Text Mining of Audience Opinion in eSports Events

Bruno Omella Mainieri* Pedro Henrique Cacique Braga Leandro Augusto da Silva Nizam Omar

Universidade Presbiteriana Mackenzie, Programa de Pos-Graduação em Engenharia Eletrica e Computação, Brazil

ABSTRACT

Chat in video game tournament live streams is a rich source of data relating to the championships themselves. With that in mind, this study sets out to analyze Riot Games' *League Championship Series* (LCS) Twitch chat, with the express goal of understanding the structure of opinionated expression through chat. This knowledge allows identifying the most popular teams and players among fans based on text messages, which would make for most interesting options for potential investors and sponsors. Employing techniques from the fields of Big Data and Text Mining over the *Hadoop* platform, mention counting and opinion measuring operations are applied to a set of over half a million chat messages collected over three weekends of competitions.The processed data is then graphed, revealing the existence of fan favourite teams and widely cited players, even among those which under-performed in the tournament, in addition to contributing to the understanding of the prevalence and impact of chat *emotes* in the expressing of sentiment.

Keywords: text mining, big data, eSports, data visualization.

1 INTRODUCTION

Professional video games competitions - a phenomenon named *eSports* - have seen an expressive growth in the past years, both in frequency of events and in viewership numbers.[8][10] The few hundred fans gathering at small lan houses at the turn of the century became millions of enthusiasts, filling up stadiums and venues around the globe. Players grew in popularity alongside the games themselves, many of which have achieved celebrity status among the fan base. Companies are already sponsoring teams and individuals, in the same way they would a professional athlete. [10]

Born in the digital age, the *eSports* phenomenon is deeply connected to technology and new media, and makes use of these as means to promote ever greater interaction among spectators, as well as between them and competitors. Through online live streaming platforms, hundreds of thousands of users watch casual, practice and competitive matches daily, making their hobby a social activity part of their everyday lives.[10][11]

This constant fan participation through digital media creates opportunities for measuring and analyzing the opinion of the public: it becomes possible to capture and process each spectator's every interaction, in a manner analogous to being able to hear what each person in a stadium is saying during a sports match. This information could be of value for potential sponsors looking for good players or teams to represent their brands.

In order to carry out such an analysis, Data Analytics and Big Data technologies are employed to collect, store and process data in distributed fashion. Specifically, the Hadoop platform[1] can be cited, being used for distributed data storage as well as R applications for processing.

With that in mind, this project's goal is to be a proof of concept for the viability of this kind of analysis applied on a specific live streaming platform, Amazon's *Twitch*.

*e-mail: bruno.omella@hotmail.com

This paper is structured as follows: the second section presents the concepts relevant to the objects of study, including employed technologies. The third describes the particular example tackled in this research. Obtained results are analyzed in the fourth section. The fifth and final section discusses the value generated by this analysis as well as possible future works in the same field.

2 FUNDAMENTAL THEORY

With the goal of measuring viewer opinion in *eSports* streams, it is fundamental to understand the concepts that make up this scenario: the live streaming platform and the content being broadcast, as well as the technologies necessary to undertake this analysis.

2.1 Live streaming

A recent development in media vehicles enabled by the growth of fast, easily accessible internet connection, live streaming is to *eSports* what television is to traditional sports. The process itself consists of three fundamental parts: the broadcaster, who creates and shares the content; the platform, which enables the real time transmission of said content; and the viewers, who consume the content and, often times, give instant feedback through means also provided by the platform, such as chat rooms.

Live streaming is not exclusively employed in *eSports*, being also used for a myriad other types of content, but its influence over the growth of video game competitions is notable: major events often draw viewership numbers exceeding tens of millions of fans from all over the globe.[4][8] As such, this type of broadcast has begun to appeal to investors and sponsors in recent years.

2.1.1 The Twitch Platform

Twitch is a live streaming platform owned by Amazon in which users can broadcast any gaming or lifestyle related content in real time for other users to watch. Any user, person or company, can create their channel, through which can be made broadcasts, known as streams. Each channel has a live chat room, in which users may publish comments about the contents of the stream in real time.

It is also a major platform in the eSports broadcasting business: based on data from Twitch's public API, market research firm Newzoo reports over 80 million hours of competitive gaming content were watched on Twitch in April 2017[7], with major events nearing or surpassing one million concurrent viewers.[2] During these broadcasts, tens of thousands of messages are published every hour, many of which refer to the matches being played, competing teams and players.

