Abstract

This paper examines Myers-Briggs personality classification through social media posts and Natural Language Processing Techniques.

Final Project

An Evaluation of personality types through text analysis

Bates, Neal

nabates@syr.edu

Haskins, Sandra

[smhaskin@syr.edu](mailto:smhaskin@syr.edu)

Khoury, Maryse

[mehjoury@syr.edu](mailto:mehjoury@syr.edu)

Ortiz, Jonathan

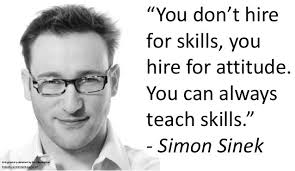
Jortiz13@syr.edu

Sunday 4:00 – 5:30pm

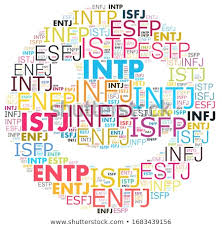
**Introduction:**

 Big Data revolutionized the way most businesses function and gain competitive advantages. Walmart and Amazon are key examples of growing a data knowledge base and taking that advantage (McAfee, 2011). Depending on the area of expertise Big Data might have a different meaning, some business functions are only interested in collecting and warehousing the data. But to be successful like Walmart or Amazon, Big Data must be interpreted in a way to find a competitive advantage. These companies do this second step by utilizing Artificial Intelligence or Machine Learning (Fox, 2020).

 Another arena that has been revolutionized in a post internet era is the Public Relations (PR) job market. A key contributor to that change is the rise of user-generated-content, on social media platforms (Solis, 2009). Social media has empowered the individuals into becoming influencers and is forcing PR and marketing professionals to not only recognize the impacts of these influencers but also to incorporate them as part of their public relational and marketing strategies (Solis, 2009). In the cases of Walmart and Amazon, they too utilize social networking to influence marketing and their public images. Both utilize social media sites of Facebook and Twitter, to help shape their brand and seek out new markets.

 However, the paper is not to discuss how Walmart or Amazon manage their public relations. Rather, this paper will discuss how a Human Relation (HR) department can find personalities that mesh well with the current workforce and culture through social media postings. “Through his research and experience, Dr. Kustis [Industrial Psychologist and Management Consultant, Dr. Gary Kustis] found that bad hires lead to 80% of employee turnover and cost 2-3 times the salary level of the position that they were hired for.” (Madigan, 2016)

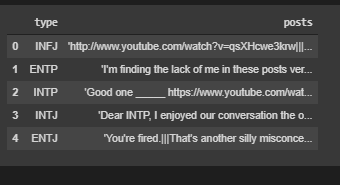
A powerful process to determine if the social media post fits a particular personality consistently is by utilizing Natural Language Processing (NLP). This could be beneficial to a company if they are looking for a particular type of personality for leading a group. Or possibly, if they are looking for a member that can help bridge several groups together into a project portfolio. Often a resume, does not capture these attributes, but a cover letter or social media postings could.

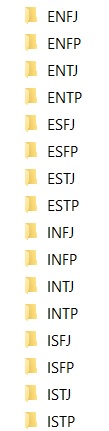
In order to demonstrate that potential, this paper will analyze public postings on various social media sites through Natural Language Processing analysis and determine the personality type for Myers-Briggs. However, not all Machine Learning algorithms are created equal and some are better suited to a task or dataset. Hence, by the end of this paper the reader should understand different NLP techniques, like Multinomial Naïve Bayes, Bernoulli, Naïve Bayes and Support Vector Machine Classifiers (SVC), Multi-layer Perceptron’s (MLP) and how different forms of vectorization can impact a result for such classification techniques. This paper will also cover discovery techniques to see there are other methods that can reinforce these categorized findings.

**Analysis and Models:**

**About the Data**

The dataset comes from a Kaggle about determining the personality type of an individual given a corpus of their social media posts. There are 50 posts per person and their personality type is labeled. The data was provided as a zip. There were 8,675 rows/individuals included in the dataset. Figure 1 below is presenting the first 5 rows of the data. There are two columns, first being the personality type label, according to Myers-Briggs and the second contains the individual’s corpus of their 50 posts on a social media site.





