Data Mining & Analytics Final Project

Text Mining Social Media Posts for Personality Classification

By Christian Carreras

CSC 558 Kutztown University

Table of Contents

4.2a Source of Dataset

After much consideration, a dataset has been selected to be the goal of the data analytics project. The dataset proposed is located on a dataset and machine learning competition site called Kaggle.[1] Data instances are composed of an anonymous volunteer's fifty most recent social media posts and where they fall on the spectrum of a personality test. The test used in this dataset is called the Myers-Briggs Type Indicator Test (MBTI).[2] The purpose of the MBTI is to make complex psychological types understandable and useful to all. There are sixteen types one can be classified as according to the test. Types are a combination of subtypes which include introversion (I) or extroversion (E), intuition (N) or sensing (S), thinking (T) or feeling (F), and judging (J) or perceiving (P). For example, a person who is introverted, intuitive, thinking and judging would fall under the type of INTJ. To eliminate any ambiguity, one can only have a

[1] https://www.kaggle.com/datasnaek/mbti-type MBTI Myers-Briggs Personality Type Dataset on Kaggle
[2] http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1

single trait from a subtype pair. That is, one cannot be both an introvert and extrovert, they must be one or the other. The same applies to all the other subtype pairs.

4.2b Intended Goal of Project Analysis

The goal of this dataset would be to classify a person to a personality type solely based on what they post to social media. This information can be useful in numerous settings and will be touched upon later. What is unique about this dataset is that it can come from any social media or forum site that is open to the public such as Facebook or Twitter and be applied to any person/profile that has several posts. Since such information is widely available in the public domain of the web, this dataset can be replicated or improved upon with a little time and possibly with web-scraping to expedite the process. There has been some prior analysis but not much more than putting together code to try and get the most accuracy. Not many, if any at all delved deeper into what words or combination of words define a person as who they are. That is where the true answer lies and what will be the focus of this project.

When the analysis is complete, some correlation between certain words or groups of words to personality type should be found. If there is no correlation it should be understood why one does not exist. A model should be created with a favorable degree of accuracy above a random guess and the classification rule ZeroR. A random guess being *1/n* where *n* is the number of classifications in personality type or one its subtype pairs. ZeroR being a classifier that chooses the majority class. Upon completion, this analysis should serve as a stepping stone in text mining as well as how one's words define who they are face to face.

4.2c Steps Taken to Prepare Dataset

It is worth noting that the dataset in its original form could not be used for analysis, thus some cleaning and preparation was necessary. Tools used to clean the data were Microsoft Excel, Notepad++ and Weka. In its initial state the dataset had every anonymous volunteer's social media post as one attribute with the string "|||" being the separation between posts. This string was removed and each post was split up into fifty different attributes by replacing the separator string with a comma and adding 49 other attributes for the other posts. After this, all instances of single quotes were removed as not to mess with analysis. All links were replaced with the string "__URL__" to signify that the post in question shared a link. This string in particular was used to make sure it would not match any word already present in a post. The links were replaced by doing a find/replace all using a regular expression in Notepad++. The expression used was "http[s]?\:\/\/.[a-zA-Z0-9\.\/_?=%&#\-\+!]+" in order to capture any form of link from any site. Another problem located within the dataset was that some instances did not have all fifty posts. To remedy that the string "__NAN__" was placed in all missing post fields as that string would not naturally occur. This was done by using Excel's Go to Special…->Blanks, typing "__NAN__" and pressing Ctrl->Enter. While Excel was still open, double quotes were put around all the posts using the \"@\" cell format. However, this causes triple-double quotes to appear around the fields when viewed from a text editor. This was simply fixed by using find/replace all in Notepad++ to turn all the triple-double quotes into a single-double quote. Even at this point, more problems arose from the dataset. Curly quotes appeared in the dataset most likely from how the data was scraped or where the data was scraped from. Again, another simple fix with find/replace all to turn them all into single quotes. At this point the dataset could be opened in Weka... well, almost. Weka should be opened from command-line with the argument "-Dfile.encoding=utf-8". This is since there are utf-8 characters present in some of the posts and Weka must adjust accordingly. Once the dataset was open in Weka it was converted

