EMOTIONS DETECTION IN SOCIAL MEDIA POSTS

Pedro Lima Rodrigues (Universidade de São Paulo)

Renato de Oliveira Moraes (Universidade de São Paulo)

Hugo Watanuki (LexisNexis Risk Solutions)

David de Hilster (LexisNexis Risk Solutions)

With the proliferation of the internet, the number of users expressing their opinions and perceptions about different entities on social media has grown. So far, research on identifying emotions in text using natural language processing has mainly focused on the English language, using Machine Learning approaches. This results in a limited understanding of the phenomenon, especially in other languages such as Brazilian Portuguese.

This study aims to detect emotions in Twitter posts related to a Brazilian soccer team using a novel natural language processing approach. To this end, a test corpus composed by tweets were studied in detail to create simple rules for how a human would perform this task and to create a taxonomy of emotions specific to soccer fans.

This logic was implemented in NLP++, a computer language specifically designed for encoding the way humans process text that differs from common machine learning techniques and requires only a small number of examples to be implemented.

As an output, a graph was generated showing the percentages of each emotion at each moment of the soccer match using a data visualization library from HPCC Systems.

The results of this study seem to indicate a viable approach to reproduce human thinking in the detection of emotions on a large scale, and NLP++ proved to be a powerful tool for dealing with this challenge in the natural language processing field.

Keywords: big data, natural language processing, emotion analysis, soccer team, social media.

1. Introduction

In recent years, social media has become increasingly present in the daily lives of people around the world. The number of active social media users worldwide has reached 4.7 billion, or 59% of the global population (KEMP, 2022). These people spend an average of 2 hours and 29 minutes per day using social media.

In Brazil, this reality is no different. There are over 171.5 million social media users, which represents 79.9% of the Brazilian population (KEMP, 2022). Brazilians spend an average of 3 hours and 49 minutes per day browsing social media, making it the second country with the highest daily time average.

According to Kemp (2022), the social media platforms with the most users in Brazil are YouTube (138 million), Instagram (122.4 million), Facebook (116.6 million), TikTok (74 million), LinkedIn (54 million), and Twitter (20.5 million).

Through these social media platforms, users express their opinions, perceptions, and emotions about various entities such as products, brands, people, and soccer teams.

Although often used interchangeably, it is worth noting that the terms "emotion" and "sentiment" have different conceptualizations. According to Gonsalves (2020), "emotion" refers to a reaction to a stimulus, which is momentary and does not involve the filter of thought. The "sentiment" concept involves a cognitive component, meaning it is a thought action that gradually constructs a feeling. Therefore, usually what can be detected in a short post in social media is an emotion, not a sentiment.

Emotion analysis has shown to have several interesting applications, such as in companies to monitor consumer perception of a product, in the government to detect social uprisings, and even in politics, where it can be used to predict the outcome of presidential elections (EMOTIVE, 2022). The EMOTIVE project, developed at Loughborough University, was used in the UK and US presidential elections and was able to correctly predict their results (LOUGHBOROUGH UNIVERSITY, 2016).

Given this context, different artificial intelligence models using various techniques have been studied to monitor emotions in social media, such as Support Vector Machine (SVM), Decision Tree, Random Forest, Neural Networks, and Naïve Bayes. Research on emotion identification in texts has mainly focused on the English language (DOSCIATTI et al., 2013). In addition, to train their models, these supervised machine learning techniques depend on large databases with labelled data and therefore require intensive human labor to label the data.

In this study, the NLP++ computational language was used. The NLP++ language and its IDE, VisualText, were developed to build computer programs that are "Digital Human Readers". This language uses syntactic matching of pattern in the text, linguistic and world knowledge, and the ability to create knowledge on-the-fly to perform a task. With a small amount of text, a human can encode the linguistic and world knowledge required to perform a task. If a problem arises, it is possible to explain why the failure occurred and correct it since 100% of the code is visible ("glass box"), and the system can be easily improved over time (DE HILSTER, 2022). This traceability and explanation capability of results is an especially critical feature for highly regulated sectors such as the financial and insurance markets. This aspect, in particular, represents an important advantage of the technique explored in this study compared to traditional machine learning approaches.