*Fig 1: Training CSV as a Data Frame*

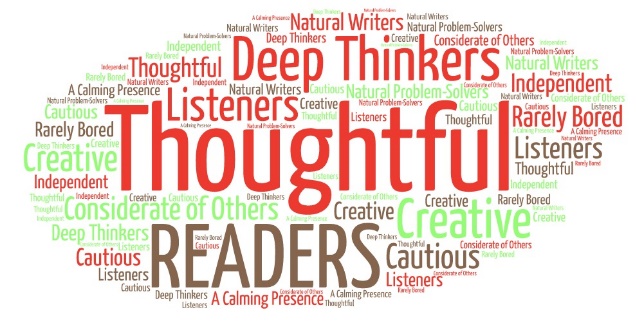
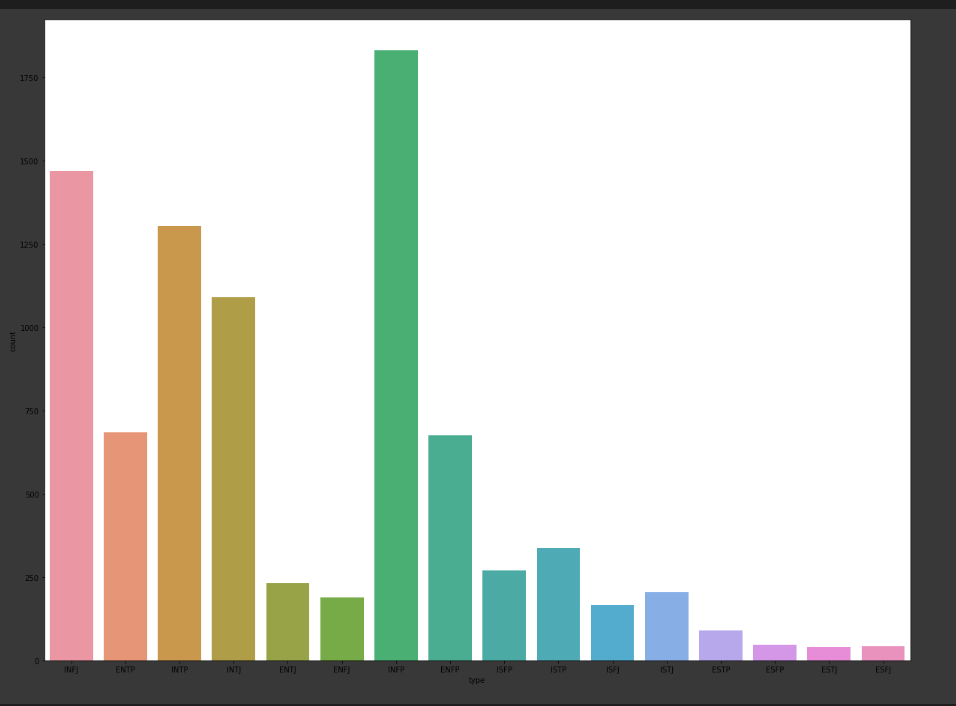
Figure 1 is displaying the dataset. What is unique is that there are a lot of multimedia links in the discussion board. These multimedia files will need to be removed as the research group did not have enough time to check each media type or come up with a systematic way to see what the multimedia was about. Thus, these multimedia files are treated as a (unigram) and in the more mature steps of the classifications were removed.

Figure 2 is presenting the frequency distribution of personality types for the dataset. The most obvious takeaway is that there is not an equal distribution between sentiment types. The most common type of personality is “INFP”. This label is not surprising given Figure 2 shows that it makes up ~19% of the total labels size. This topic will also be covered in greater detail in the model and results section.

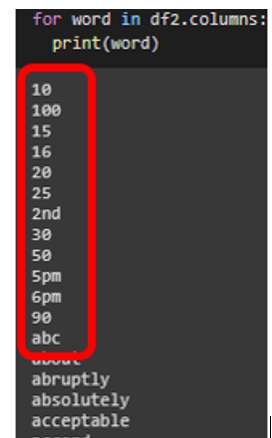


*Fig 2: Frequency Distribution of Personality Types in the Dataset*

Additionally, as Figure 2 presents, that is the distribution of the dataset values. Myers-Briggs did not provide any distributions on how much of a population is a given personality type. Hence, for the analysis it is assumed that the sample provided matches the larger population as a whole. This assumption will enable the trained algorithms from this dataset to be applied the population as a whole and those media posts.

**Transforming the Data**

A key component of any computerized form of text analysis is that the words must be transformed into a numeric values. To do that, the first step is to take each corpus of social media posts and read them into a single matrix called a “count vectorizer”. What this does is take every word and their counts for the number of times they appear in a post, across all posts. Another additional transformation is that a lower was applied to all the data. The reason why this is an important step for Natural Language Processing (NLP) is because a computer will interpret a word with a capital first letter and the same word with no capital letters as different words. Unless they are normalized by being all lower or some other form. In some areas of NLP, it is not suggested to make all things lower but for this case, lowering will help in getting the most accurate view.



*Figure 3: Example Stop-Words*

On top of lowering all the words in the dataset, additional filters are often applied to datasets, like removing stop-words. Figure 3 displays, words from the posts and the red circle identifies some of the stops-words. The reason why a data scientist would want to remove stop-words is because they generally do not assist in determining a classification. Things such as numbers, addresses, or name should not be considered a valid word for this form of text classification as these attributes are defined by a particular personality type. Normally those words will be utilized as describing a fact of where or when the topic occurred for a post. This setting aspect of the post is something relatively neutral when it compares to personality itself. Consequently, to assist the algorithms in predicting the various personality types, removing stop-words were part of the more mature prediction methods.

As mentioned in introduction, this paper will show various algorithms and count vectorizer models. Thus, instead of filtering the stop-words solely, this paper will compare results of a normal frequency model of Count Vectorizer to a TF-IDF Vectorizer. A TF-IDF Vectorizer takes the words that are found most common across all documents receive a lower weighted score. Meaning, hopefully it will do a superior job for identifying different personality types by reducing the scores of the remaining common words.