to a .arff file. Now that the dataset is in arff format, some fine grain modifications and the addition of derived attributes could be accomplished. Weka.NominalToString needed to be used to change the posts back to string since weka imported them as nominal values by default. Then Weka.StringToWordVector was used to split up all the posts into separate word attributes. Weka.Reorder had to be used to place the personality_type nominal attribute last while keeping the relative order of all the word vectors unchanged. The dataset was then randomized with Weka.Randomize and Weka.SortLabels was used to alphabetically sort the attribute personality_type's nominal values in a more logical way. At this point the dataset was finished being cleaned but it was taken one step further for analysis sake. Separate .arff files were made that had each pair of subtypes in the MBTI as the target attribute. This was accomplished by using Weka.Copy on the personality_type attribute, Weka.RenameAttribute on the newly created attribute and Weka.MergeWithManyValues with ignoreClass set to true to merge each pair in a subset together. Finally, removing the personality_type attribute ensured it would not interfere with classification of the new attribute. In the end, five separate .arff files were created; four for every subset pair and one with them all combined.

4.2d Application of Potential Analysis Results

With the dataset finally prepared, there should be applications for results. All it takes is a quick glance to realize that analyzing this dataset can prove most useful in both commercial and research settings. For commercial use, an employer can apply this analysis to better sort incoming applications based on the sender's social media. Social media itself can use this to adapt to one's personality type and respond in different ways to be more accommodating. Certain groups can flag like-minded individuals to mingle and work with. Matchmaker/dating, recruiting, entertainment and countless other companies could find a function for this analysis

because it is always helpful to better know and understand an individual before meeting and communicating with them. Research could also find this helpful as it can give anyone interested in understanding and finding a person's personality type a starting point or inspiration. Research fields including computer science, psychology, sociology, business, social work, etc. all can use what is found from this dataset to improve some aspect of their work. Results could range from inconclusive to eye-opening but it is difficult to deny the potential for corporations and academia alike to capitalize on such a unique opportunity for the benefit of science.

4.2e Anticipated Machine Learning Techniques

To get the analysis where it is wanted several different machine-learning techniques are being planned to be used to decipher this dataset. Nominal classification using numerous algorithms and methods is the strategy where the target attribute is the personality type. Algorithms such as rules, trees, bayes, regressions and neural networks are some of the ways this dataset will be classified. Bagging, clustering and other methods may be used to find correlations within the data and improve accuracy of a model. This relates back to the potential applications by trying to find words or groups of words that correlate with a certain personality type and making the model favorably accurate and to a minimum description length. Tools that will be used to reach this goal are Weka, TensorFlow, Python machine-learning libraries and visualization technology present within Weka, TensorFlow and Python.

4.2f Additional Important Project Details

There were some runner-up datasets that should be discussed. First off was a dataset of all international FIFA soccer games starting since 1872.[3] The dataset included date, home country, away country, location of the game, type of game and the score of each team. A new nominal attribute was derived from both score attributes to signify whether the home team won, lost or tied. Since the new derived attribute was made the target attribute, both score attributes had to be removed so that they could not be used to classify the game outcome. It was found that the prediction of a game's outcome is much more complicated than just the when, who and where and as such been abandoned. Another candidate dataset was a list of all players in the FIFA video-game database.[4] This was intended to be used to predict a player's stats, pay, or club based on dozens of attributes. However, most of these attributes assigned a numeric value based on how proficient in that area of the game the player was. Since it is not really known how they arrived at such numbers this dataset was also abandoned. The last runner-up dataset was composed of video game sales for consoles since the late 1970s.[5] It could have been used to predict sales of a new game or predict the publisher/designer of the game based on sales. Yet this dataset was doomed from the beginning because of its numerous missing fields that would have taken too much time to be filled in by one person in the time allotted. Eventually the social media MBTI dataset was picked for its interesting subject matter and its possibility to be applied in multiple scenarios. In a society where we define a person by what they *do*, it would be even more fascinating if we could define a person by what they *say* or *write*.