Finally, NLP++, besides having already been successfully applied to emotion analysis involving the NASDAQ stock market, also allows scalability to process large volumes of text through an NLP++ plugin available on the HPCC Systems supercomputing platform (HPCC SYSTEMS, 2022). The HPCC Systems is a completely free and open-source big data processing and analysis platform created and maintained by Lexis Nexis Risk Solutions (LNRS), a global data and analytics company.

2. Objectives

The overall objective of this study is to develop an approach to identify the emotion contained in social media posts written in Brazilian Portuguese by using human encoded linguistic and world knowledge. To this end, Twitter posts about a Brazilian soccer team were processed using NLP++ and HPCC Systems technologies.

Despite not having as many users as other social media platforms, Twitter was used in this study because the context and the format of the post utilized in this platform, namely tweet, better match the type of emotion that the study aims at focusing. Twitter is mainly used as a second screen, allowing users to comment and debate what they are watching on TV, such as news programs, reality shows, and soccer games. Plus, Twitter also has some operational advantages in comparison to other platforms, such as ease of access to their historical tweet database, making it an attractive source of data for studies of this nature.

In parallel, soccer is a national passion in Brazil that stimulates intense emotions in fans, who express themselves in various ways, moments, and channels. Among the various Brazilian

soccer teams, the Sociedade Esportiva Palmeiras was chosen to facilitate the extraction of tweets, as the team's name would not be easily confused with other entities.

3. Methodology

This paper used a positivist approach to construct a model to identify the emotions on tweets. The model generated uses a set of rules trying to represent the way a human brain would identify the emotions on tweets. The model development has followed four steps:

a) Data collection from the social media. The Project use a service available on internet to collect the tweets about Palmeiras soccer team on Twitter;
b) Adoption of an emotion taxonomy. A first analysis of the tweets led to the creation of a taxonomy of emotions on this kind of message;
c) Development of the emotion detection model. With the two elements above – data collection and emotions taxonomy – a model to classify the emotions was created;
d) Model evaluation. With the data collected during few soccer matches and processed by the model, the observed results were evaluated.

Actually, the model built is a prototype that was supposed to operate in a context with massive volumes of real-time data and to be able to incorporate improvements without using large training databases with labelled examples. These were the reasons for the technological choices of the project.

The HPCC Systems platform is a parallel processing environment for massive volumes of data, and NLP++ is a computational language that may represent the rules of human expertise.

4. The emotion detect model

This section details how the model was developed. It describes some important decisions made during the model building.

4.1. Capturing the test corpus

The first task performed in the project was to capture tweets related to Palmeiras. Twitter itself offers the API tool for developers interested in using their data for academic purposes, called Twitter API v2 (TWITTER, 2023). After registering on the platform, a Bearer Token is provided, which is important to authorize data extraction.

For an initial analysis of tweets, 100 tweets were captured every 15 minutes from the start to the end (2 hours after the start) from random Palmeiras soccer matches, since the number of

tweets generated during the match is significantly higher than at other moments. For each game, nine extractions of 100 tweets each were performed.

The output file type of the API is a .json file, however, the NLP++ IDE requires files in .txt format for testing and improving the analyzer. Therefore, each tweet from the .json file was transformed into a .txt file using Python code.

4.2. Detailed analysis

The collected tweets were meticulously analyzed. In this stage, it was important to create a taxonomy for emotions. The use of a taxonomy in the study is important, as it better characterizes the emotions being considered and delimits the scope of the classifier. There are some well-known taxonomies such as Ekman (1971), Plutchik (1980), Drummond (2004), and Izard (2009), however, since the context of this study is very specific, a taxonomy for football fans has appeared to be relevant. The emotions that were most frequently present in the tweets composed this taxonomy, which contains the following emotions: support, anger, happiness, and funny.

Given the nature of NLP++, the next step was to think about how a human performs the task of classifying each text into these emotions. At this point, some simple rules were created by observing the patterns contained in the tweets.

Among these stipulated rules, some examples are highlighted next.

4.2.1. Funny

There are some specific expressions to represent a laugh, such as "kkkkk", "hahaha", "ksksks", among others. The program must be able to detect in a tweet if there is any type of laughter and, if so, classify it as "funny".

In the specific case of Palmeiras, any mention to the fact that the team has not yet won a world championship can also be classified as "funny".

4.2.2. Anger

Although some swear expressions can be used in other contexts, such as celebrations and cheers, in general, they can be classified as "anger".