Still of note in this platform is the presence of *emotes*: small images which are displayed embedded in text messages when certain words are typed in. Each of these *emotes* has a certain meaning behind it, which may vary depending on context of use. Some *emotes* represent opinions or feelings, such as approval or joy, while others have very specific meanings, referring to famous players or community figures. These *emotes* play a large role in the way meaning is transmitted through chat messages, as will be detailed further in this paper.

2.2 eSports

As mentioned above, *eSports* is the general name given to the phenomenon of competitive video gaming, practiced in similar fash-

ion to traditional sports, in that organizations form teams composed of professional players to participate in tournaments.[4] Except for the clear distinction of its digital nature, it closely resembles sports in the way it is organized, as well as its social and business impact: video games which become competitive titles, not unlike popular sports, amass large numbers of casual and hobbyist players, from which emerge a small fraction of competitors. As organizations form to enable and regulate structured competition, matches and tournaments begin attracting spectators, teams and players gain fans, and business have opportunities to thrive in this scenario.

Due to its relation to commercially developed games, the history of *eSports* is often linked to the games themselves. In it's current form *eSports* often has it's origins traced back to early 2000s South Korea, where Blizzards' *StarCraft* became not only a commercially successful game, but also a popular competitive title - and one which proved to be a favourite among viewers.[10] Now, the major titles which drive forward the industry include Valve's *Dota2* and *CounterStrike: Global Offensive*, Blizzard's *Hearthstone* and *Overwatch* and the most played and most viewed of all, Riot Game's *League of Legends*.[13]

As the popularity of eSports events grew among fans, companies began venturing into sponsorship deals with teams, players and organizers, in the hopes of targeting this enthusiastic audience, composed mostly of males around 20 years old.[11][14] Sponsorship revenue industry-wide is expected to surpass $250 million in 2017.[10]

2.2.1 The *League Championship Series*

As the most watched video game in the world, *League of Legends*, or LoL, regularly draws in hundreds of thousands of viewers even for small events - reaching past a million for premiere ones.[13] In this game, two teams of five players each compete in matches that last between 20 and 50 minutes in average, during which the participants must employ strategy and technique to gain advantages and ultimately destroy the enemy base.

One major event was chosen as object of study for this particular research: the North American and European *League Championship Series*[9] (LCS) for the Spring season of 2017. Serving as qualifiers for international events in each of these regions, the LCS is a regular event consisting of two seasons each year. During these seasons, matches are played every weekend, as teams compete against each other to achieve the top ranks in their region. At the end of a season, the playoffs occur: a tournament to determine the final rankings of teams in the region for that season.

2.3 Big Data

In its simplest form, Big Data is data which cannot be processed or otherwise handled by traditional, centralized computing systems. The theoretical representation of this type of data is usually given by the "4 Vs of Big Data": data which has great Volume, high Velocity, ample Variety, and needs Veracity. Furthermore, a fifth V is often included: the goal of obtaining Value.[3][5]

In this case, the Volume of messages is not so exceedingly large as to require specialized infrastructure to handle, but being only a small proof of concept to eventually be applied to much larger data sets, it makes sense to propose a system capable of handling far more data. The Velocity is given by the pace in which messages are published, seeing as the projected system is capable of capturing them in real time. Variety comes from the unstructured nature of chat messages, and Veracity is a concern for any attempt of valid analysis. Finally, the goal is to obtain Value for interested sponsors and investors, who want to make informed decisions when choosing teams and players to work with.

Technically, Big Data is often described in terms of a stack of technologies, which work together to capture, store and process data in a distributed and scalable manner[3], as illustrated in Figure 1.

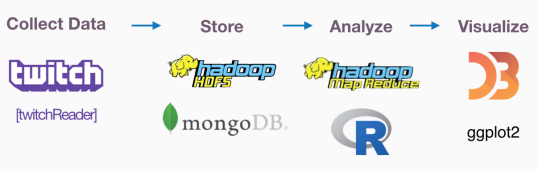

Figure 1: Simplified model of employed infrastructure

2.3.1 Distributed Storage and Processing

Hadoop Distributed File System (HDFS) handles storage by breaking data down into blocks, which are distributed among the machines in a cluster. As such, overall storage capacity can be increased by adding more nodes to the cluster as needed. The same scalability is offered to processing via MapReduce, in which functions, such as filtering and counting, are applied to each block separately ("Map") before being combined into the final result ("Reduce").