---

[3] https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017
[4] https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset
[5] https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

5.2a Additional Data Collected for Analysis

Extra data collected was used to compare the breakdown of the demographics in the dataset to the demographics of personality types in real life.

https://www.careerplanner.com/MB2/TypeInPopulation.cfm

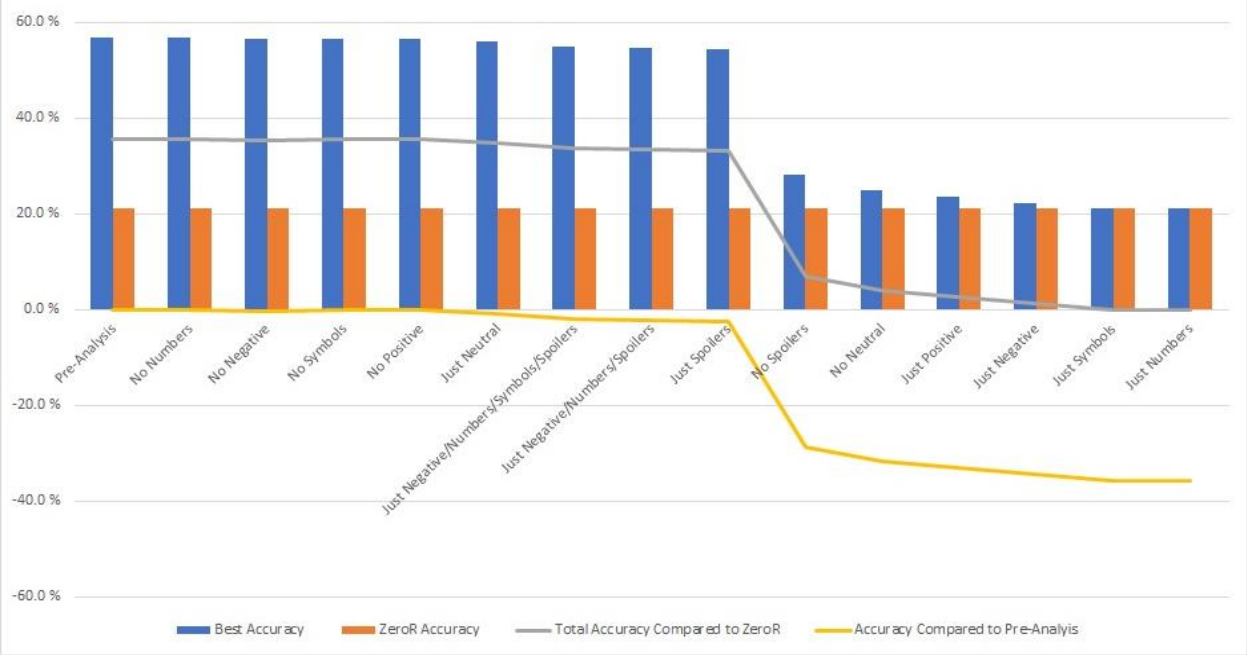
http://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm?bhcp=1

5.2b Results Compared to Project Goal

To some degree the goal of the project was a success. However, it was not a success in the way originally planned. The original goal states that the "...goal of this dataset would be to classify a person to a personality type solely based on what they post to social media" and "when the analysis is complete, some correlation between certain words or groups of words to personality type should be found." Another goal was that "a model should be created with a favorable degree of accuracy above a random guess and the classification rule ZeroR". All of these goals have been met but it did not play out as assumed. From the start, the word vectors of the model were grouped into categories. The categories delegated were positive, negative, neutral, symbols, and numbers. First symbols and numbers were joined into two separate derived attributes "__sym__" and "__num__". What was found was that this decreased accuracy and was therefore undone. Symbols and numbers were kept as their individual attributes. These attributes represented number digits zero through nine and symbols such as ampersand, equals sign, or even a combination of symbols such as emojis. Word categories on the other hand are a little more complex than symbols and numbers. Each category was tested by observing the accuracy of just the words under that category and the accuracy of the model with only those words removed. Accuracy was tested with the full 16 classification model and one of the binary

