre rules regarding content, format and language”. This analysis shows that using an Inquirer dictionary yields results beyond simple positive or negative sentiment. It is also highly relevant because of the specificity of the corpus it analyses. The scope of the corpus analysed in the Experiments chapter of this paper will have a similar corpus scope, and it is promising that multiple Inquirer categories were still discovered. A similar or higher number of categories should be found in the subject of NFL scouting reports.

2.2 Text Classification with Machine Learning Algorithms

Text classification is one of the core topics of this paper, and shall be the subject of one of the experiments performed in the experiments chapter. Miao et al. (2018) showed that the Support Vector Machine (SVM) algorithm was the most accurate for multi-label classification, albeit only slightly more accurate than the much quicker Naive Bayes (NB) algorithm. Due to this the SVM algorithm is only seen as appropriate for use with smaller datasets. The experiment is relevant to this paper due to the fact that both the SVM algorithm and Naive Bayesian algorithm was implemented by the scikit-learn package and produced desirable results. More information on scikit-learn can be found in section 3.3. Another one of its strengths is the use of precision and recall in addition to the F1-score. This gives context to the F1-score. A weakness of this paper is that the smaller classes have better results than the big classes, which would lead to a reduction in expected accuracy under practical circumstances.

Dharmadhikari et al. (2011) also posited that SVM is one of the most effective text classification algorithms due to its ability to manage large spaces of features. The two papers also agree its unwieldy computational demands. The papers findings also takes into account NLP techniques and how it can be used to improve classification results. A major weakness of this paper is the lack of experiments showing its findings, focusing entirely on existing literature. Another algorithm that can be used for text classification is Stochastic Gradient Descent (SGD). It performs well with sparse and high dimensional data (Prasetijo et al., 2017). Madhfar and Al-Hagery (2019) found that on large, multi-label datasets, both SGD and Logistic Regression had the highest F1-score in comparison to SVM

and NB approaches.

2.3 Dataset Imbalance

Data imbalance occurs in many real world domains. Given that 70% of NFL players are non-white, this applies to the problem domain of this paper. Song et al. (2013) shows that imbalanced data affects the performance of normal classification, with balanced datasets achieving a higher average F1-score. It was found that the difference in classification performance between balanced and imbalanced data grew as the degree of imbalance distribution grew. One weakness of the experiments by Song et al. (2013) is that it did not involve any text classification, something that is a core component of the experiments proposed in later in this paper. Oversampling the minority class, especially in datasets with a large discrepancy provided good results (Batista et al., 2004). Furthermore, Batista et al. (2004) also found that random over-sampling was less computationally expensive than other methods to balance datasets.

2.4 Existing Work on Sports Media

American sports, with American Football and the National Basketball League (NBA) specifically, are in a unique position, with the majority of their players being non-white but other positions, be it coaching positions, executive positions or media positions being of a white majority ¹. This has caused the two sports to be scrutinized more than most for bias in their respective media. Studies have been conducted into the language used by television announcers or commentators. They give play-by-play description of what is happening at any given time during the game. More often than not they are paired with an analyst, whose role is to

¹<https://bit.ly/3gxzsEu>

2.4 Existing Work on Sports Media

provide expert analysis and background information about a player or team. An article by (Rada and Wulfemeyer, 2005) analysed comments made by both the play-by-play commentator and the commentator tasked with analysis in American Football and Basketball collegiate games. To sample the American Football it took only a quarter of each game over a period of weeks as their data sampling method. For the basketball it used a Division 1 Men’s Championship basketball tournament, obtaining over 55 hours of basketball coverage. They hypothesized that black players would receive:

- a higher portion of negative comments than white players.
- more comments on their physical attributes than white players.
- more negative comments regarding both their intellect on and off the field.
- both more negative statements about their character and would receive more negative personal interest stories than their white counterparts.

In each case the hypothesis was supported by their experiments. This article’s strength lies in how thorough it is despite its small sample size. The study would have however benefitted from a larger dataset, perhaps over several years. In this paper it has been established the effect this implicit bias can have on viewers, discussion which is missing from Rada and Wulfemeyer (2005)’s article.

There are instances of overcoming the problem of a small dataset. It was shown by (Merullo et al., 2019) that large scale analysis of American sports media was not only possible, but produced coherent results. 1,445 American football (both collegiate and NFL) games were automatically annotated with mentions of players and linked with metadata. To reduce susceptibility to noise ARK TweetNLP was used as it is more robust than conventional part-of-speech tagging methods. One of the outputs of this study was their *FOOTBALL* dataset, a large scale sports

2.4 Existing Work on Sports Media

commentary corpus annotated with the race of the player, which is useful for any future work in bias of American football media. This paper is included here due to its sound technical foundations, and how it is possible to produce definitive results from processing the language associated with sport. It also only furthers the need to analyse scouting in this manner.

Chapter 3

Methodology

In this chapter, the method of sourcing the data is described. The classification and sentiment analysis experiments are also explained here ahead of their use in Chapter 5.