In this study, this MapReduce process employs Text Mining procedures. These are concerned with deriving information from textual data via identification of patterns and trends. In order to do so, collected data, which is often varied and loosely structured, needs to be parsed and preprocessed by means such as removing stop words (which do not contribute to the meaning of the text) and stemming (reducing words to their roots).

3 Experimental Methodology

The proposed study consists of collecting and analyzing chat messages from the North American League Championship Series (NALCS) and European League Championship Series (EULCS), with the goal of identifying more popular players and teams among spectators, both in terms of raw number of mentions and proportion of positive opinions.

3.1 Data collection

For this application, the collection of data is handled by a software named *TwitchReader*, developed specifically for use in this project. It makes use of Twitch's public chat API[6] to connect to specified channels and capture messages as they are published.

Chat messages were collected from three weekends of the Spring 2017 playoffs (April 8th and 9th; 15th and 16th; 22nd and 23rd). Just over 48 hours of chat were collected, resulting in a total just short of 650.000 messages. All messages were in English.

During the days data was collected, six teams participated in each region's event, as shown in Table 1. Each team played in two of the three weekends covered, but the teams that placed 5th and 6th on each region played one less match, having less time on stream.

Region	1st	2nd	Other teams
NALCS	TSM	Cloud9	Pheonix1, CLG, FlyQuest, Dignitas
EULCS	G2	UOL	Fnatic, Misfits, Splyce, H2K

Table 1: Teams on each LCS region

3.2 Data Storage and Processing

The collected text is saved directly to HDFS. Once the desired volume of data is collected - in this case, after one full day of LCS live streams - it can then be processed. In the proposed architecture, this

is done via MapReduce: data is handled while still distributed, and only the reduced values are ever pulled out of HDFS.

The operations themselves consist of text mining procedures, namely frequency counting and simple opinion evaluation. These are programmed in R via the TidyText package.[12] This package offers tools to operate over textual inputs by breaking them down to single words, placing those in large tables and then running traditional table operations, such as joins, on them. It also handles preliminary operations such as stop word removal via the same operations.

Frequency counting is done simply by filtering the lines of these massive tables, each of them containing a single word, in order to find references to players or teams, and counting the number of matches. Measurement of opinion is achieved by attributing a numerical value ranging from -5, for most negative, to 5, for most positive, for each word in the data set. This value is determined by a sentiment lexicon provided by the package, structured as a table containing common English words as well as their corresponding numerical values. Words which are not present in the lexicon receive a value of 0. *Emotes* were manually added to the lexicon, as these play a large role in conveying meaning in chat messages. Overall opinion value is then calculated on a message level, by adding the values of each individual word on the message.

4 Experimental Analysis

Initially, queries were made over the data set that had the goal to obtain a better understanding of the structure of chat messages, in particular regarding the way opinion is expressed.

The first such query aimed to measure the prevalence of *emotes* - small pictures that appear when users type in specific words into their chat messages - in the collected data. To that end, a list was created containing the 30 most frequently used *emotes*, and the amount of occurrences of those was measured against the total number of words.

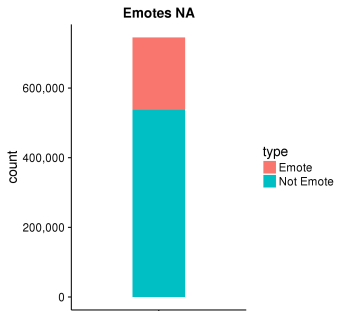

Figure 2: Proportion of *emotes* to non-*emotes* in NALCS chat

In all subsets of data this query was applied, the result was that about 25% to 30% of all words in the captured chat messages represented *emotes* (Figure 2). Given their frequent use, it was clear that any attempt at identifying meaning had to take them into account, which justified adding those to the sentiment lexicon used.

Before exploring positive or negative opinions in messages, an analysis of raw popularity was made, based solely on the number of mentions of each team (Figure 5) or player (Figures 3 and 4) within the data set.

Of interest to note is that neither region had the most mentioned players exclusively from the top teams, pointing to a possibility that fan interest is not entirely dependent on present performance.