classification models derived from the original. The binary classification model picked was the thinking/feeling model because it had the most balance out of all the derived models along with a favorable amount of accuracy above ZeroR. To give an idea of what constitutes a negative, positive or neutral word, here are some examples. A positive word evokes positive feelings in the subject of the sentence in most or some contexts such as caring, love, smile, or friendship. Positive could also mean a positive adjective like happy, healthy, fun, comfortable or wonderful. Negative words are just the opposite. They are words like angry, anxiety, bad, evil and terrible that conjure unfavorable emotions and outcomes. Neutral words are just that; neutral. They do not have any emotional leaning one way or another. Examples would be about, currently, during, eye, here or room. No context can ever make neutral words not neutral as they are usually vague or meaningless by themselves. Now that words were split into categories, testing could begin to find any correlation with personality type. After testing each category, it was noticed that no change was improving the accuracy but in some cases the accuracy dropped dramatically. In some cases, accuracy dropped down by 35%. This loss was noticed when the attributes only consisted of only positive or negative. At this point, it was deduced that there was a subset of words that made the model substantially more accurate and they had to be part of the neutral words. After reading through the list of neutral word attributes a set of words stood out. These words, that could be called “spoiler words” were words that identified the personality type classifications. In other words, these words “spoiled” the classification for the model by telling them the result and thus made the model more accurate. To test this theory, all attributes except these spoiler words were removed and lo and behold, very little accuracy dropped. Now it would be a lie to say that positive, negative and neutral do no play a part in the accuracy. Although it would not be an exaggeration to say that their role is insignificant to prediction. In

this case insignificant only means a difference of (~2%).

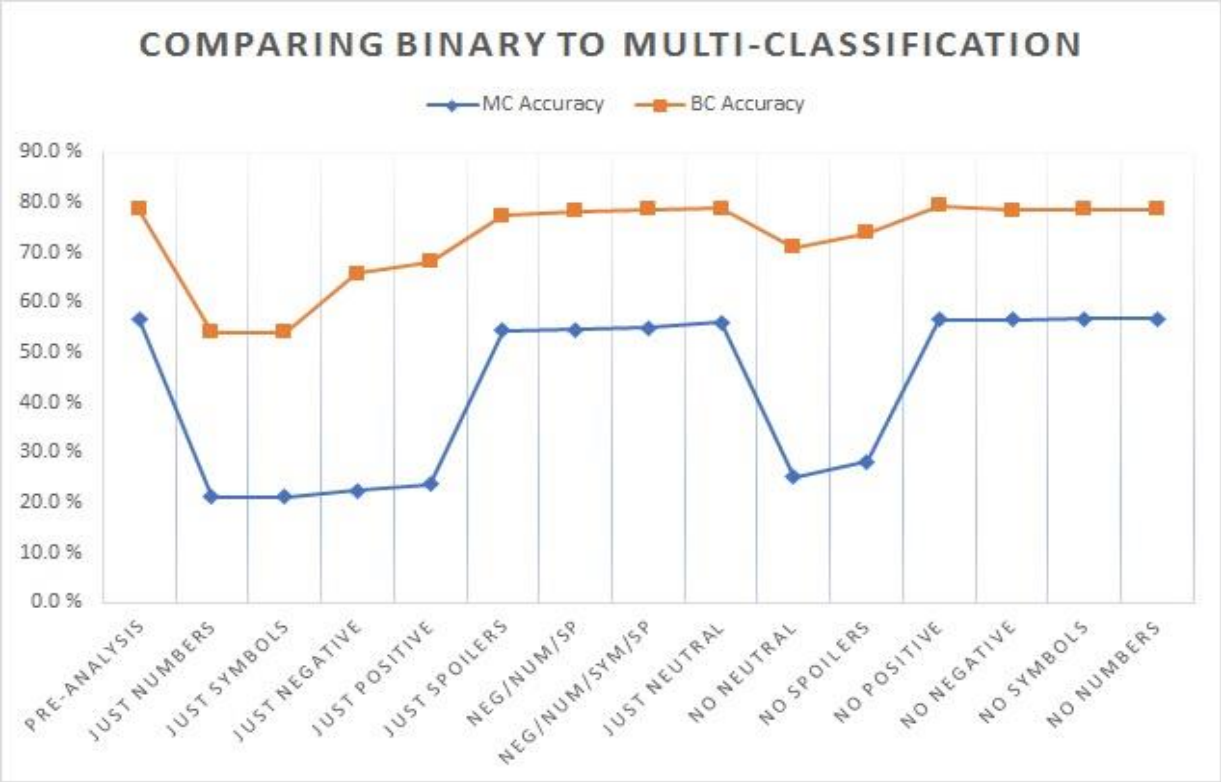


Some of this difference could be mitigated by adding back words in addition to the spoiler words but what is the point if no real predicting power is coming from them? With that said, the goal to find a correlation between words and personality type were found but it was not natural. More likely than not the correlation speculates that either some or most of the instances in the dataset were fabricated or the anonymous volunteers who handed in their posts were not so anonymous. Why would so many “random” and “anonymous” people put their personality type in their social media posts right around the time this dataset was collected? This was not coincidence. The volunteers must have personally known the data collector or known the purpose of the dataset to post their personality type seemingly out of ignorance, malevolence, or by order of another. All said and done this project did have a fair amount of accuracy above ZeroR. Yet knowing how that goal was met is bittersweet to say the least.