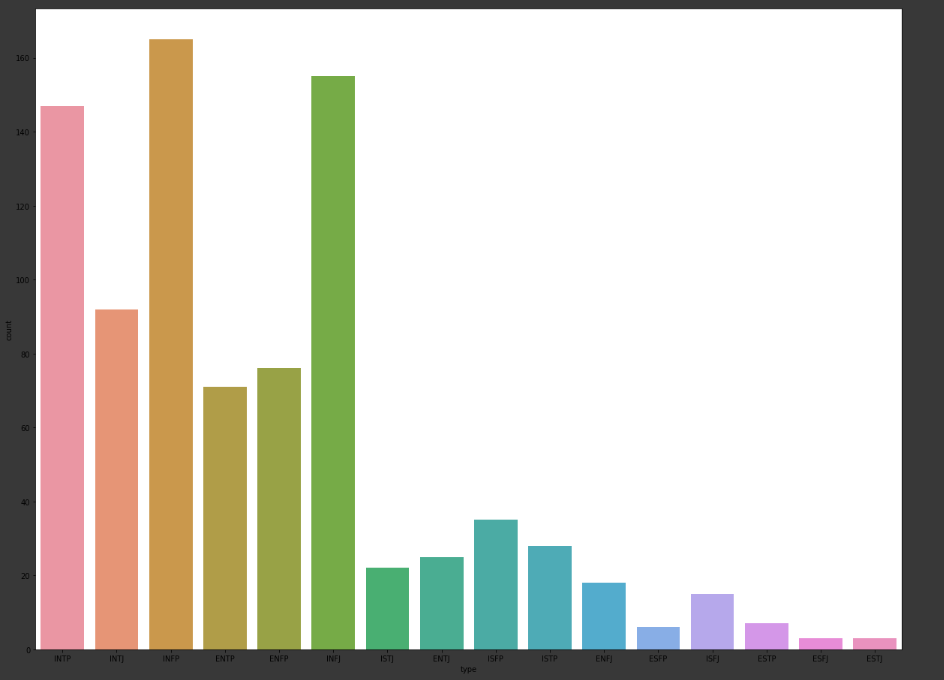
Another common NLP process worth mentioning is a stemmer or lemmatizer. A stemmer and lemmatizer are both taking a word and trying to apply a normalized form. For instance, a simplistic stemmer will remove the plural of the word, or other conjugations. And as mentioned in the above any character differences between words, the computer would treat those words as different. A lemmatizer is also showing a normalized form, but transform the word, back to the root form. The issue is that in both cases, it is not advised as by stemming or lemmatizing, the emotional significance would be impacted. Say, words ‘happy and ‘unhappy, both appear in different posts. Depending on the stemmer and lemmatizer both would take the word ‘unhappy’ and replace it with the word ‘happy’. Given that the context of that post, a personality might appear more positive. However, given that the corpus for each individual is 50 or more posts, it is less likely to make this error across all the posts for an individual. Thus, the a stemmer was also applied to compare the results, for this analysis.

**Method:**

Several key things to keep in mind about NLP is that normalization is normally key. In other words, how often does a word pop-up for a document. For example, a novel might contain one million words, and the word dragon is used only once. Making its normalized score to be 0.00001% for the word dragon. While other stories about where a dragon is a major plot point, is used 0.73% of the time. Based off that knowledge, it would be very unlikely that the first story would be about a dragon when compared to the norm. Now although that example was talking about the subject of the stories, a similar technique can be used to describe the personality type of the user based on their social media posts. Thus, the need for normalization could be significant for an NLP investigation. However, since the dataset is looking at phrases, the researcher is not positive that a normalization model would apply. Thus, using a frequency Count Vectorizer and a TF-IDF function which use normalization process, can be tested to see if it matters for algorithms’ predictions.

On top of normalization another part of most NLP processes is Distance measurement. In other words, when looking at word frequency, there might be a pattern where one type of label uses the words more common than the other. To measure those frequency there are multiple different methods, but for the sake of the brevity, euclidean distance (L2 norm) was used for all tests within this paper.

A method that will be used to determine how well the various algorithms and processes were predicted will be by doing a, holdout method, utilizing 85/15 split. Figure 4 below demonstrates the test analysis labels. What is evident is that it keeps the same proportional distribution of the entire dataset. Normally cross-validation would be suggested on-top of any hold out method to prove the results were not an issue of “lucky” sample. Due to the fact the research team ran over 20 various prediction models and kept the results in one python file for ease of the reader, the server ran out of memory to do cross-validation on all the models. However, given the large size of the dataset not doing cross-validation is less likely; due to the large sample size and training set.

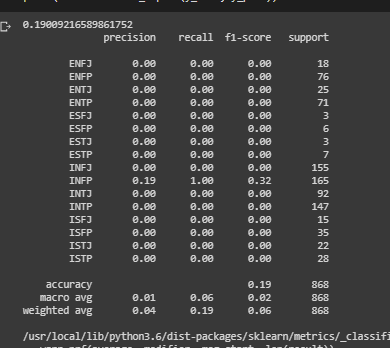


*Figure 4: Testing Labels Distribution*

The metrics that will be evaluated for the holdout method, are accuracy, precision, recall, and f-1, for how well an algorithm and model can predict using the dataset. The results will display some of the confusion matrix, and prediction values for certain models to help compare what they do well and areas for improvement. For all the models and tests, please see the attached python script. This summary of key models and methods will allow for ease of comparisons between the other methods. The best model would then be optimized further and the coefficients for the words will be displayed for each category.