3.1 Data Gathering

The data needed to perform the experiments proposed was not available from any resource online. Thus, web scraping techniques had to be implemented. It was decided upon that player data should be obtained from the official NFL.com website ¹. Beautiful Soup ² (BS) was used to obtain all player information from the most recent (2021) draft. Here is where the player in-depth profile URL link was also captured, which will also be parsed to find information to make up the scouting report column of the datasets. As the web page in question uses Javascript to load its content, Chromium and Selenium was used to connect to the web page and allow the content to load before obtaining the information needed to create the datasets. The page's Javascript loaded every player name,

¹<https://www.nfl.com/draft/tracker/picks?year=2021>

²<https://beautiful-soup-4.readthedocs.io/en/latest/>

3.1 Data Gathering

player position and player profile URL link under the same CSS tag, allowing for functions to iteratively grab the data. Part of the web page can be seen in Figure 3.1. The player scouting report was obtained in a similar manner, using the aforementioned URL link with BS to get the player strength and weaknesses. The final column of the dataset, player race, was manually added, as there is no record of it on the web pages parsed. Each row in the dataset consists of many sentences pertaining to one player. To test the the different machine learning algorithms on how much information they need, the master dataset was split into 1-sentence, 2-sentence, and 5-sentence datasets. For more information on all the datasets created, see Chapter 4.

1		JACKSONVILLE JAGUARS	Trevor Lawrence	QB	Clemson
2		N.Y. JETS JETS	Zach Wilson	QB	Brigham Young
3		SAN FRANCISCO 49ERS	Trey Lance	QB	North Dakota State
4		ATLANTA FALCONS	Kyle Pitts	TE	Florida
5		CINCINNATI BENGALS	Ja'Marr Chase	WR	LSU

Figure 3.1: NFL.com's 2021 draft table

3.2 Using the Harvard General Inquirer to produce data

The Harvard General Inquirer dictionary is available to download for use from the harvard.edu website ¹. Using this file, a Python dictionary of key-value pairs was created, the key being a stemmed word from the GI dictionary file and the values being each sentiment associated with it. With this dictionary, new datasets were created by applying the dictionary to the datasets created in the section above. This created a GI tagged version of the 1-sentence, 2-sentence, 5-sentence, and master dataset. The 1 sentence GI tagged dataset is what shall be used to perform the sentiment analysis experiment described later in the chapter. The Harvard GI dictionary was also used to create an integer tagged version of the 1-sentence, 2-sentence, 5-sentence and master dataset. The dataset(s) columns were every GI tag found in the master dataset and the players ethnicity. Each row contained either a 1, indicating the gi tag was present in some sentence, or a 0, to indicate no presence of the tag. This was done in an attempt to improve the accuracy and f1-score Of the GI tagged version of the aforementioned datasets. This collection of datasets shall be henceforth referred to as the binary-GI tagged group of datasets.

3.3 Classification

This section describes what machine learning tools were used to implement the classification experiment. The experiment's aim is to use different algorithms on the different datasets created to try and to correctly classify between white and non-white instances. From this we can see what level of accuracy we can

¹http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

achieve, a higher score insinuating that language used can separate players by their ethnicity. FLOWCHART

3.3.1 Pre-processing

Text pre-processing was performed on the scouting report sentences in each dataset before this experiment was performed. Specifically:

- stopword removal with the Natural Language Toolkit (NLTK) library.
- removal of any punctuation with NLTK.
- the stemming of words also with NLTK.

After this pre-processing, each report instance was then converted to a matrix of token counts through vectorization.

3.3.2 Data Imbalance

The data imbalance in the datasets in this thesis was rectified through oversampling of the minority, in this case white player, class. To do this, the `imbalanced-learn` library (Lemaître et al., 2017) is used. Imbalanced-learn is an MIT-licensed library that uses scikit-learn that provided tools to deal with imbalanced classes during classification problems. It is an open source library that has both under-sampling and over-sampling methods. It is also fully compatible with scikit-learn features. In this experiment, an imbalanced-learn pipeline was used. Over-sampling was implemented to oversample the minority class and thus even the dataset.

3.3.3 Algorithms

To explore how many sentences are needed to produce the most accurate result, different ML algorithms were used. All algorithms were implemented with scikit-

learn (Pedregosa et al., 2011) due to its accessibility. It has a huge library of classification, regression, and clustering algorithms. The four algorithms used are:

- *Logistic Regression*: In the case of this experiment, this classifier models the probability whether a scouting sentence(s) instance is either a white player or a non-white player.
- *Naive Bayes*: This is a simple classifier that assumes strong independence between features. However in this experiment, models are created on one feature, the scouting report sentence(s).
- *Stochastic Gradient Descent*: This algorithm works well for text classification due to the sparse nature of text classification (Prasetijo et al., 2017).
- *Support Vector Machine*: Similar to SGD, SVM also works well for text classification.

Each of these algorithms were applied to all datasets created for classification. For details of the settings for each of these algorithms, see the chapter on experiment settings.

3.3.4 Evaluation

Due to the fact that the data collected was small (see Data chapter), evaluation techniques are used to combat this. 10-fold cross-validation was used over hold-out validation as k -fold cross-validation produces more accurate results (Yadav and Shukla, 2016). F1-scores are implemented as part of the accuracy evaluator. It was chosen over accuracy due to how important precision and recall are in ML problems. The average F1-score was taken from the output of the k -fold cross-validation to produce results for each algorithm.

3.4 Sentiment Analysis

The second experiment proposed in this paper is to perform sentiment analysis on sentences categorized by the Harvard GI. This includes the Harvard IV-4 dictionary and the Lasswell value dictionary. The 1-sentence dataset shall be tagged with the sentence's GI category and the player's race that the sentence pertains to. There are several steps involved in this method, described below. FLOWCHART

3.4.1 Obtaining all GI Tags

Firstly, the aforementioned GI Python dictionary of key-value pairs is used. A dataset is then parsed through and each key occurrence for both white and non-white player reports causes its value (its sentiment) to be kept track of. It is important to note that one word can have multiple GI tags. For example, *Confident* has 7 unique GI tags associated with it:

- Positive - Positive words.
- Strong - Words implying strength.
- Power - Subset of strong, indicating a concern with power, control, or authority.
- Pleasure - Words indicating the enjoyment of a feeling, including words indicating confidence, interest, and commitment.
- EMOT - Words related to emotion.
- Overst - "Overstated", Words indicating emphasis in realms of speed, frequency, causality, inclusiveness, quantity or quasi-quantity, accuracy, validity, scope, size, clarity, exceptionality, intensity, likelihood, certainty and extremity.