5.2c Machine Learning Results

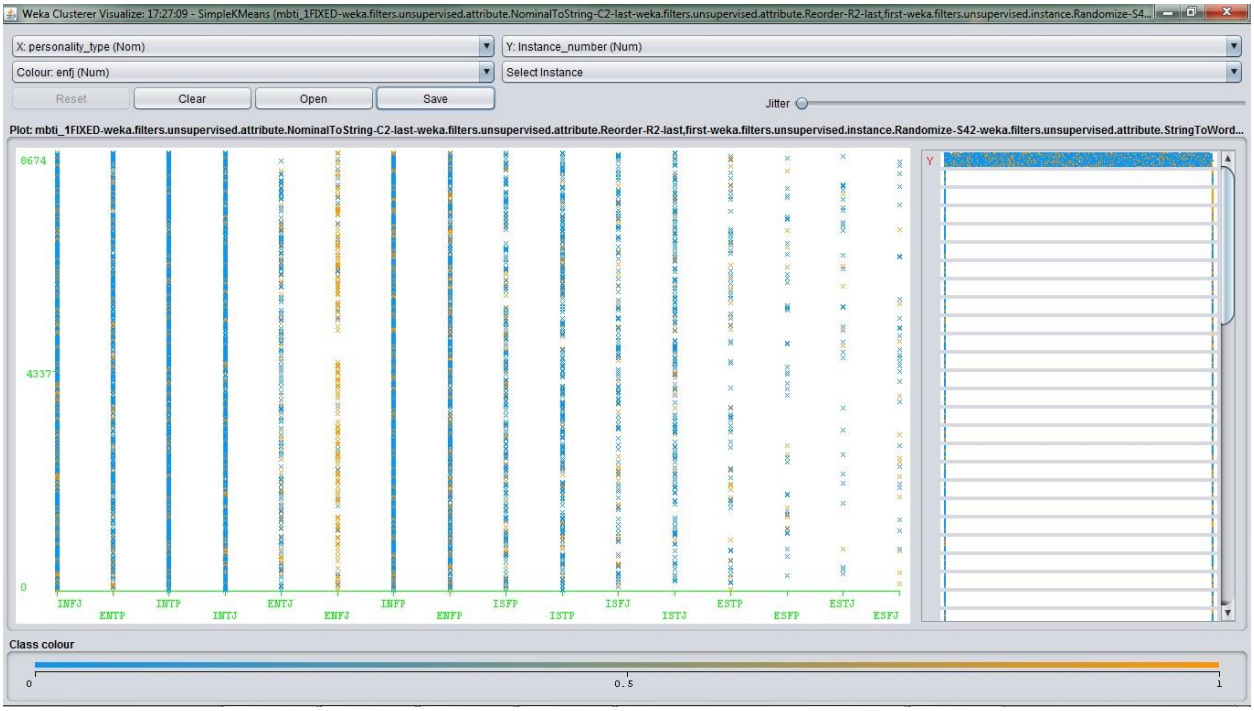
To meet the goals of the project there had to be an application of machine learning. Starting out, the dataset was put through trials of multiple algorithms to find which one returned the greatest accuracy. There were two versions of the dataset being used with these tests. One was the multi-classification model that classified all the personality types at once and the other was the most balanced out of all the files created in section 4.1 of assignment four. This other file was a binary classification model that predicted between a thinking and feeling individual. But before anything could be done, a baseline had to be made to compare to. ZeroR had to be ran to get the minimum accuracy to compare to when using other algorithms. With no modifications done the model scored 21.1182% on the multi-classification model and 54.1095% in the the binary classification model. After minimum accuracy was secured through ZeroR, all other classification algorithm could be tested with. A natural successor to ZeroR, OneR was up next. OneR scored 28.6686% in the MC (multi-classification) and 63.2046% in the BC (binary classification) models. Trees were up next and they did not prove to be too successful in increasing the accuracy to substantial levels. J48 clocked in with an accuracy of 37.9366% in the MC model and 67.5504% in the BC model but it is worth noting that it clocked in much later than other algorithms when it comes to time. The same could be said for RandomForest which had an accuracy of 34.9049% in the MC model and 74.2594% in the BC model. RandomForest did however perform faster and better than J48 with binary classification. The last tree used was RandomTree and it performed last or next to last in both models. It performed worse than ZeroR in multi-classification at 17.3833% and slightly better in binary classification at 56.3804%. Last to be used were Bayes algorithms. These algorithms performed the best out of all the algorithms for both files. NaiveBayes scored 41.5101% in the MC model and 77.0836% in the BC model. BayesNet took home the trophy for both files though by scoring 56.8069% in the MC model and 78.536% in the BC model. BayesNet represented the highest score for both files