Additionally, one last thing a reader should know is what the ‘majority rule’ score is for this dataset. The ‘majority rule’ is if the algorithms just predicted the label that is most common. In this dataset it is the label “INFP”, which is 19 % is the majority rule.

Figure 5 below is showing a Multinomial Naïve Bayes, with no normalization/stop-word/stemmer applied and it predicted by Majority Rule only. Multinomial Naïve Bayes (MNB) is the fastest method to predict given the analysis. And if it would have predicted well, then there would have no further need for additional analysis. But since it stuck to the majority rule addition algorithms and models were used to improve upon that method.



*Figure 5: Multinomial Naïve Bayes, Normalization/Stop-word/Stemmer Not Applied*

Three different discovery techniques were used during this analysis. The first was K-means clustering, the second was Latent Dirichlet Analysis and the third being K-Nearest Neighbor (KNN). Ideally, there would be 16 perfect clusters of topics, or one cluster per personality. Similarly, for distinct topics, if the topics were about the personality themselves. This issue will be discussed in greater detail below in the results and area for improvement.

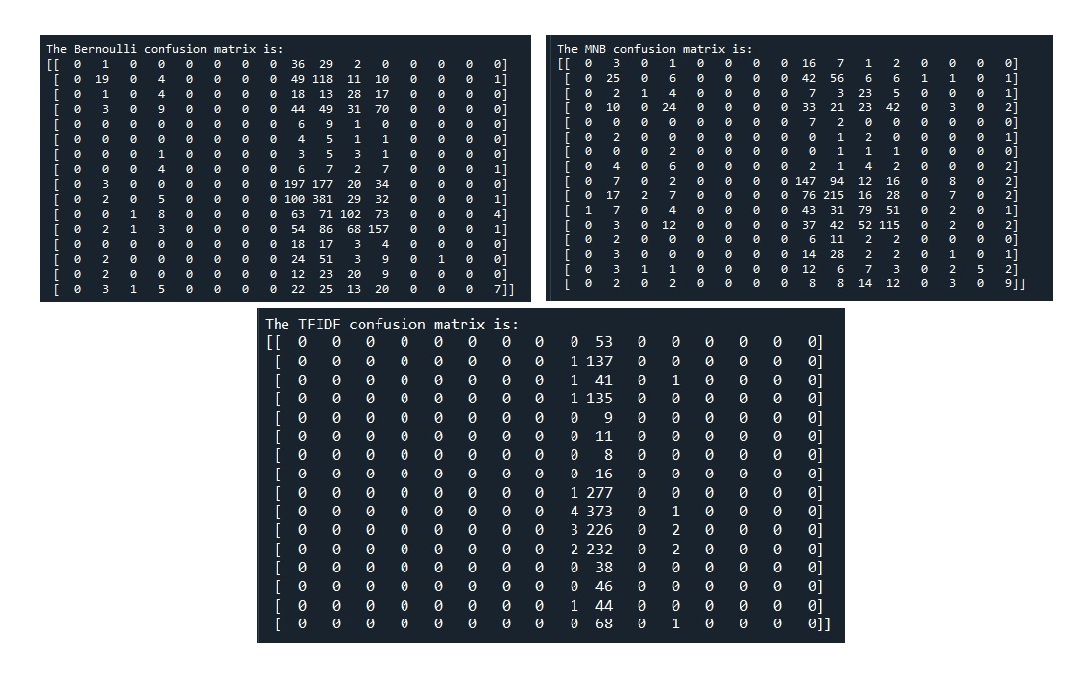
**Results:**

Table 1 below, is showing the results for the some of the better preforming models. All the models were able to achieve accuracy with doubling to more than tripling the majority rule’s score. For these models, removing stop-words and a stemmer was applied for each of their best results. However, one thing that was unique for the findings is that Naïve Bayes bucked the trend of having the best results be the TF-IDF. It preformed the best with a normal CountVectorizor. The best preforming models overall were Support Vector Classifiers (SVC) and a deep learning neural network of Multilayer Perceptron’s (MLP). In both of these cases, they were able to ~ double the accuracy of the Naïve Bayes by using TF-IDF vectorizers.

