University of Strathclyde Glasgow

Assignment 1

CS982: Big Data Technologies

Ilia Iliev

201332409

1. Introduction to the Problem

Dota 2 is arguably one of the most popular online games and is being played by millions of players ever since its debut in 2013. To illustrate how popular the game is, current estimations show about 13 million unique players logging into the game each month (Wolmarans, 2016) with a peak of more than 1.2 million concurrently in-game players (Dota 2 - Steam Charts). And yet, for some people, Dota 2 is a lot more than a game; there are millions of dollars involved in competitions - both in prize pools and team sponsorships. For instance, the largest yearly event where only the top players are invited compete currently has a prize pool of 24 million USD - a number that will continue to increase in proportion to the amount of interest in the game (The International 2017) . Even though the esports scene is just emerging, Dota 2 is one of the driving factors and it shouldn't be a surprise that many people are attempting to apply analytical approaches to the game.

While the game itself is very complicated and almost impossible to fully master, in it's core, there are two teams of 5 players competing against each other. Players can interact with each other and battle with the opponent team over objectives until one team comes out on top. It should be noted that not only knowledge of the game but also teamwork is vital for winning a game of Dota 2. And while the very top players usually compete with the same rosters, the vast majority of players tend to play with people they don't personally know; the teams are formed by the game matchmaking algorithm. This raises a very interesting question where people chances of winning are very heavily dependent on collaborating with random players. While the chat feature is intended to discuss strategies and objectives, it is also often used to vent out frustrations between the team or taunt the opposing team. This project will attempt to look at and analyse how the chat is related to the team's chances of winning.

2. Description of the problem

This report looks at a dataset of chatlogs gathered from 50 000 games. Since each game has two opposing teams, there are a total of 100 000 chatlogs, each corresponding to a team of 5 players. In each game, there is only a single winning team, so there are 50 000 losers' and 50 000 winners' team chatlogs in this dataset. This project will try to use various data from those chatlogs and try to find common factors that might lead to winning or losing of a match.

There are certain specifics related to the chatlog that need to be addressed before preparing the dataset for further analysis. Because the game is very fast-paced, most players don't have the time to fully explain their intentions and resort to typing a shortened version of common phrases, which may appear nonsensical to an unfamiliar person. For instance, 'gg', the most common phrase found in those chatlogs, is an abbreviation for 'good game' - is usually typed when the game appears to be over, as a sign of respect. Or, alternatively, a player could say 'gg' prematurely, to imply that someone made a misplay so bad that the game might as well be over. There are also ways of taunting - for instance, after a particular flashy outplay, it wouldn't be uncommon for the player to taunt his opponent by typing '?' in the chat. Another very common taunt is 'ez' an alternative way of saying 'easy', usually implying that winning didn't take much effort. Using those phrases in such manner has a few implications that can be used to further analyse the chatlogs.

In Dota 2 chatlogs, sentence structure tends to be ignored and shortened phrases and abbreviations are a lot more common. Most of those phrases are typically used in certain scenarios, in order to illustrate how an individual player is feeling about what is happening in the game. This raises an interesting problem that only by looking at the chatlog, there is a possibility to infer what has happened in the game and what is the chance that the team ended up winning or losing the game.

3. Summary Statistics of Analysed Data

This section shows the various considerations of the type of data used for the training of the supervised and unsupervised algorithms. As the original format of the data needed to be pre-processed and brought the right format, here will be discussed some of the decisions regarding how this is done.

The original dataset contains the chatlog in its raw format - as it was typed by the player. However, it is common that people type each word or phrase on a new line, even when they are typing out a whole sentence. Moreover, as it was justified in section 2., common phrases and abbreviations will be analysed rather than whole sentences. That is why each line of the chatlog is 'tokenised' (Himmelstein, 2016) and words are considered as independent of each other. This is a simplification that makes sense for the vast majority of the cases but it should be noted that it is not perfect and that there is a degree of natural language, even if it is considerably less

than in traditional text. This is why this project will not attempt to analyse every single word encountered in those chatlogs; instead, only words and phrases that are proven to directly correlate with win rate will be considered for analysis.