without any modification to the original dataset. Binary classification performed better on all accounts but that is to be expected as it is always easier to predict two classes compared to sixteen. Not much more accuracy was gained after the pre-analysis tests. The multi-classification model never had a higher score and only had one additional score match the highest. Contrary to those results, the binary classification model had four additional scores that were higher than the preliminary tests. Those four were the following: a test with no symbols that scored 78.5821% (0.0461% higher than pre-analysis), a test with just neutral words that scored 78.7435% (0.2075% higher), a test with no positive words at 79.2622% (0.7262% higher) and a test with only negatives, numbers, symbols and spoilers at 78.6282% (0.0922% higher than the pre-analysis). These results show that no change increased accuracy by any significant amount but also supports that the highest scores always had the spoiler words included. Thus, the spoiler words were the key to the accuracy seen in the models. Here is a chart comparing the scores of the binary model to the multi-classification model.

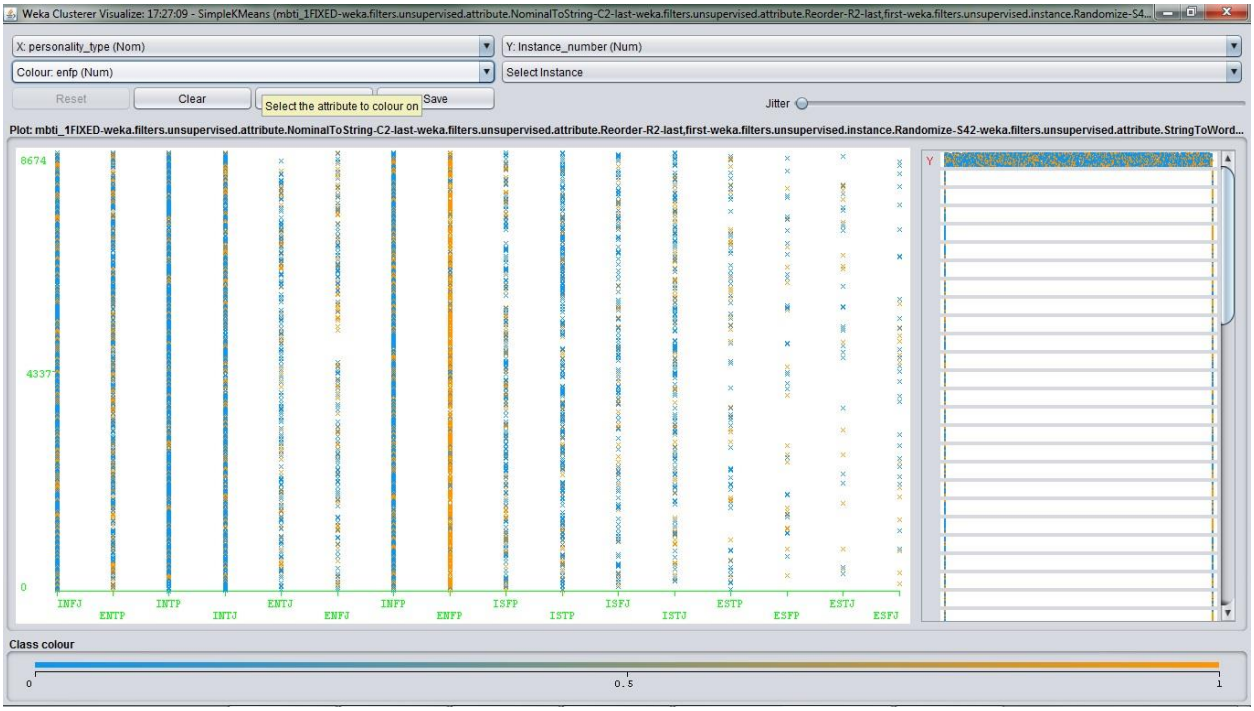


5.2d Results of Additional Techniques

To confirm suspicions, clustering was used alongside visualization in Weka to show that spoiler words were being used to predict the personality type and that they were indeed coming from individuals from within that personality type. SimpleKMeans was used with sixteen clusters for the sixteen class multi-classification model. What it presented was a clear picture of these spoiler words were grouped together with their matching personality type.



Here in the Weka Visualization tab shows that mostly all the instances of the word ‘enfj’ were from people that were classified as ENFJ or Extrovert, Intuitive, Feeling, Judging.



Here is another picture showing most of the instances of ‘enfp’ falling under ENFP. This trend keeps going on for every personality type. This cannot be a coincidence. Clustering helped prove that this dataset was biased and ultimately contaminated for one reason or another. Unfortunately, nothing more could be done than prove this dataset was not genuine.

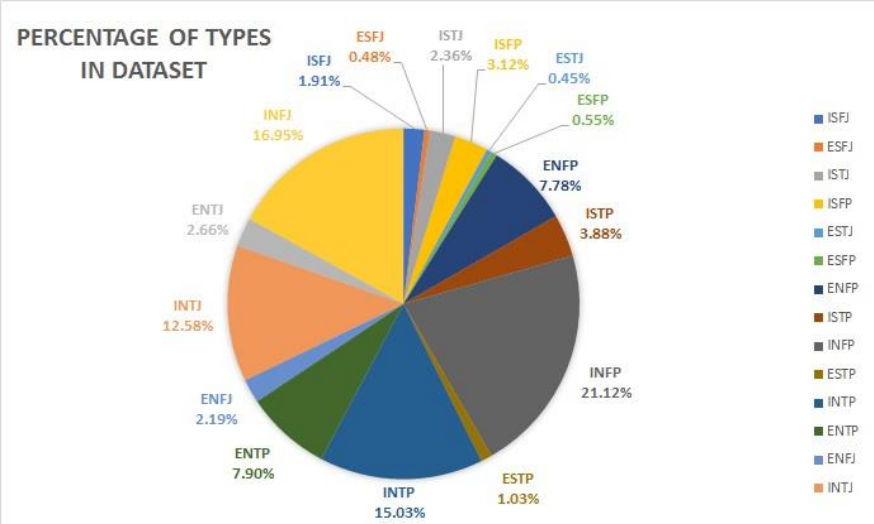