- WlbTot - Words in well-being, relating to the health and safety of the player.

For each of these GI tags in both ethnicities, the percentage of their occurrence was then calculated, and these figures were made into a dataset.

3.4.2 Tagging the dataset

From the dataset created above, GI tags deemed irrelevant were removed from the dataset. Using this dataset, the 1-sentence dataset, and the GI python dictionary, every row in the 1-sentence dataset was parsed through and all GI sentiments for that sentence was found and added to a new dataset, consisting of the player sentence, the GI tag, and player ethnicity. With this new data, the sentiment for each GI tag for both white and non-white players can be found, as well as the difference between them. For more information on all datasets created for sentiment analysis see Chapter 4.

Chapter 4

Data

This chapter describes all data gathered for the purpose of experimentation.

4.1 Classification Data

For the classification experiment, 4 different datasets were created. As mentioned in the Methodology chapter, each dataset has a standardized structure. Column names of:

- Player ID - Given to provide player anonymity, each player has their own unique identifier number. This becomes important in the 1-sentence, 2-sentence, and 5-sentence datasets because as the data becomes more divided, player sentences can still be grouped by their ID.
- Position - Abbreviated player position. In this data, 17 player positions arose:
 - C - Center.
 - CB - Cornerback.
 - DE - Defensive End.

- DT - Defensive Tackle.
 - EDGE - Edgerusher.
 - FB - Fullback.
 - G- Guard.
 - K - Kicker.
 - LB - Linebacker.
 - LS - Long Snapper.
 - OT - Offensive Tackle.
 - P - Punter.
 - QB - Quarterback.
 - RB - Running Back.
 - SAF - Safety.
 - TE - Tight End.
 - WR - Wide Receiver.
- Scouting Report Sentence(s) - These are sentences taken from a players strength and weaknesses list as described on NFL.com.
 - Player Race - player race is denoted as either **W** - White or **NW** - Non-White.

4.1.1 All Sentence Dataset

This is the first dataset that was obtained, and is referred to in this paper as the master dataset. All other datasets pertaining to the classification experiment is derived from it. It contains:

4.1 Classification Data

- 259 player rows in the format discussed above.
- Each player has an average of 17 sentences in their report. Player ID 1 has the most sentences at 31, player ID 225 has the least sentences associated with them at 4.
- Of the 259 players, 213 have been classified as NW and 46 have been classed as W.

ID	Position	Scouting Report Sentence(s)	Race
248	G	Showed off durability with 42 career starts. Played both tackle positions and right guard. Punch is accurate and fired with leverage. Makes quick adjustment when front moves. Works to sustain and finish his block. Effective when asked to pull, find and land his block. Tackle experience improves protection potential. Doesn't cut hands loose unnecessarily in pass pro. Core power and ankle flexion to battle bull-rushers. Average hand quickness into first contact. Not much knock-back pop. Allows opponents to punch and play off of him. Leg drive won't be able to push pros around. Pass slides are more like heavy stomps. Head fakes tilt him off-balance. Struggles to recovery and redirect rusher off track. Athletic pass rushers cause him to over-correct.	W
159	OT	Very durable, making 40 consecutive starts for Nebraska. Has starting experience at both tackle positions. Quality lateral foot quickness. Able to set out on top of rushers when he needs to. Throws combos like a boxer to maintain his feel for the rusher. Choppy, controlled footwork to mirror and slow inside counters. Athleticism to open hips and recover at top of the rush. Maintains pad level and knee bend. Plays with body control and second-level adjustments. Rolls hips through initial contact. Uses inside hands into the frame. Accelerates feet through angle blocks to create block security. Defenders successfully cross his face with slants. Needs better landmarks and angles in first attack. Base will narrow out on move blocks. Sustain would benefit from better reset of hands and post-block footwork. Below-average length diminishes opportunity to cinch up after punch. Plays with nose out past toes a little too often. Would benefit from better weight distribution to prevent falling forward. Upper body gets ahead of his feet at times on lateral movements. Rush anchor is below average.	W
164	SAF	Position and scheme versatile. Starting experience at cornerback and safety. Maintains eye balance throughout the route. Plays with good reads and positioning from the post. Instinctive and rangy, firing off the hash over the top. Smooth transitions to route match from off-man. Above-average ball skills. Possesses hands and body control to take it away. Intercepted Ohio State QB Justin Fields twice in 2020. Diligent getting head around to find the football. Field intelligence to line up defense on back-end. Very good feel for space from short zone. Rarely out of position. Top-end speed appears to be average. Delay in downfield trigger coming out of his pedal. Run pursuit might need more patience as a safety. Tackling will need more work. Inconsistent to break down and center up. Wrap-up tackle strength is very average.	NW
147	TE	Three-year starter with adequate size. Highly decorated as in-line and slot option. Appeared to play with an elevated level of confidence. Improved route work and quickness out of breaks in 2020. Takes tight angles on out-breaking routes. Stems and creates openings for quarterback on short slant. Scrambles open on off-schedule pass plays. Absorbs heavy strikes and continues on after the catch. Adequate acceleration after the catch. Can be a punishing runner with the ball in his hands. Drops head into contact and blocks with erratic technique. Block sustain will require much more work. Gradual route bends between hashes give positioning away. Balance is volatile in pattern work. Tight hips restrict intermediate opportunities. Needs more aggression at the top of the route. Excessive body-catching and bobbles. Not a very reliable option when catch is contested.	NW
24	RB	Great size with the demeanor for the game. Runs decisively and with urgency. Rarely fumbles. Able to gather and cut on short notice. Downhill wiggle to slalom around bodies. Change of direction is crisp. Makes large number of tacklers miss for being a physical back. Has tools to handle inside/outside duties. Follows and bursts to daylight off of lead blocks. Violent finisher who is ready to thump when challenged. Plus balance to keep run on track through first contact. Talent as route runner and pass catcher was on full display in 2020. Soft hands with above-average ball skills. Willing and able in pass protection. Runs with inconsistent feel for tempo. Lacks speed to threaten for the big play. Gets in a hurry, hindering his ability to find developing blocks. Runs with low knee action and average burst between tackles. Average acceleration out of his cuts. Takes on heavy contact on a consistent basis. Running style could lead to challenges with durability.	NW