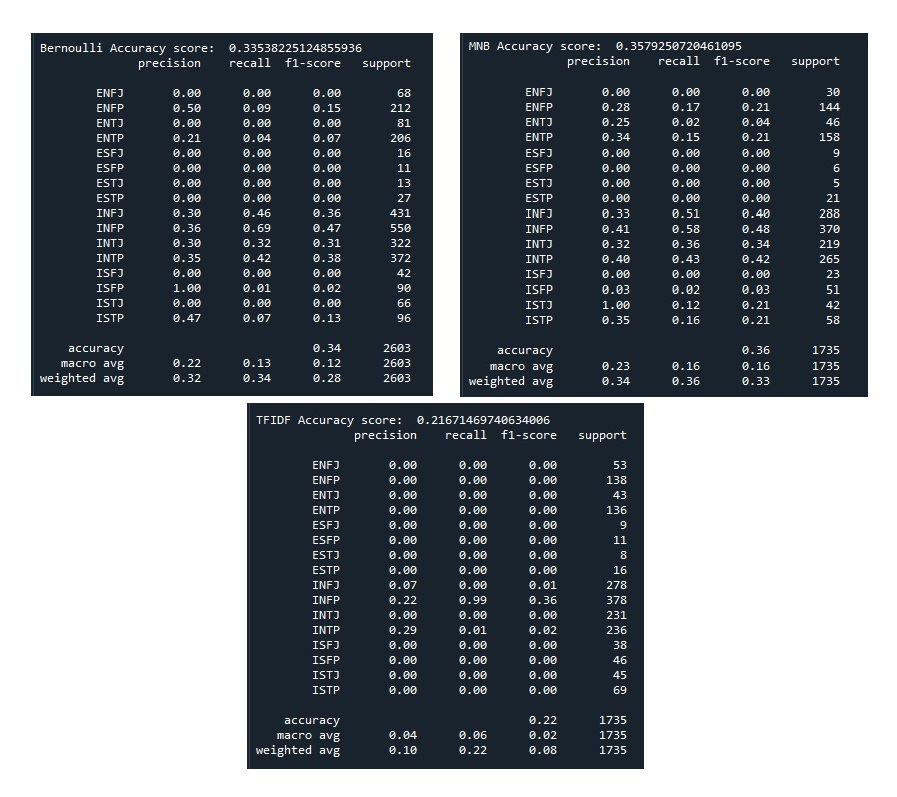
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Boolean | CountVect | TF-IDF |
| Naïve Bayes | 34% | 36% | 22% |
| SVM Linear | 51% | 59% | 64% |
| MLP | 67% | 69% | 74% |

*Table 1: Machine Learning Results on Accuracy (All Scores Rounded to Nearest Whole Integer)*

Figures 6 below presents the results of the Multinomial Naïve Bayes with Bernoulli, Count and TFIDF vectorizers. Figure 6a presents the confusion matrixes for each vectorizer. Looking at the confusion matrixes, there is no values concentration over the confusion matrix accuracy plane. This topic of reading a confusion matrix is discussed in greater detail below. The main takeaway is that although MNB does significantly outperform a majority rule model, it does allow for areas of improvement.



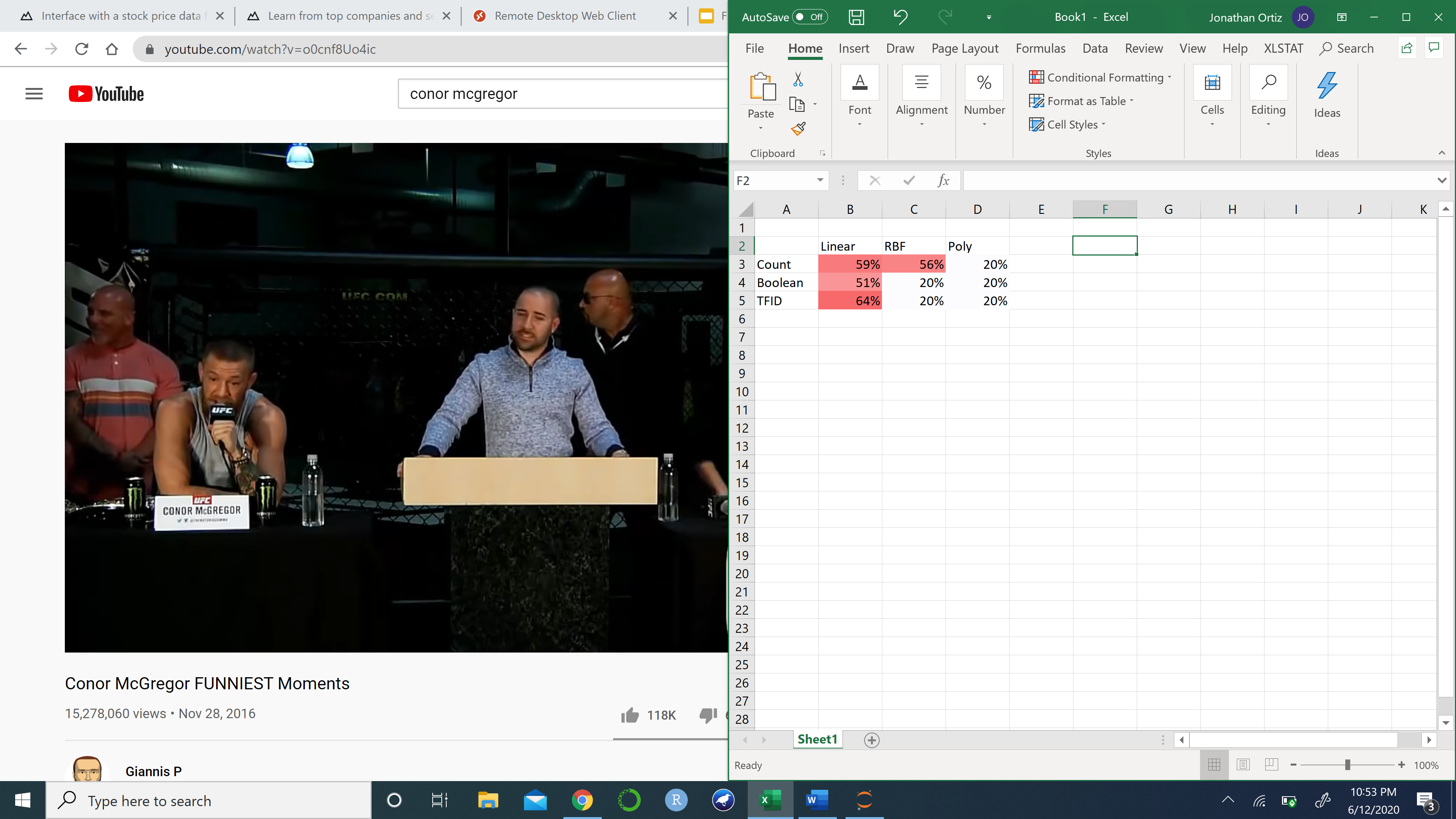
*Figure 6 a: Naïve Bayes with Bernoulli, Count and TF-IDF ion Matrix*