A small amount of games do not include anything to analyse, as no player has used the chat feature. That is why those games are not used for the rest of the report and only games with some sort of chat log are being considered. Consequentially, games that had no chat at all were more likely to result in losses so the final dataset contains a tiny bit more wins than losses (51.4% wins and 48.6% losses). Moreover, Figure 3.1 illustrates the relationship between chat length and win rate:

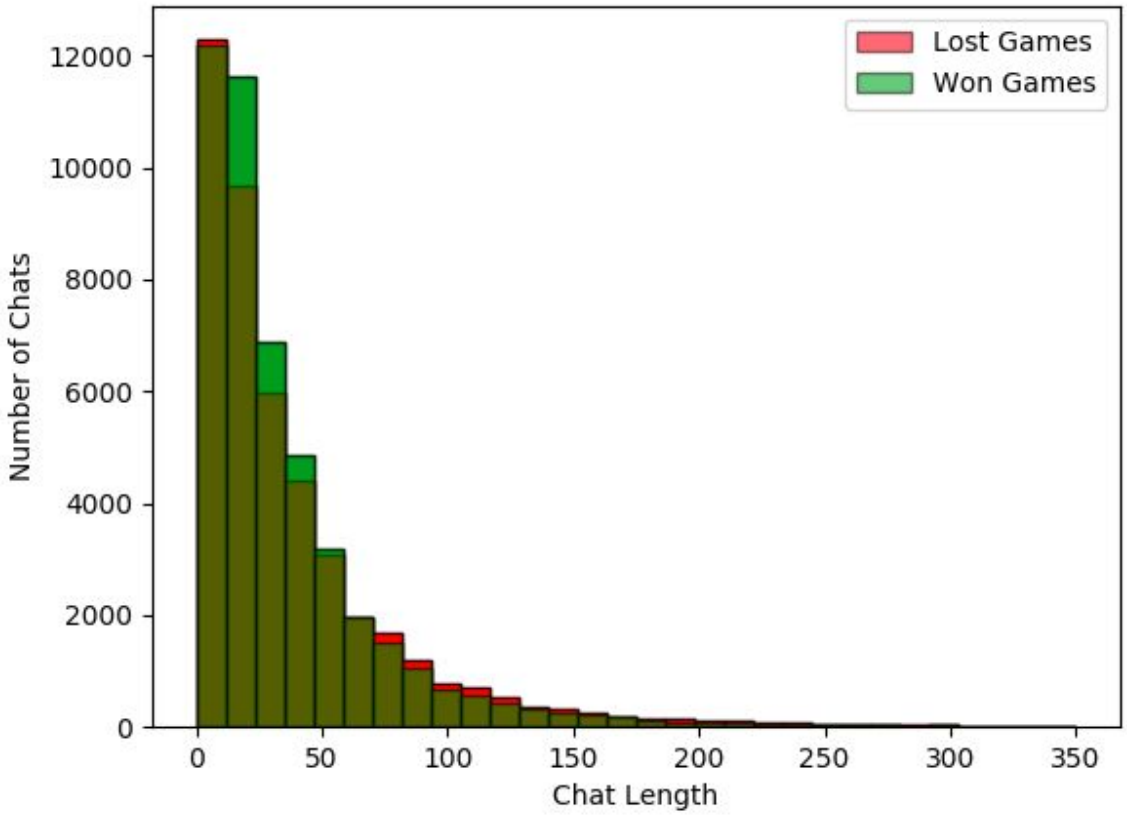

Figure 3.1 - Word Count in Chat Related to Win Rate

As can be seen from Figure 3.1, there is correlation between chat length and win rate. Teams that use the chat without overusing it are more likely to win. This can mean that there is some

degree of communication that might improve the chances of winning. However, for chats that contain more than 60 words, there is a higher chance to result in a loss rather than win. Potential explanation is that long chats are typically indicative of back and forth insults being thrown between the teammates, which can be demoralizing.

