

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 17, 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: _____

Abstract

To do after results have been determined

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

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	21
3.3.2	Data Imbalance	22
3.3.3	Algorithms	22
3.3.4	Evaluation	23
3.4	Sentiment Analysis	23
3.4.1	Obtaining all GI Tags	24
3.4.2	Tagging the dataset	24
4	Data	26
4.1	Classification Data	26
4.2	Sentiment Analysis Data	26
5	Experiments	27
5.1	Classification Settings	27
5.1.1	Data Imbalance	27
5.1.2	Algorithms	27
5.1.3	Evaluation	27
5.2	Sentiment Analysis Settings	28
6	Results	29
6.1	Classification	29
6.1.1	Experiment Results	29
6.1.2	Interpretation	29
6.2	Sentiment Analysis	29
6.2.1	Experiment Results	29
6.2.2	Interpretation	29
7	Conclusion	30
7.1	Thesis Evaluation	30

CONTENTS

7.1.1	Contributions	30
7.1.2	Weaknesses	30
7.2	Future Work	30
	References	34
	A Code	35

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 (Wright, 2016)	9
3.1	NFL.com's 2021 draft table	20

List of Tables

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 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 (Miklius, 2019). Currently there are 350 Division 1 universities in America (NCAA, 2019), 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. 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 (Kelly, 2021), 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

1.2 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 Viklund (2009) 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 (Lapchick, 2020), but in reporting roles in mainstream media (Lapchick, 2018). 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 by Lapchick (2018) that 85% of sports editors were white, and 80.3% and 82.1% of columnists and reporters were white, respectively.

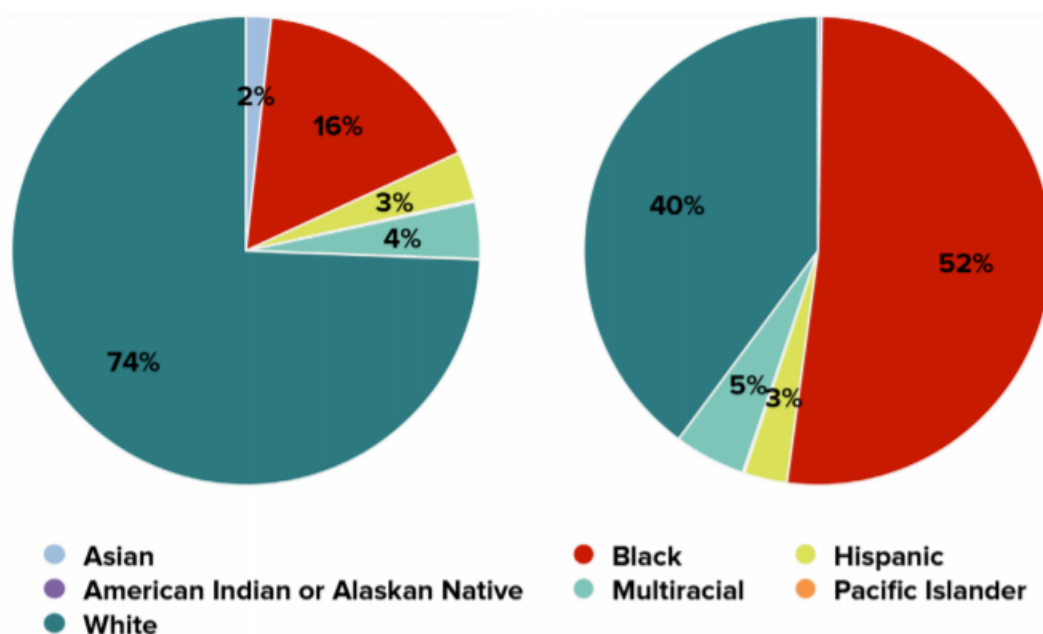


Figure 1.3: Comparison of Racial Representation Among Enrolled Vs. Disciplined Females (Wright, 2016)

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 study by Wright (2016) 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.

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 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 models created 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 the 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 (Lapchick, 2020). 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 provide expert analysis and background information about a

2.4 Existing Work on Sports Media

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 commentary corpus annotated with the race of the player, which is useful for any

2.4 Existing Work on Sports Media

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. 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 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.

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 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.

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

- 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.

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.
- Ovrst - "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

To do before monday

4.1 Classification Data

How many sentences, players, datasets, etc. *** Lots of R tables to show data

4.2 Sentiment Analysis Data

Chapter 5

Experiments

Your goal is to give a complete description of your experiments, sufficient for another researcher to read your document and reproduce your results.

5.1 Classification Settings

can also talk about what and how preprocessing was done

5.1.1 Data Imbalance

imb pipeline settings

5.1.2 Algorithms

75/25 split, vectorization,

5.1.3 Evaluation

eval settings - f1 score, cross Val Score settings, stratified fold settings

5.2 Sentiment Analysis Settings

Chapter 6

Results

Results first, using figures and tables, with little commentary and no interpretation. Then analysis and interpretation.

6.1 Classification

6.1.1 Experiment Results

6.1.2 Interpretation

Remember to show examples

6.2 Sentiment Analysis

6.2.1 Experiment Results

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

7.1.2 Weaknesses

what is a significant difference for sentiment analysis score?

7.2 Future Work

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
- Mari Castañeda. The power of (mis)representation: Why racial and ethnic stereotypes in the media matter. Technical report, University of Massachusetts Amherst, 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, 2002. 10
- Danny Kelly. Nfl draft guide. <https://nfldraft.theringer.com/mock-draft>, 2021. 8
- Richard E. Lapchick. *The 2018 Associated Press Sports Editors Racial and Gender Report Card*. TIDES, 2018. 9
- Richard E. Lapchick. *The 2020 Racial and Gender Report Card*. TIDES, 2020. 9, 16

REFERENCES

- Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-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
- Michael Miklius. Nfl draft prep: How successful are first round picks? <https://football.pitcherlist.com/pessimists-guide-to-the-nfl-draft/>, 2019. 8
- 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
- NCAA. Division i schools. <https://www.ncaa.org/about/division-i-schools>, 2019.

REFERENCES

- 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. 22
- 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. Prasetijo, 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 sgd. 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*

REFERENCES

- 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
- 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
- Pat Viklund. *Brains versus Brawn: An Analysis of Stereotyping and Racial Bias in National Football League Broadcasts*. PhD thesis, Boston College, 2009. 9
- Robin A. Wright. Race matters... and so does gender. <http://kirwaninstitute.osu.edu/implicit-bias-training/resources/race-matters.pdf>, 2016. vi, 9, 10
- 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