When it comes to teams, the results are somewhat different. In North America, the top 2 teams were clearly the most mentioned by fans, while in Europe the most mentioned team (Fnatic) actually placed third. Going back to Figure 4, it stands to reason that Fnatic

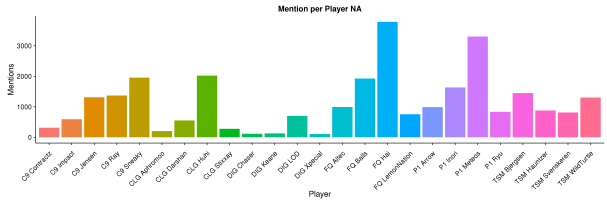

Figure 3: Player mentions in NALCS

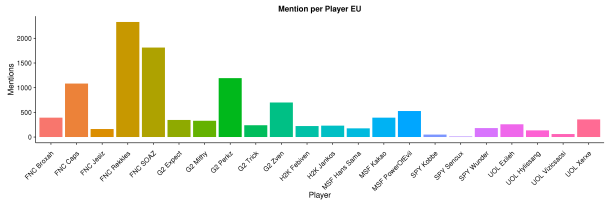

Figure 4: Player mentions in EULCS

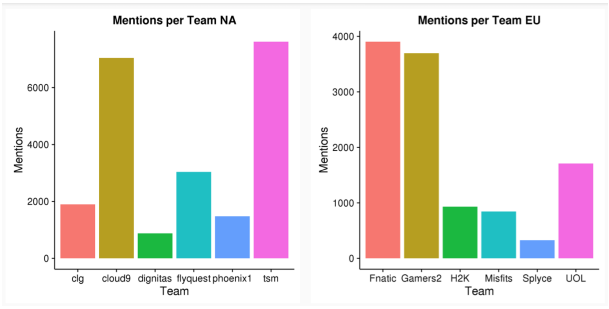

Figure 5: Team mentions in NALCS (left) and EULCS (right)

and its players are very popular among the spectators, despite not even going to the Grand Finals. This kind of information, can be of great value to a potential sponsor: it seems, from this limited sample, that Fnatic is a consistently popular team, and while it is impossible to affirm based on this simple analysis alone, it could be an indicator of strong fan loyalty to the team.

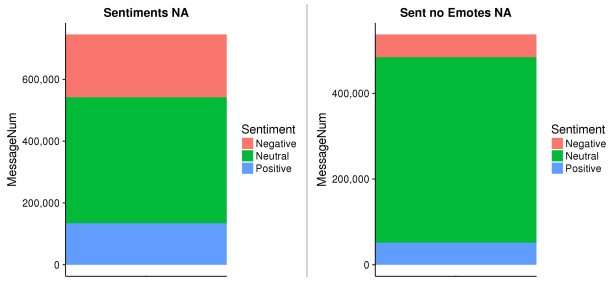

Figure 6: Sentiments in NALCS chat with *emotes* (left) and without *emotes* (right)

After that, queries were made to understand the impact of *emotes* in the measurement of opinion. While undoubtedly a very simplistic approach, the employed method, given the simple structure of chat messages (being composed mostly of a single sentence of under 5 words) and the apparent ease of identifying meaning through emote usage, was expected to give satisfactory results for this initial research. In order to validate the belief about the influence of *emotes*, two different measurements were made: one including

them, and another discarding (Figure 6).

The results from this procedure revealed that a significant fraction of all opinionated messages (averaging around 60% on the LCS data set) rely on *emotes* to convey their meaning, corroborating the theory that these small images are integral part of the form of expression within these chat rooms.

Finally, the aspects of mention counting and sentiment measurement were combined to identify viewers' opinions on each team. This was done by calculating the overall opinion value of each message (including *emotes*) and attributing that to any teams also mentioned in the message. Once again, this approach is rather simplistic, and doesn't handle, for example, messages in which a viewer compares two teams and expresses how one is much better than the other. In exploratory queries, however, it was noted that less than 1% of all messages mention more than one team, and these kinds of semantic structures are uncommon in chat messages.