Initially, the problem considered taking into account the presence of foreign characters into a chatlog. Dota 2 is an international game and there are many people that don't speak English and only converse using foreign characters (e.g. cyrillic). There is a possibility that this might impact the communication between the team and, consequently, the win rate. Figure 3.2 below shows the amount of chats that contain foreign characters:

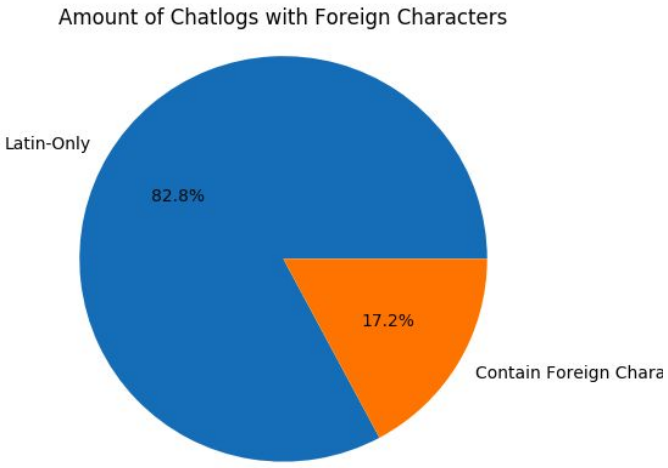

Figure 3.2 - Amount of Chat with Foreign Characters

As can be seen from Figure 3.2, a little more than 17% of the chats contain some form of foreign characters. This means that the amount of foreign characters is statistically significant and there could potentially be useful. However, as this project is concerned with win rates, Figure 3.3 looks at how the presence of such foreign characters impact the win rate:

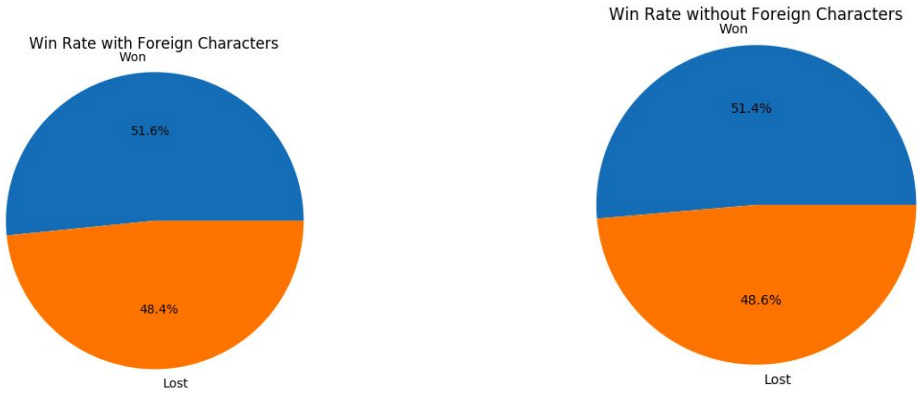

Figure 3.3 - Win Rate Comparison for Foreign Characters

As can be seen from Figure 3.3, chats that contain foreign characters are just as likely to win as chats that do not. In other words, the existence of such column in the dataframe is not supported through statistics and, despite the initial consideration, the idea of looking at presence of foreign characters is dropped.

To complete the dataset, each token from the chatlogs is considered in terms of winrate. However, as there are tens of thousands unique words and some of them have only a few instances, words at least once in at least 10% of the whole dataset (10 000 occurances) are considered. Figure 3.4 below plots how those words are related to the win rate: what amount of chats that contain the word/phrase did end up in a win:

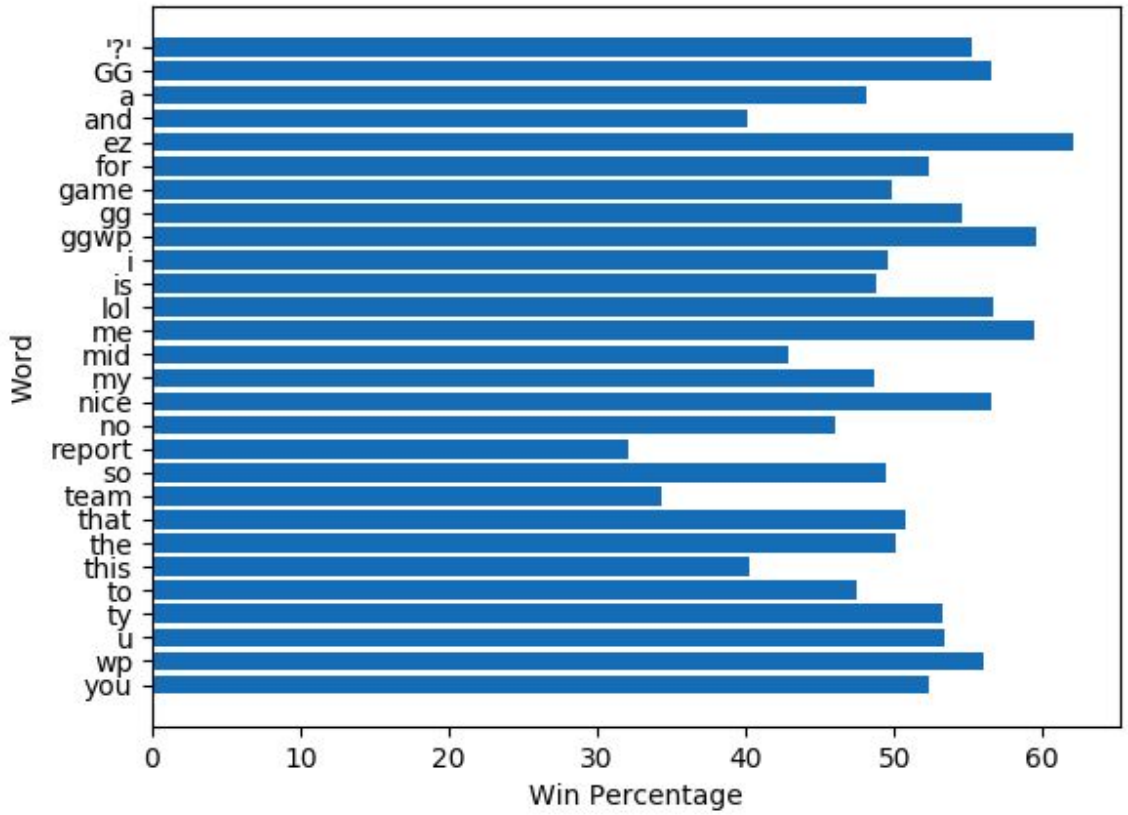

Figure 3.4 - Win Rate Comparison for Most Common Words

From figure above, it can be seen that some of the most common words have direct impact on win percentage while others only have a minimal impact. This is likely due to the fact that some words cannot be considered on their own as they are usually part of a sentence structure. An example is ‘so’ which is usually used as a word to start off a sentence. Such words cannot be analysed on their own and a more complex methods must be used. As a consequence, all words that impact the win rate by only a small amount (range is 45-55%) are taken out of the dataset. The result is shown on Figure 3.5:

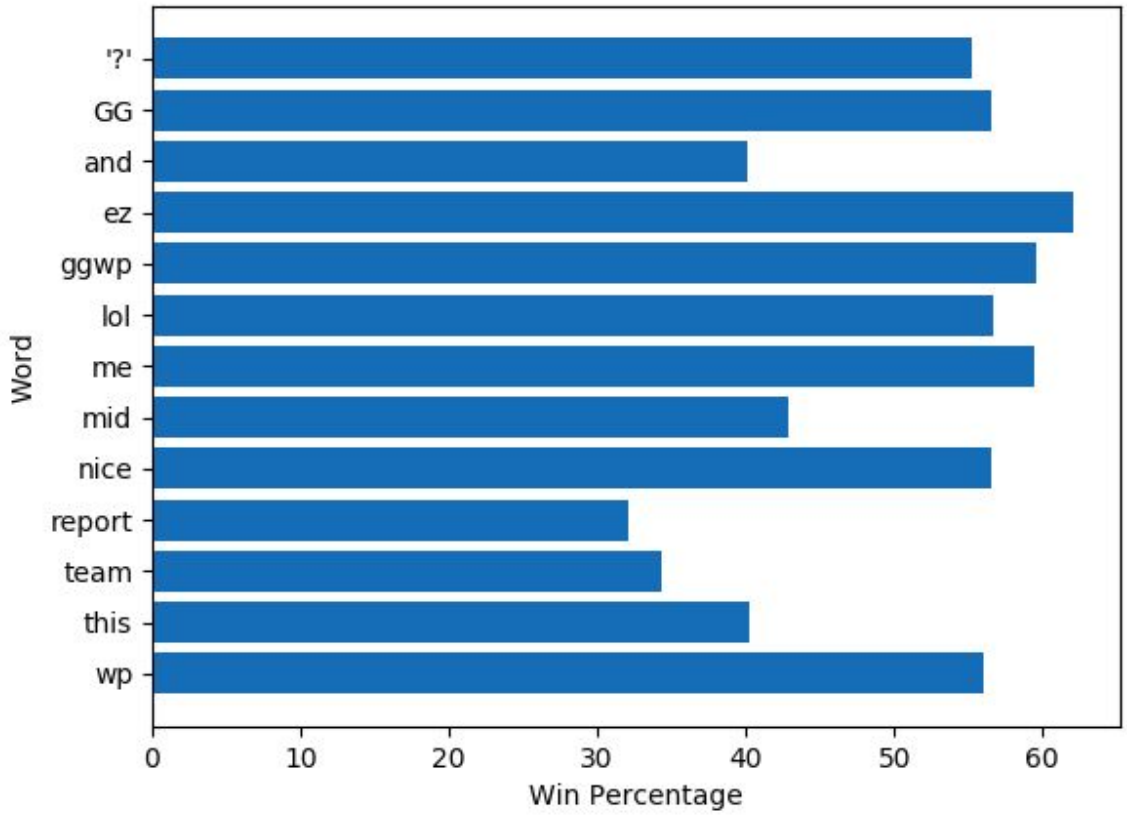

Figure 3.5 - Words with Win Rate Used for Further Analysis

Alternatively, Figure 3.6 shows the amount of chats with the chosen words and phrases:

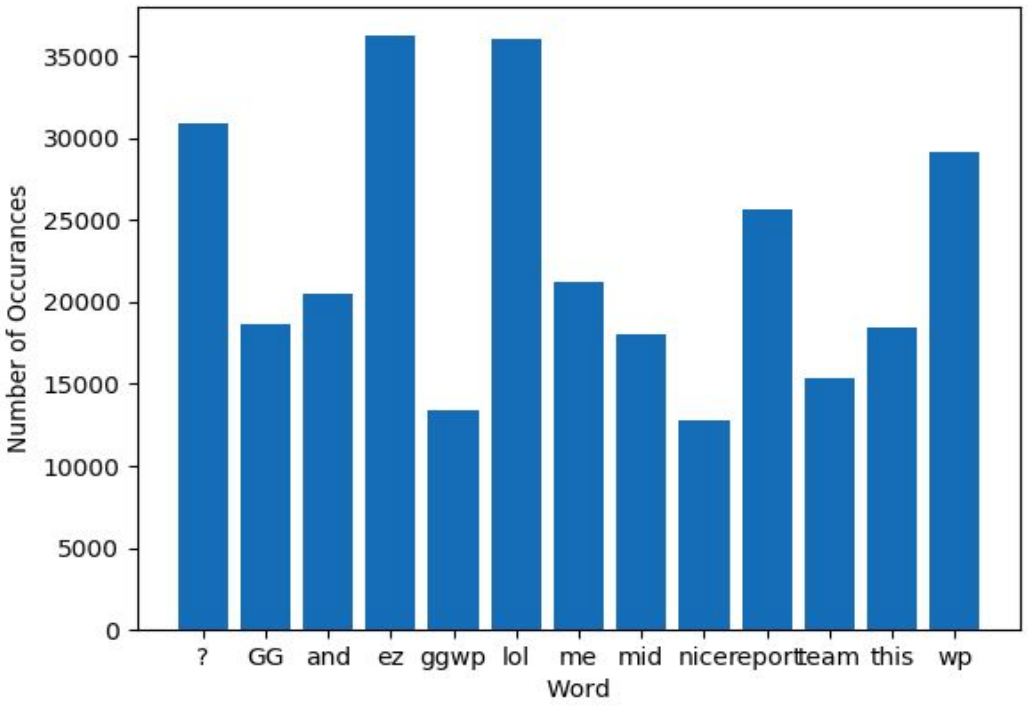

Figure 3.6 - Number of Chats that Contain the Chosen Words

With the words from the figure above, the dataset is complete and learning algorithms can be implemented. The final dataset has 88925 chatlogs, each associated with a chatlength and a field for one of the common words. If a word is present in the chatlog, it has a value of 1 or, alternatively, a value of 0 if the word was not found in the chatlog. This is used to set the baseline comparison - since there are 2 possible outcomes, a random classifier would have accuracy of 50% so anything more accurate than can be considered as a degree of success.

4. Unsupervised Learning - K-means

In order to better illustrate what the dataset looks like, K-means clustering is used in an attempt to find clusters of wins and losses. This means that there are 2 clusters that should be found. The data is divided in 75-25% split for training and testing. Since K-means is very prone to scaling, all features are automatically scaled. Since the problem has 16 dimensions and one column to predict, in order to visualise the results, the features are reduced to only 2 dimensions using PCA. However, this is purely for visualisation purposes and the entirety of the data is used to train the clustering algorithm.

Initially, the dataset is fed in the algorithm in its entirety. The results are shown in Table 4.1 and Figure 4.1:

Table 4.1 - K-means Results for Full Dataset

	Precision	Recall	F1-Score	Support
Losses	0.48	0.87	0.62	10844
Wins	0.45	0.16	0.17	11388
Total	0.47	0.48	0.38	22232

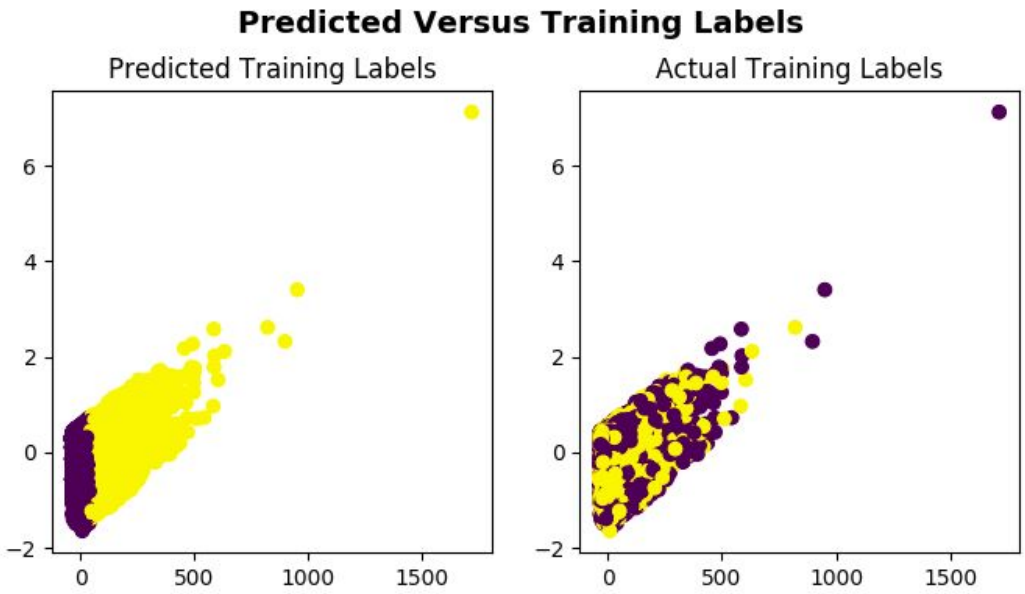

Figure 4.1 - Visualised PCA clusters for K-means on Complete Dataset

In an attempt to understand the results shown above, K-means clusters have very little overlap with the actual cluster of wins and losses. In fact, the results are essentially random. There are various reasons for this. The dataset uses both categorical (binary) data to represent presence of common phrases as well as numerical data for chat length. In other words, the results shown above do not account for this scaling as it is impossible to scale numerical and categorical data at the same. For this reason, the chat length is removed from the dataset, so that only

categorical data is present and the algorithm is re-run. Results are shown on Table 4.2 and visualised in Figure 4.2:

Table 4.2 - K-means Results for Categorical Dataset

	Precision	Recall	F1-Score	Support
Losses	0.58	0.29	0.38	10870
Wins	0.54	0.80	0.65	11464
Total	0.56	0.55	0.52	22232

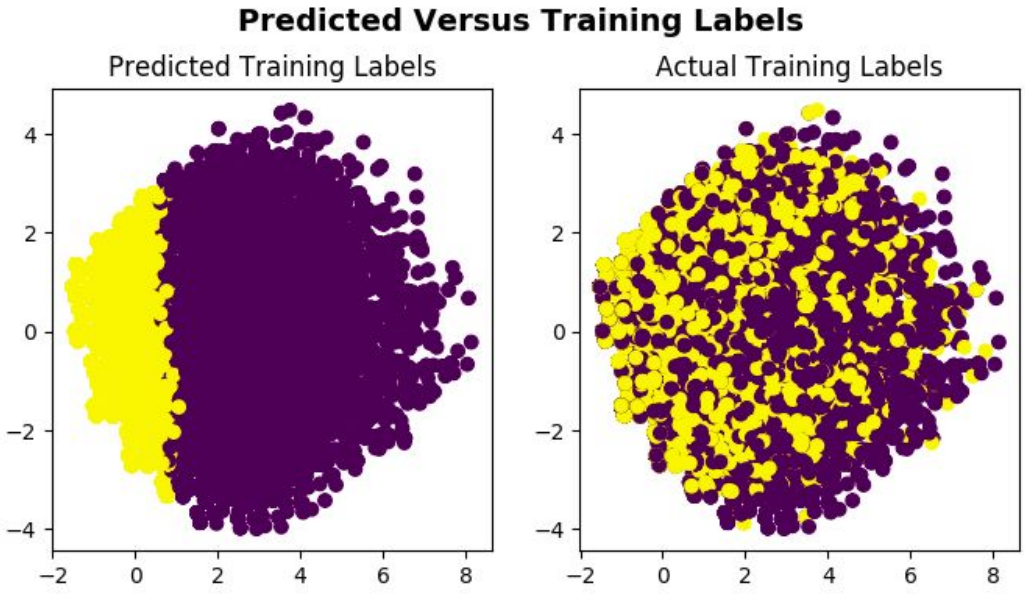

Figure 4.2 - Visualised PCA clusters for K-means on Categorical Dataset

As can be seen from Table 4.2, the results are no longer random and there is some degree of correlation between the found clusters and wins/losses, so there is an improvement. Unfortunately, the results are comparatively poor and the found clusters do not represent wins and losses. To get a better understanding of what is happening, Figure 4.2 shows what the reduced dimensionality dataset looks like. While there is a slight pattern to be found, for the most part, the wins and losses data is very heavily mixed together and it is difficult to find clusters using K-means. That is why an alternative approach to finding the wins and losses clusters is taken.

5. Supervised Learning - Neural Network

Since data appears to be too complex to detect clusters using K-means, a more advanced technique is used. Artifical neural networks were chosen for their properties - they work well with complex data of many dimensions that are not correlated in a linear fashion. Moreover, as the dataset has a very large amount of chatlogs, there is sufficient amount of training data to implement a neural network solution (Forbes, 2016). Table 5.1 below illustrates the results from applying ANN to the chatlog dataset:

Table 5.1 - ANN Results for Full Dataset

	Precision	**Recall**	**F1-Score**	**Support**
Losses	0.62	0.64	0.63	10934
Wins	0.64	0.62	0.63	11400
Total	0.63	0.63	0.63	22334

As can be seen from the results above, precision and recall seem to be give consistent results for both wins and losses. The result is evaluated at ~63% which is considerably better than random and has substantial improvement over K-means which had issues with the recall rate. The testing set is further reduced using PCA and correct and predicted labels are shown on Figure 5.1 below:

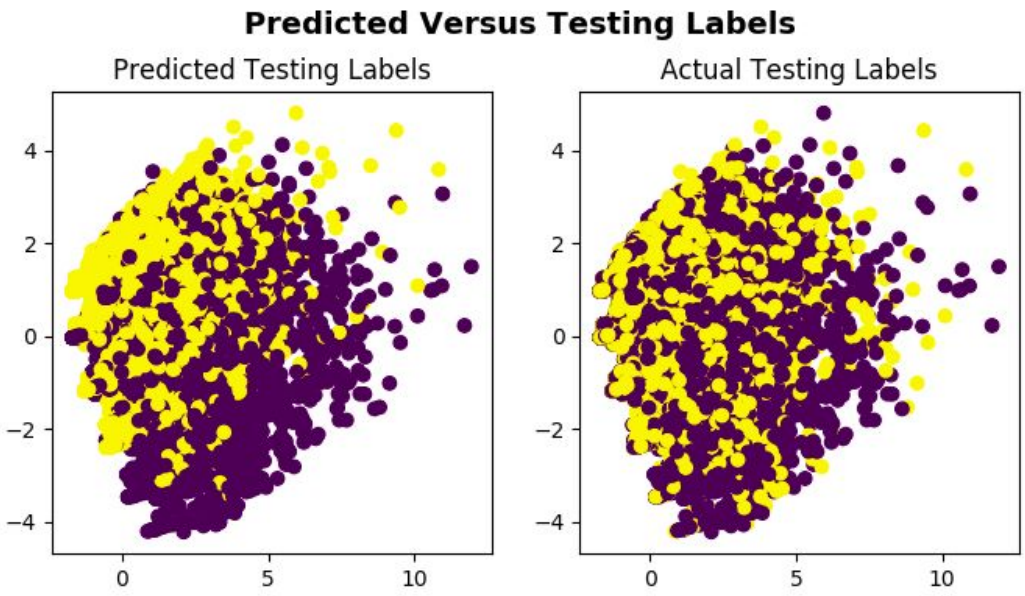

Figure 5.1 - Visualised PCA clusters for ANN on Full Dataset

As can be seen from the figure above, ANN manages to find the general trend for wins and losses. It struggles to produce more accurate results because there are isn't a defined border to separate wins from losses; instead, the data appears to scrambled. However, relating the results to the statistical analysis, the majority of the data extracted from the chatlogs does not deviate by more than 10% from the random point of 50%. This means that it is unlikely to produce a better result by using the explained methodology at looking at common phrases and abbreviations.

6. Reflection and Conclusion

This report introduced the problem of looking at Dota 2 chatlogs and trying to infer win rates by looking at common patterns. This involved pre-processing of the data to extract relevant fields that has a direct correlation to win rate - chat length and a list of common phrases. Those features are then used to create a custom dataset to be fed in one unsupervised and one supervised learning algorithm.

In an attempt to find clusters of wins and losses the K-means algorithm was used. While the results are better than the random baseline, the algorithm struggles in distinguishing between

two clusters as there isn't a clear separation between wins and losses. It was concluded that this problem is better suited for a supervised learning technique.

A more advanced algorithm - Artifical Neural Networks - is used to better distinguish between wins and losses with an accuracy of ~63%. When looking back at the chatlogs and correlating phrases to win percentages, it can be seen that only a small portion of the chat information deviates by more than 10% from the truly random point of 50%. In conclusion, an algorithm that gives the correct result 63% of the time is considered a success for the chatlog dataset.

7. References

"Dota 2 - Steam Charts." *Steam*. Steam, n.d. Web.

Forbes, Brett. "Why Aren't Artificial Neural Networks Used for Everything?" N.p., 18 Feb. 2016. Web.

Himmelstein, Daniel. "Split (explode) Pandas Dataframe String Entry to Separate Rows." *StackOverflow*. N.p., 09 Oct. 2016. Web.

"The International 2017." *The International 2017*. Valve, n.d. Web.

Wolmarans, Kyle. "Dota 2 vs. League of Legends: Updating the Numbers." *CriticalHit*. N.p., 22 Sept. 2016. Web.