Figure 6b presents the Naïve Bayes accuracies demonstrating and confirming the confusion matrixes results in the following percentages.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Boolean | CountVect | TF-IDF |
| Naïve Bayes | 34% | 36% | 22% |

*Figure 6 b: Naïve Bayes with Bernoulli, Count and TF-IDF Accuracies*

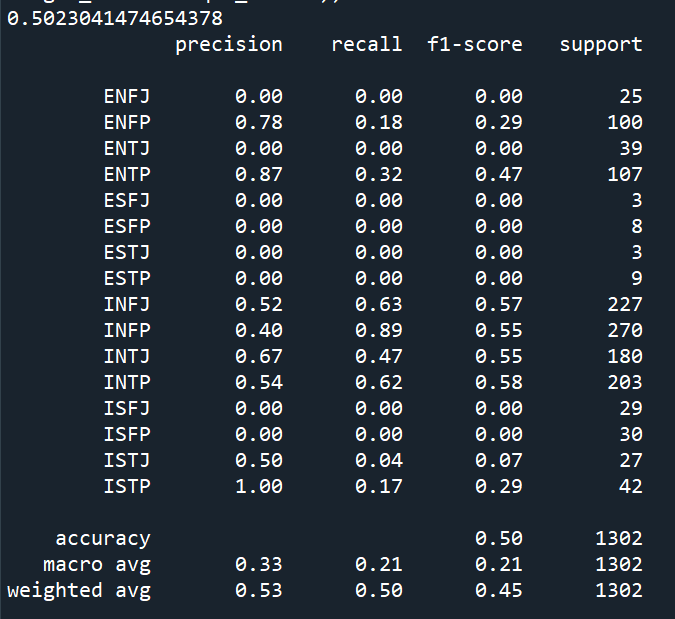
Continuing off the lessons learned from Naïve Bayes, the Results of the support Vector Machines are displayed in the table below. It is important to note the models that have 20%, would predict all the personality as majority rule type. The TFID vectorizer performed the best with 64 % accuracy.



*Table 2: Vectorizers and Kernels Results*

Random Forest was ran on the Boolean, CountVectorizer and TF-IDF data. The CountVectorizer term frequency data provided the best results with 50% accuracy. This model had difficulty differentiating between personality types that contained some of the same key components. This model also had difficulty identifying personality types that were under represented in the training data.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Boolean | CountVect | TF-IDF |
| Random Forest | 41% | 50% | 47% |

**A picture containing rain

Description automatically generated**



Figure 7: Results of Random forest with CountVectorizer Data

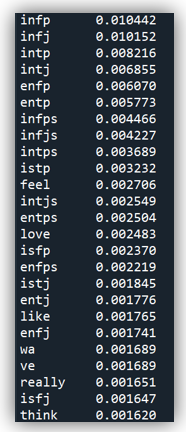
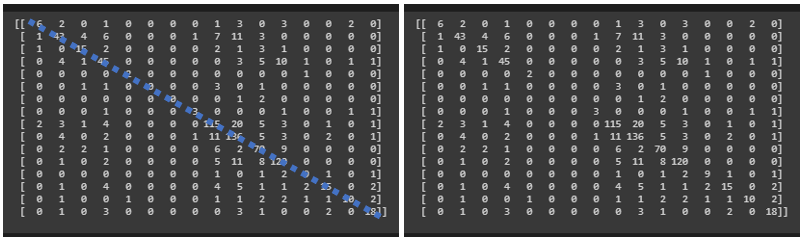


Figure 8: Random Forest Top Features

Figure 9 below is presenting a confusion matrix for the best preforming model for the dataset (MLP). There are two confusion matrix, represented for this model, as the first one acts as a guide while the second displays the values. To read a confusion matrix, one starts in the upper left corner and makes their way to the bottom right corner. This path for correct values is represented on the left confusion matrix as a blue line. The greater the values along this diagonal line the greater the accuracy for that model for predicting that personality type. For any point that is not along the diagonal line, it is a missed guess by the machine learning program. And as the table 1 displayed it was 74% accurate for its best model, and there are not a single cluster of misrepresented features, so the neural network is able to predict across label categories.