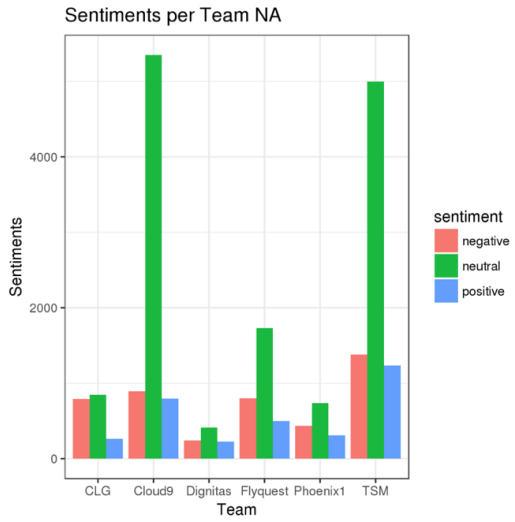

Figure 7: Opinion on NALCS teams

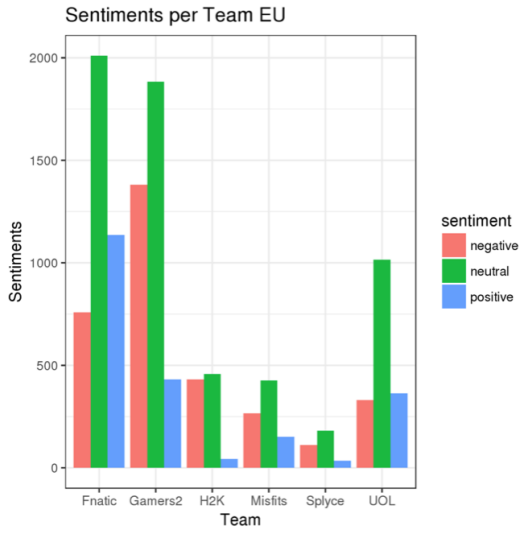

Figure 8: Opinion on EULCS teams

While the overall frequencies of mentions are the same as in the simple popularity count, the proportions of positive to negative messages reveal much, particularly in Europe: while champion Gamers2 has far more negative mentions than positive ones, third place and apparent fan favourite Fnatic has the most positive balance of all teams - in spite of their performance. This adds a new range of considerations to be made: the number one team in the European LCS has the worst opinion evaluation of all participants, while several teams which placed further down the ranks are spoken of in more positive terms. Once again, this initial analysis is not enough to determine the cause of these observed phenomena.

5 Conclusion

Addressing the expressed objective of being an initial investigation on the possibility of employing Text Mining concepts to *eSports* chat, the conducted project was able to identify more mentioned teams and players, as well as measure the opinion of viewers on each team in satisfactory fashion, resulting in rough indicators of popularity and fan loyalty, and a grasp on the dynamics of opinion expression in these environments. It was also capable of demonstrating the potential of distributed and scalable architectures for this kind of endeavor.

As expected, the methods employed were rather simplistic, and not able to provide more than initial estimates. The opinion measuring technique, in particular, can be vastly enhanced through the use of tools and procedures from Natural Language Processing, perhaps going as far as identifying reasons for positive or negative opinions of certain teams or players. Difficulties could arise, however, when trying to approach the extremely informal syntactic structure of most messages.

Additional information could be acquired by also capturing and analyzing time data: by keeping the time stamp of each message and then comparing to a time line of events from the actual content being broadcast, it might be possible to find causal relations to trends in chat, such as spikes in popularity of players when they make impressive plays. One could also try to identify trend starters by keeping meta data relating to the user names of viewers posting messages and looking for patterns of opinion changes based on which users posted to the chat.

References

[1] Apache Software Foundation. Apache Haoop. Available in: https://hadoop.apache.org/

[2] K. Beck. Counter-strike major breaks twitch record with over 1 million concurrent viewers. Mashable, 2017.

[3] A. Gandormi and M. Haider. Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management*, 35:137-144, April 2015.

[4] J. Hamari and M. Sjoblom. What is eSports and why do people watch it?. *Internet Research*, 27(2):211-232, 2017.

[5] M. Hilbert. Big data for development: A review of promises and challenges. *Development Policy Review*, 34:135-174, 2016.

[6] Twitch Interactive Inc. Twitch Developer API - Chat and IRC. Available in: https://dev.twitch.tv/docs/v5/guides/irc/

[7] Newzoo. Most Watched Games on Twitch. Available in: https://newzoo.com/insights/rankings/top-games-twitch/

[8] A. Paradise. The importance of streaming to e-sports. 2017.

[9] Riot Games. LoL eSports. Available in: http://www.lolesports.com

[10] E. J. Schultz. Are you game? Brands are discobering loyal fanbase for esports but marketers need to play by the rules. *Advertising Age*, 88:12, April 2017.

[11] Y. Seo and S. Jung. Beyond solitary play in computer games: The social practices of esports. *Journal of Consumer Culture*, 16:635-655, OCtober 2014.

[12] J. Silge and D. Robinson. *Text Mining with R*. O'Reilly, 2017.

[13] P. Tassi. Monstrous vieweship numbers show League of Legends is still esports king. Forbes. Available in: https://www.forbes.com/sites/insertcoin/2015/12/11/monstrous-viewership-numbers-show-league-of-legends-is-still-esports-king

[14] T. L. Taylor. *Raising the Stakes: E-Sports and the Professionalization of Computer Gaming*. The MIT Press, 2012.