Figure 4.1: Random sample of Master dataset rows

Some of this master dataset can be seen in Figure 4.1.

4.1.2 1-Sentence Dataset

This dataset is the master dataset split into 1-sentence rows. It consists of 4511 player rows. Of those rows, 3726 have been annotated as NW and 785 have been classed as W. An example of this dataset can be seen below in Figure 4.2.

ID	Position	Scouting Report Sentence(s)	Race
152	SAF	Sudden and explosive leaper at high point	NW
180	SAF	Broke right collarbone twice and dislocated the shoulder on the same side	NW
47	CB	Twitch for more plays on the football if he squeezes routes a little tighter from off-man	NW
59	WR	Has experience outside and as a big slot in LSU offense	NW
8	CB	Tunnel vision caused issues for him in coverage near goal line in 2020	NW
105	LB	Better athlete than linebacker	NW
240	DE	* Active hands in the passing lane	NW
6	WR	Eleven of 20 career touchdowns went for 50-plus yards	NW
11	QB	Needs to improve pocket mobility for clean launch points	NW
228	CB	Willing to submarine pulling linemen to set an edge	NW
246	EDGE	Rangy, with some closing speed to pursue	NW
142	G	Has starting experience at guard and tackle	W
3	QB	Film junkie with high football IQ and an NFL frame	NW
4	TE	Rare combination of size, speed, athleticism and elite ball skills	NW
246	EDGE	Might need to rush with a hand in the ground	NW

Figure 4.2: Random sample of 1-Sentence dataset rows

4.1.3 2-Sentence Dataset

This data has also been garnered from the master dataset, with each row having 2 non-repeating sentences. As some players had a number of scouting report sentences not divisible by 2, a random, previously seen sentence of theirs was randomly added to the data to complete that 2-sentence row. This was done by using modulo to get the sentences that needed supplementing. There are 2448 rows in this dataset, 2022 of which are classified as NW and 426 as W. See Figure 4.3 for a 2-Sentence data example.

4.2 Sentiment Analysis Data

ID	Position	Scouting Report Sentence(s)	Race
64	QB	* Tardy safeties will find a willing challenger over the top, * Talented downfield passer with touch and accuracy	W
251	CB	Injuries have been an ongoing concern, Plays too low when defending high/low route concepts	NW
57	WR	Brings legitimate deep-ball danger to the field, Breakaway speed as home-run hitter with a stack of long TDs to his name	NW
212	SAF	Indecisive with sticky feet as open-field tackler, Gets turned around tracking deep routes	NW
96	EDGE	5 sacks over last 19 games, Bursts into gaps with pads low and motor revved	NW
116	EDGE	Fails to generate much torque and push as power rusher, Needs to supplement basic rush attack by setting up counters	NW
243	SAF	Strikes pass catchers with force to jar pass loose, Good physicality taking on blocks in space	NW
177	LB	Adequate upper-body strength to unglue from blocks, Subtle level changes to keep blockers guessing	NW
71	CB	Good foot quickness and agility, Desired short-area burst on lateral transitions	NW
199	DT	Rarely ends up on the ground as run defender, Flashes strength to recover from early disadvantages in the trenches	NW
108	CB	" Very good feel for timing and playing receivers hands", Good physicality and play strength in run support	NW
74	CB	False steps at the top of his drop from off-man, Narrow base in space creates bumpy transitions to shadow route breaks	NW
253	DE	" Rush plan isnt well-prepared", " Doesnt get to spin counters often enough"	NW
2	QB	Offers potential as full-field reader, Arm looked substantially more lively in 2020, with highlight-reel throws on the move	NW
157	WR	Small deceleration coming out of route stem, Long, skinny legs are slower to change direction	NW

Figure 4.3: Random sample of 2-Sentence dataset rows

4.1.4 5-Sentence Dataset

Similar to the 2-sentence dataset, this data was also pulled from the master dataset. Modulus was also used here to find rows lacking the 5 sentences needed. A random sample, the size of the need of the row was taken from that players own pool of sentences. Steps were taken to ensure none of the sampled sentences were already part of the row in need. There are 1052 row in this dataset, 866 of which have been classified as NW and 186 classed as W. An example of this dataset is below.