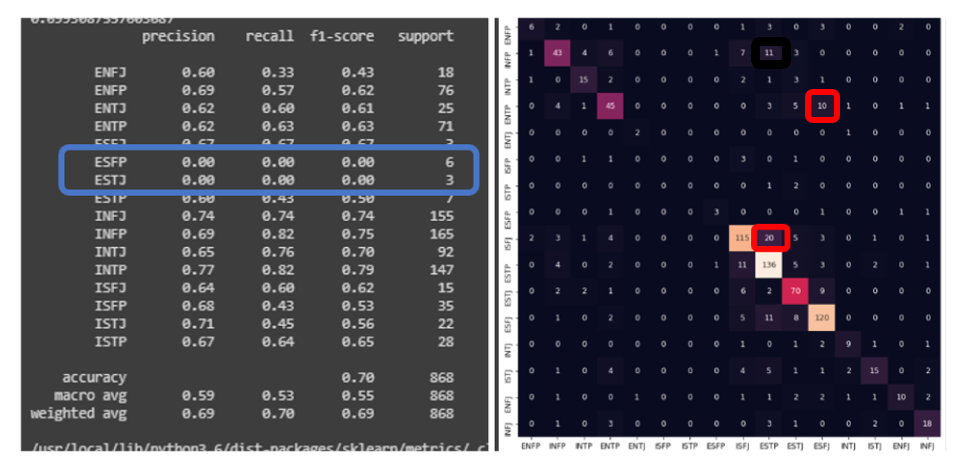


*Figure 9: MLP with a Porter Stemmer, Stop-words and TF-IDF ion Matrix*

Figure 10 below is displaying the TF-IDF scores for the best preforming model. In table one it is the MLP TF-IDF score of 74%. One thing to note is that this score is 70% instead of 74% as it does not include the stemmer. The reason for why this is signficant is that it does cost ~ 3% in accuracy to not have the data stemmed. The application created for entering ones own social media/cover letter was not able to take the stemmer as part of its process. Given more time the research team probably could have trouble shooted and fixed this error. However, with 70% accuracy it is still the best preforming predictor regradless of not having the stemmer.

Figure10 also displays 2 key things a business should be aware of when implimenting this prediction method. As shown on the left graphic for the scores for this MLP, “ESFP” and “ESTJ” were not predicted by any machine learning algorythm. They did not have a signficant footprint in the dataset to distningish themselves (1%). Another thing a company should be aware of, when using this application is that although it is 70% accurate, it is over 99% accurate when determining the 3 of the 4 personality traits. For example the largest error occurs with the two follow senarios, the value is supposed to be “ISFJ” and the tool predicted “ESFJ”. Another example is it supposed to be predicitng “INFJ” and it predicted “INSJ”. Hence, a company might find this tool to be extremely valauable or damanging depending on the importance of the most common error for that prediction type.

The right side of the graph is showing a heat map to the confusion matrix so that way it is easily determeined whether or not that 1 personality trait is significant for the cultural of the team. For exmaple, a particular team might care that they are getting a ISFJ instead of a ESFJ personality, as the “Extroversion” was a needed piece. Like in the case of feeling a sales role, having extroverted sales people help boost sales of a product. But an internal team, might not care if the person is “Intreverted” or “Extraveted” as they are looking for the other qualities, like being “Sentitve” to the details and being able to “Feel” out the auidance for project. Thus, depending on the usecase the prediciton application might be an excell tool for finding associates with personality traits to match or boster ones culture.



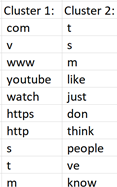
*Figure 10: MLP Metric Scores and Confusion Matrix Heatmap*

Figure 11 below is presenting the application function. When running the application it will prommp the user to place their text. Then it will take the text, put it through the TF-IDF vectorizor where stop-words are removed. Then it will take that vector and run it through the MLP to get the personality of the text, and prints out the text and personality prediction. Figure 9 uses sample text from a presidential speech, but multiple other examples can be found in the suppliment code, taken from twitter to other famous quotes.