5.2e Application of Actual Analysis Results

With the fact that this dataset was not as well composed as originally thought, these results cannot be used as previously intended for a commercial or research setting. They can be however used as a guideline for looking for fraudulent or biased data in text-mining. Biased data could devastate a project and it would be wise to recognize it before it is too late. Thus, any company or person wishing to get into text-mining could find this helpful to figure out if the data they are mining was dreamt up by some individual or is disproportionate for some purpose.

5.2f Additional Important Analysis Details

After the project and when mid-project assumptions were proven correct, the distribution of the personality types within the dataset was compared to the frequency in real life. What was unearthed was just as troublesome as the spoiler data found halfway through the project. Some of the rarer personality types in the real world made up most of the personality types in the dataset. On the flip side, most of the majority personality types in the real world were the minority in the dataset. This is quite the coincidence and conundrum. Perhaps the data collector was lucky enough to happen upon a large enough group of people who make up only 1.5% of the population to fill 16.95% of the dataset. That is 1,470 people in one place that agreed to have their social media posts placed in a dataset for text mining. The numbers do not add up. Again, on the other side, the personality type which most people fall under in reality is only represented by 166 people. This shows that the dataset is severely unbalanced to the ratio seen in life. If the track record so far had been immaculate, this could have been a great amount of coincidence. However, seeing that spoiler words were found in the dataset, there is no way this dataset can convince anyone that it is not skewed out of portion. One can hope a reputable MBTI personality to social media post dataset will show up in the future because this unfortunately turned into a sad excuse for finding bias within a text mining dataset.

The pie chart below shows the distribution of personality types in the dataset.



The next pie chart shows the distribution of personality types in real life.