4.2 Sentiment Analysis Data

4.2.1 The GI dictionary

The dataset which the Python GI dictionary helps create is GI tag comparison dataset. It consists of 4 columns:

4.2 Sentiment Analysis Data

ID	Position	Scouting Report Sentence(s)	Race
113	LB	Strong pro day effort with good speed, vertical and bench numbers, Inconsistent diagnosis of run play development, Can be a step slow flowing with the pace of outside zone, Gets trapped behind climbing blockers in space, Allows lead blockers to get into his play-side shoulder	NW
139	OT	Tremendously long, Has 85-inch wingspan and 35-inch arms, Punch is well-timed and has some snap on it, Keeps scrambling in recoveries and will find wins, Lateral foot quickness to make back-side/play-side reach blocks	NW
156	DE	Goes along for the ride on lateral block engagements, Unable to sit down firmly and anchor as an edge setter, Lacks foot quickness to work quickly around a block, Below-average hands to open a rush lane for himself, Way too tall on rush counters and is easily punched and stalled	NW
65	SAF	Possesses pro frame with good overall size and length, Soft hands and takes aggressive routes on the throw, Finished college career with 13 interceptions, Sees and responds to passing game much better as split safety, Attacks crossing routes like a heat-seeking missile	NW
120	RB	Very little wear and tear with 165 carries at Oklahoma, Running style is willful and belligerent, Adequate burst for lateral cuts from downhill track, Requires legitimate run support from cornerbacks to slow him, Carries first tackle try with him as he falls forward	NW
68	OT	Bursts out of stance as run blocker, Above-average sense of positioning in run game, Accelerates feet through down blocks, Reliable climbing to second level from combo block, Good block adjustments to sudden movement in space	NW
116	EDGE	Elongated first three steps gain plenty of ground into rush turn, " Stalks and closes quarterbacks with efficiency once hes in the pocket", Will punch, lift and shuffle around the block at point of attack, Plays with some force in his hands, Appears to have gained more weight since 2019, which helps his cause	NW
120	RB	Needs ball tucked tight to frame to prevent fumbles, Limited one-cut ability due to tight hips, " Loses forcefulness when feet are slowed or hes spilled outside", Runs with blurred vision as inside runner, Will unexpectedly take run off designed play track	NW
1	QB	Speed, wiggle and toughness for zone-read and called runs, Understands how to protect himself, Feet stay calm when working from pocket, Subtle pocket slides to stay pass-ready and on platform, Internal clock helps him stay on schedule	W
60	LB	Long and thin through hips and legs, Below-average base to withstand angle blocks, Will have trouble constricting run lanes with force, Needs to be quicker shadowing lateral run cuts, Glued to bigger blockers once they find his frame	W

Figure 4.4: Random sample of 5-Sentence dataset rows

- Sentiment - A GI tag which appears at least once in the master dataset. There were 156 GI tags found in.
- % of words with sentiment in white reports - The percentage of words that some GI tag is associated with in W reports.
- % of words with sentiment in non-white reports - The percentage of words that some GI tag is associated with in NW reports.
- Relative % difference - The relative difference in occurrences of a GI tag in W and NW reports

From these 156 tags, 56 were deemed as relevant, a sample of which can be seen Figure 4.5.

4.2 Sentiment Analysis Data

Sentiment	% of words with sentiment in white reports	% of words with sentiment in non-white reports	relative % difference
Ovrst	3.79	4.04	-6.39
Vice	2.14	3.54	-49.30
WlbPsys	0.23	0.24	-4.26
SV	0.79	0.79	0.00
AffTot	1.55	0.92	51.01
Virtue	8.57	9.25	-7.63
Exprsv	1.42	1.71	-18.53
Goal	1.61	1.21	28.37
Quan	2.83	2.59	8.86
TrnLoss	1.09	1.10	-0.91

Figure 4.5: Random 10 tag sample of GI tag comparison dataset

4.2.2 GI Tagged Dataset

The GI tagged dataset is an amalgamation of the 1-Sentence dataset and the dataset filtered above. Given that one word can have many GI tags associated with it, this dataset contains 29137 rows. Each row has:

- Sentence - The player report sentence in question.
- Tag - A GI tag associated with a word in that sentence.
- Race - The players ethnicity.

See Figure 4.5 for an example of a sentence that has multiple GI tags associated with it. This was the data used to perform sentiment analysis on a tag-by-tag basis.

4.2 Sentiment Analysis Data

Sentence	Tag	Race
Elite size, athleticism and play traits	Power	W
Elite size, athleticism and play traits	PowAuPt	W
Elite size, athleticism and play traits	PowTot	W
Elite size, athleticism and play traits	Active	W
Elite size, athleticism and play traits	Solve	W
Elite size, athleticism and play traits	IAV	W
Elite size, athleticism and play traits	Nonadlt	W
Elite size, athleticism and play traits	IndAdj	W
Elite size, athleticism and play traits	Know	W

Figure 4.6: Example of a sentence with multiple GI tags

Chapter 5

Experimental Settings

Described below is the complete experimental settings of the classification and sentiment analysis experiments. All experiments were completed using Python and the settings have been written to reflect this.

5.1 Classification Settings

The classification experiment is the most setting intensive of the 2 experiments. All 12 sentence datasets were used here. Following from the model described in Chapter 3, the following settings were used to produce results. Before any experiments were run, text preprocessing was performed on the player report of the datasets that were read in:

- With the NLTK library, stopwords data was imported and set to 'english'.
- All punctuation was removed using the `string` library.
- Stemming was then done using NLTK's `PorterStemmer` function.

Each row of data was processed like this, the model extracting the processed player report and its label (W, NW). Scikit-learn's *train_test_split* function was

then used to split the data into training and test sets. This function had 2 settings, the train/test size split, which was set to 75/25, and `shuffle`, a boolean on whether or not to shuffle the data before splitting. This was assigned to `True`. An instance of scikit-learn's `CountVectorizer` was then used. The function was given a "strict" setting in dealing with character it does not recognise. It was also set to also only vectorize unigrams. The vectorizer was then trained on the training set features and applied to.