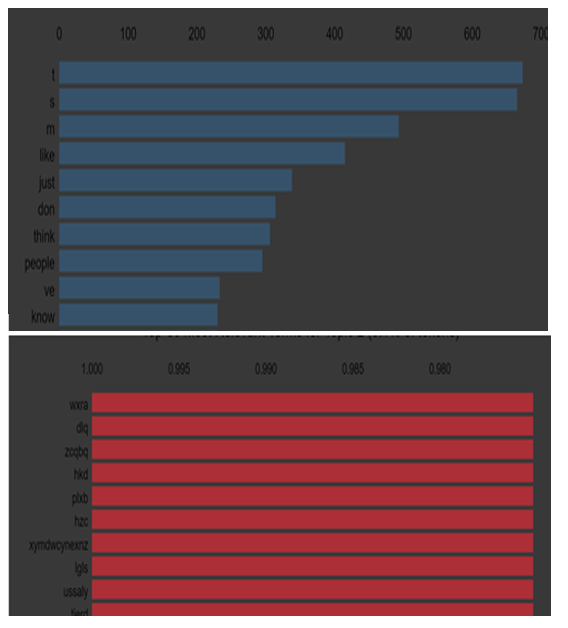
*Figure 11: Example Application Prediction*

A key necessity for buidling this application or other type of predciton model applications is that labeled data must be present. As the discovery models did not reinforce that personality types inherinetly. The results of the various clustering analysis and topic models were similar in the quantity of clusters, but differ in their respective groupings. Figure 10 shows the top features in the K-means clustering method. While figure 13 shows the top features for the Latent Dierchlet Analysis (LDA). Both results show that two clusters are prevelant. However, the LDA topics and top features did not match the K-Means features/topic. Therefore, the topics covered appear to be less around personality types and greater around social media posts or other topics discussed in the forms.



*Figure 12: K-Means top 10 terms*

Figure 12 below is showing the top 10 features from the two topics LDA determined. What is also previlent is that these topics have few terms overlapping. Like the K-Means example it appears to be finding two distinct topics being discussed in the social media posts. However, these topics do not apear to be aligned to a personality type.



*Figure 13: LDA top 10 terms*

However, one discovery model that does seem to reinforce personality clustering of posts appear to be the K-Nearest Neighbor (KNN) model. The data was ran through a frequency vectorizer and KNN model. The KNN model utilized the following parameters for k = 1 through 25. The best performance was found with k=21. The KNN model then run with k=21 and the weight set to distance which assigns more importance to points that are closer to each other.

A close up of a map

Description automatically generated

Figure 14: K Valve vs. Accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Boolean | CountVect | TF-IDF |
| KNN | 12% | 40% | 46% |

The TF-IDF data provided the best results using KNN though it is still less than 50% accurate. This model had a difficult time differentiating personality types that contain several of the same key components.

**Conclusions:**

Big Data revolutionized the way most businesses function and gain competitive advantages. In a post internet era the rise of user-generated-content, on social media platforms allow for easy access into the personality types of candidates seeking a position with said companies.

As the work market matures the more competitive the search for appropriate professionals occurs. A business not only needs to acquire and retain unique skillsets, another need emerges for retaining good candidates. These candidates must also be a good cultural fit for not only the team but to help reinforce the mission and vision of the company as a whole. Maintaining the work environment and culture that energizes employees to the business’ mission and goals ignites the road to not only current but future successes.

Achieving this level of energy can only come from within. Having teams that are compatible yet complement each other with different capabilities and various skillset puts successful businesses on the right track. This paper demonstrated that these candidates can be determined by their social media posts.

Although the methods demonstrated were successful, there is still room for improvement. As shown in the discovery, having a dataset curated towards a specific topic would help flush out the differences between personality with greater success. More time should be allotted to data cleansing and stop-words extraction, as the research group found that after every round of cleaning a new layer of effort could have been done. One last area for improvement is that applying the stemmer to the application. The issue appears to be on calling out the variable as a global variable, but not enough time to find a solution to that issue.

This paper demonstrates how public postings on various social media sites that are processed through Natural Language Processing analysis can predict the Myers-Briggs personality types. It also demonstrated several different methods for analyzing and processing text. The best preforming algorithms were used in the application that a company should use to make sure a prospective employee fits their cultural and environment. Thus, the application is designed to make a work culture better and compatible to maximize not only employee happiness, but productivity.

**Work Cited:**

Fox, Jon (2020). Week 1: Asynchronous Lecture – Introduction to Big Data

McAfee, Andrew & Bryjolfsson, Erik. (2011). Harvard Business Review: Big Data: The management Revolution. Retrieved from <https://wiki.uib.no/info310/images/4/4c/McAfeeBrynjolfsson2012-BigData-TheManagementRevolution-HBR.pdf>

Sentiment Analysis with Python NLTK Text Classification. (n.d.). Retrieved from http://text-processing.com/demo/sentiment/

Solis, B. (2009). Putting the Public Back in Public Relations. Retrieved from https://books.google.com/books?hl=en&lr=&id=xLXinA8LbTMC&oi=fnd&pg=PR9&dq=public+relations+and+social+media&ots=FutfdF3W5K&sig=IScGN4uV6joUR8pzGP0gOdTCNfE#v=onepage&q=public relations and social media&f=false

Terry, Neil, & Butler, Michael (2008). The Determinants of Domestic Box Office Performance In Motion Picture Industry. Retrieved from <https://pdfs.semanticscholar.org/0d55/44616785fdd71f11dd83abbe189219c7299c.pdf>

The Cost of a Bad Hire (Newton guest post by Danny Madigan)

<https://www.paycor.com/resource-center/the-real-cost-of-a-bad-hire>