5.1.1 Data Imbalance

Data imbalance was corrected using imbalanced-learn's *Pipeline* classes. It requires a list of inputs to transform and sample the data. In this experiments case, the class was given a sampling strategy to resample all classes but the majority class, and the ML algorithm to be used. Each ML algorithms settings are described in the following subsection.

5.1.2 Algorithms

Scikit-learn is used to implement all 4 ML algorithms used in this experiment. The library provides the ability to alter many aspects of each algorithm, so each algorithm's parameter were kept as default unless otherwise stated.

- LR - This algorithm uses L2 regularization with a Limited-memory BGFS (Broyden-Fletcher-Goldfarb-Shanno) algorithm as its solver. The maximum amount of iterations for the solver to converge has been set to 100.
- NB - The scikit-learn class `MultinomialNB` was used to implement the Naive Bayesian algorithm. This is because it is able to use the vectorization for text classification. The additive Laplace smoothing parameter was set to 1.

- SVM - The `SVC` (Support Vector Classification) class was used to imitate an SVM algorithm. A Radial Basis Function (RBF) was used as the kernel as it produced the best results. L2 regularization was also used.
- SGD - The SGD algorithm is implemented using the `SGDClassifier` class. The algorithm's loss function was the hinge loss function. Again, L2 was the penalty used. The alpha value used was set to 0.0001. This was used to compute the learning rate as the learning rate parameter was set to "optimal". The training data for SGD was also shuffled after every epoch.

For each algorithm, the parameter `random_state` was set to a constant integer number, for result reproducibility.

5.1.3 Evaluation

10-fold cross validation was implemented using the `RepeatedStratifiedFold` class. 10 splits were specified for the number of folds. 3 cross-validator repeats were also specified. Score evaluation was then done through the `cross_val_score` function. This used the pipeline built, and accepted a data list of the sentences to classify and the data list of player race as its target variable. 2 scoring metrics were recorded, algorithm `accuracy` and algorithm `f1-score`. These scoring functions were passed into the `cross_validate` evaluator. The result was the mean of the 10 scores, 1 from each of the folds performed. Experiment results and analysis is discussed in Chapter 6.

5.2 Sentiment Analysis Settings

The sentiment analysis experiment uses a Python implementation of the Harvard GI dictionary, its creation discussed here. The `pysentiment2`¹ library was used

¹<https://pypi.org/project/pysentiment2/>

5.2 Sentiment Analysis Settings

to tokenize each player report column in each row. This library was used over other tokenization methods as it contains the Harvard GI dictionary as its library, an instance of which can be called for the purpose of tokenization. This was used to tokenize every player report sentence in the 1-sentence dataset, an example of which can be seen in Figure 4.4.

This dataset, along with the manually filtered GI tags list, is the input to the model that gets the average sentiment for each GI tag for all W and NW sentences associated with that tag. Each row in the dataset is read and the sentence's polarity calculated using `pysentiment2`'s tokenization and sentiment score functions. These functions were used in their default settings. The model then adds this polarity score to either the W polarity score or the NW polarity score. Once the dataset has been iterated through, it outputs the average W and NW polarity for the GI tag in question, as well as the difference between the two polarities. The model produces this result for every GI tag it is given.

Chapter 6

Results

Results first, using figures and tables, with little commentary and no interpretation. Then analysis and interpretation. The results chapter presents the findings from both experiments described in the Methodology and Experiments chapters.

6.1 Classification

This section presents results in relation to the classification experiment. Overall, 12 different datasets were passed through the model, totaling 96 outputted results between the accuracy and f1-score. Both higher accuracy and F1-Score indicates better performance of the algorithm. From our results we can interpret how much information different algorithms need to produce the best results possible.

6.1.1 1-Sentence Results

In the context of the 1-Sentence datasets, Table 6.1 shows that the SVM algorithm produced the best classification accuracy results across all 3 different dataset types. LR, NB, and SGD all had similar scores across the 3 dataset types. Overall, the raw sentences version of the 1-Sentence dataset gave the highest accuracy,

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.71	0.645	0.565
Naive Bayes	0.681	0.628	0.579
Support Vector Machine	0.806	0.723	0.682
Stochastic Gradient Descent	0.706	0.636	0.562

Table 6.1: 1-Sentence Classification Accuracy

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.313	0.27	0.301
Naive Bayes	0.339	0.286	0.291
Support Vector Machine	0.181	0.239	0.251
Stochastic Gradient Descent	0.294	0.261	0.284

Table 6.2: 1-Sentence Classification F1-Scores

with the GI Tagged 1-Sentence giving the 2nd highest results.

The NB algorithm gave the single highest F1-score, and was consistently high overall. Again, SGD and LR shared similar results. SVM produced the worst F1-Scores in Table 6.2 for each of the 3 dataset types.

6.1.2 2-Sentence Results

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.754	0.691	0.603
Naive Bayes	0.746	0.675	0.603
Support Vector Machine	0.83	0.77	0.738
Stochastic Gradient Descent	0.759	0.689	0.616

Table 6.3: 2-Sentence Classification Accuracy

From Table 6.3, again the SVM algorithm has the best accuracy results, in this case for the 2-Sentence datasets. Each of the other 3 algorithms scored

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.333	0.257	0.311
Naive Bayes	0.396	0.293	0.3
Support Vector Machine	0.19	0.218	0.253
Stochastic Gradient Descent	0.308	0.246	0.294

Table 6.4: 2-Sentence Classification F1-Scores

similarly in each of the 3 dataset types. The raw sentence dataset type produced the highest accuracy in all 4 machine learning algorithms. The Binary-GI tagged sentences had the worst accuracy performance for each algorithm.

Table 6.4 indicates that the NB algorithm produced the highest F1-Scores in the Raw Sentences and GI Tagged Sentences datasets. LR and SGD had similar results here too. SVM produced the lowest F1-Scores in each dataset type. Like the 1-Sentence dataset, the raw sentences dataset type performed the best.

6.1.3 5-Sentence Results

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.816	0.732	0.723
Naive Bayes	0.833	0.739	0.686
Support Vector Machine	0.842	0.742	0.771
Stochastic Gradient Descent	0.816	0.738	0.735

Table 6.5: 5-Sentence Classification Accuracy

From Table 6.5, it can be seen that all 4 algorithms performed well on the raw sentences dataset, each scoring over 80% accuracy. SVM and NB had the highest results overall, though all 4 algorithms scored similarly. Raw Sentences again had the highest accuracy, with the GI Tagged and Binary-GI sentences reporting similar accuracy results for each of the algorithms used.

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.459	0.285	0.32
Naive Bayes	0.55	0.312	0.328
Support Vector Machine	0.253	0.242	0.249
Stochastic Gradient Descent	0.399	0.263	0.298

Table 6.6: 5-Sentence Classification F1-Scores

In Table 6.6, NB has the best F1-score in all of the 5-Sentence dataset types, followed by LR. In the 5-sentence experiment, a clear separation can be seen between the 4 machine learning algorithms. The Binary-GI dataset outperformed the GI Tagged dataset here. SVM performed the worst of the 4 algorithms in each dataset type.

6.1.4 All Sentence Results

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.824	0.774	0.723
Naive Bayes	0.829	0.829	0.686
Support Vector Machine	0.828	0.781	0.771
Stochastic Gradient Descent	0.828	0.787	0.735

Table 6.7: All-Sentence Classification Accuracy

Algorithm	Raw Sentences	GI Tagged Sentences	Binary-GI Sentences
Logistic Regression	0.431	0.266	0.304
Naive Bayes	0.457	0.28	0.262
Support Vector Machine	0.047	0.405	0.169
Stochastic Gradient Descent	0.391	0.261	0.261

Table 6.8: All-Sentence Classification F1-Scores

In Table 6.7, it can be seen that all 4 algorithms performed very similarly on the all-sentence ¹, raw sentences dataset type. NB had the highest accuracy for the GI Tagged dataset and the lowest accuracy on the Binary-GI dataset type. LR and SGD had similar accuracy across all dataset types. The raw sentence dataset type lent itself to providing the highest accuracy. Binary-GI produced slightly worse accuracy results than GI Tagged.

Table 6.8 shows that the NB algorithm produced the best F1-Score for Raw Sentences, SVM produced the highest F1-Score for the GI Tagged sentences, and LR the highest for Binary-GI. SVM produced poor results for both Raw Sentences and Binary-GI. As a whole, Raw Sentences provided the highest F1-Scores, while GI Tagged and Binary-GI gave similar scores.

6.1.5 Results Interpretation

From the results of each of the 4 (1-Sentence, 2-Sentence, 5-Sentence, All-Sentence) sentence level experiments, several interpretations can be made, firstly with regards to accuracy:

- Of the 3 dataset types (Raw Sentences, GI Tagged, Binary-GI) it is clear that the Raw Sentences version of each of the 4 sentence levels produced the highest accuracy, regardless of what algorithm was used. The Raw Sentences 4 gave a mean accuracy of 78.8%, with GI Tagged and Binary-GI only achieving accuracy levels of 72.3% and 65.7% respectively.
- The SVM algorithm stood out as having the highest accuracy overall, with the algorithm averaging an accuracy of 77.2% across all dataset types and sentence levels. The other 3 algorithms varied in their accuracy standings

¹The all-sentence datasets is the master dataset, the master dataset GI Tagged and Binary-GI'd

across the 4 sentence levels, but had extremely similar average accuracy with, SGD averaging 70.8%, LR averaging 70.7%, and NB averaging 70.6%.

- Another result we can extrapolate from the tables shown is the accuracy on a sentence level to sentence level basis across the 3 dataset types. The 1-Sentence, 2-Sentence, 5-Sentence, All-Sentence dataset had an average accuracy of 66%, 70.6%, 74.4%, and 78.3% respectively. This shows that the more data in the scouting report available to each algorithm, the better it can classify the player.

The F1-Score results shall also be discussed. One generic finding from the F1-Score is that LR, NB, and SGD had a higher recall metric score than precision score. The SVM F1-Scores always produced a balance of the two metrics. Several more interpretations can be made from the F1-Score results:

- Like the accuracy results, the Raw Sentence version of each of the 4 sentence levels gave the highest F1-Score in comparison to GI tagged and Binary-GI, achieving an average F1-Score of 0.34, compared to GI tagged's 0.27 and Binary-GI's 0.28. Binary-GI's F1-Score being similar to GI Tagged was unexpected given that GI Tagged far outperforms Binary-GI in terms of accuracy.
- The NB algorithm had the highest average F1-Score across all dataset types and sentence levels, with a score of 0.34,. LR was second best in this respect, averaging 0.32. SGD averaged 0.30. SVM performed by far the worst, averaging only 0.22. This is due to the fact that the other 3 algorithms recall was much higher than SVM's at every sentence level with every dataset type.
- Finally, the 5-Sentence type had the highest average F1-Score of the 4 sentence levels, averaging 0.33. The 1-Sentence, 2-Sentence and All-Sentence

were bunched together in terms of score, achieving an average of 0.28, 0.28, and 0.29 respectively. This means that the NB algorithm on the 5-Sentence Raw Sentence dataset had the best individual score with 0.55.

Overall, this experiment showed that the Raw Sentence type was the best for text classification regardless of sentence level. This is because of how well the word vectorizer works with sentence data. It also showed that the Binary-GI was the worst dataset type for classification. This is due to the sparse nature of the dataset, as only a 5 to 10 columns would be set to 1 on each row. The number of features was also a cause of low performance, as the dataset had 156 features, 1 feature for every GI Tag found in all the data. Something interesting from these results was the difference in result order of the algorithms between accuracy and F1-score. Accuracy results went: SVM with the highest average, then SGD, then LR, then NB. F1-Scores was the reverse of this: NB producing the highest, LR, SGD, and SVM following in that order. One of the reasons is because of NB, LR, and SGD's high recall metrics. High precision was the reason SVM was able to achieve such high accuracy.

From the classification accuracy produced in this paper, it can be concluded that **W** and **NW** players can be classified to a high degree of accuracy using machine learning methods of text classification.

6.2 Sentiment Analysis

6.2.1 Experiment Results

figs and tables

6.2.2 Interpretation

Remember to show examples

Chapter 7

Conclusion

Here you must zoom back out to evaluate the thesis. Mention limitations and weaknesses as well as contributions.

7.1 Thesis Evaluation

What i wanted to do, and what i learned

7.1.1 Contributions

datasets created, 1,2,5 and 1-GI tagged. showed ML strengths.

7.1.2 Weaknesses

what is a significant difference for sentiment analysis score? dataset small as oversampling needed. No DeepL methods for classification. choosing GI tags when filtering

7.2 Future Work

DeepL methods for classification. Classify on GI tags of sentences bigger dataset

References

- Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007735. URL <https://doi.org/10.1145/1007730.1007735>. 16
- M. Castañeda. The power of (mis)representation: Why racial and ethnic stereotypes in the media matter. In *The Power of (Mis)Representation: Why Racial and Ethnic Stereotypes in the Media Matter*, 2018. 11
- S. Dharmadhikari, M. Ingle, and P. Kulkarni. Empirical studies on machine learning based text classification algorithms. *Advanced Computing: An International Journal*, 2:161–169, 2011. 15
- B. Greenberg, Dana E. Mastro, and Jeffrey E. Brand. Minorities and the mass media: Television into the 21st century. In *Minorities and the mass media: Television into the 21st century*, 2002. 10
- C. Keith Harrison. Chapter seven: Brains, brawn, and pigskin balls: Racism and athletic manifestation in society. *Counterpoints*, 107:103–121, 2000. ISSN 10581634. URL <http://www.jstor.org/stable/42975921>. 9
- Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-

REFERENCES

- learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL <http://jmlr.org/papers/v18/16-365.html>. 22
- Mokhtar Ali Hasan Madhfar and Mohammed Abdullah Hassan Al-Hagery. Arabic text classification: A comparative approach using a big dataset. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5, 2019. doi: 10.1109/ICCISci.2019.8716479. 15
- Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O’Connor, and Mohit Iyyer. Investigating sports commentator bias within a large corpus of american football broadcasts. *CoRR*, abs/1909.03343, 2019. URL <http://arxiv.org/abs/1909.03343>. 17
- Fang Miao, Pu Zhang, Libiao Jin, and Hongda Wu. Chinese news text classification based on machine learning algorithm. In *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 02, pages 48–51, 2018. doi: 10.1109/IHMSC.2018.10117. 15
- K. Mouthami, K. Nirmala Devi, and V. Murali Bhaskaran. Sentiment analysis and classification based on textual reviews. In *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, pages 271–276, 2013. doi: 10.1109/ICICES.2013.6508366. 13
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 23

REFERENCES

- I. Pollach. Electronic word of mouth: A genre analysis of product reviews on consumer opinion web sites. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 51c–51c, 2006. doi: 10.1109/HICSS.2006.146. 14
- Agung B. Prasetyo, R. Rizal Isnanto, Dania Eridani, Yosua Alvin Adi Soetrisno, M. Arfan, and Aghus Sofwan. Hoax detection system on indonesian news sites based on text classification using svm and sg. In *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 45–49, 2017. doi: 10.1109/ICITACEE.2017.8257673. 15, 23
- James A. Rada and K. Tim Wulfemeyer. Color coded: Racial descriptors in television coverage of intercollegiate sports. *Journal of Broadcasting & Electronic Media*, 49(1):65–85, 2005. doi: 10.1207/s15506878jobem4901_5. URL https://doi.org/10.1207/s15506878jobem4901_5. 17
- Nancy Signorielli. Race and sex in prime time: A look at occupations and occupational prestige. *Mass Communication and Society*, 12(3):332–352, 2009. doi: 10.1080/15205430802478693. URL <https://doi.org/10.1080/15205430802478693>. 10
- Yale Song, Louis-Philippe Morency, and Randall Davis. Distribution-sensitive learning for imbalanced datasets. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. doi: 10.1109/FG.2013.6553715. 16
- Cheryl Staats, Kelly Capasto, Lena Tenney, and Sarah Mamo. Implicit bias review. Technical report, The Ohio State University, 2017. 10

REFERENCES

- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966. 14
- Sanjay Yadav and Sanyam Shukla. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83, 2016. doi: 10.1109/IACC.2016.25. 23

Appendix A

Code

See GitHub repository for thesis code and data (repo currently private).