In addition, when the tweet mentions the nouns "VAR", "arbitragem" (referee), "juiz" (judge), or similar ones, it can be considered a complain about some attitude of the refereeing and feeling prejudiced, which is configured as "anger". The same applies to verbs such as "roubar" (steal) and "assaltar" (rob).

4.2.3. Support

Usually, when intending to encourage and cheer for their team, users use various variations of "vamos", "bora", and "vamo" (let's go, come on), including repetition of the last letters, such as "vamooooo". Other constructions also fit into this situation, such as "pra cima" (go for it) and "faz mais um" (score another one).

Also, in the specific case of Palmeiras, fans usually use the words "time do amor" (team of love) and "time da virada" (team of comeback) to encourage the team.

4.2.4. Happiness

"Happiness" manifests itself in moments when the team is doing well in the match or when the team scores a goal. Therefore, the word "GOL" (GOAL) and all its variations fit into this emotion.

4.3. Implementation in NLP++

After this detailed analysis, the logic was implemented in NLP++. Some code snippets and outputs are detailed below.

4.3.1. Repetition of letters

A very common linguistic construction in the tweets is the repetition of letters in words. Therefore, it was necessary to create a way for the code to identify these words without having to list all possible combinations. For this, functions were created that, based on two arguments (the word to be analyzed and the benchmark word), return whether the word in question is equivalent to the benchmark or not.

The LetrasRep function is used when the last letters of the word are repeated.

4.3.2. Team name

To refer to the Palmeiras team, supporters can use several forms and all of them should be understood by the analyzer. For example, an article, such as "the", a possessive pronoun, such as "meu" (my) or "nosso" (our), and various nouns, such as "palmeiras", "verdão", "porco" or "palestra". Therefore, the code in Appendix A shows that the article and possessive pronoun are optional.

4.3.3. Cheering with "vamos" (let's go, come on)

The logic consisted of classifying as "support" when the word "vamos" (or its variations) is followed by a mention of the team. In addition, there may be the word "ganhar" and a comma between these elements. Appendix B shows the rule.

4.3.4. Insult

A similar logic was used with insults. The tweet can only be classified as "anger" if the insult is followed by a mention of the Palmeiras team. The comma between the insult and the team name is optional, as shown in Appendix C.

4.3.5. Use of the verbs "roubar" or "assaltar"

There are several ways to combine these verbs with the noun, so the code must encompass as many possible situations as possible. An anger verb with a mention of the team (with an optional adverb) sets the "anger" emotion. Another way to express it is by using the team's name first and then an anger adjective, such as "assaltado" or "roubado", which may or may not use a linking verb between them, such as "foi" or "é".

The identification of emotions is performed via a knowledge base, which is completed with each tweet and then deleted. Each time a linguistic construction that characterizes an emotion appears in the text, a point is added to the corresponding emotion. When the analysis of a tweet is completed, the count of emotions is reset and the analysis of the next tweet begins.

The complete code is available in the project's GitHub repository (RODRIGUES, 2023).

4.4. Tweet processing

The number of tweets generated during a soccer match is high, and therefore, it is important to use an adequate tool for processing massive volumes of data, i.e., a supercomputing platform. HPCC Systems was developed with the big data paradigm in mind and uses parallel and distributed processing. The NLP++ plugin was used to call the analyzer in HPCC Systems cluster and to process a large number of tweets in real-time. The ECL code was divided into 3 stages (Extraction, Transformation, and Delivery), as outlined next.

4.4.1. Extraction

The first step consisted of merging all nine Twitter files, containing 100 tweets each, into a single dataset. Next, only the core data required for the analysis were extracted from the raw data, such as the tweet id, timestamp, and text.

4.4.2. Transformation

Once the proper raw data was extracted, the second step was to finally call the NLP++ analyzer to process all the tweet texts in the dataset. At this point, a nested recordset was created with the tweets posting time in ascending order, as shown in Table 1.

Table 1 – Emotions detected in each tweet

Hour	ID	Text	Emotion	Counter
1600	001	VAMOS PRA CIMA, PORCO! 🐷 #AvantiPalestra	support	2
1600	002	@Palmeiras VAMOOOOOOOOOO	support	1
1600	003	Dudu, meu orgulho 😍😂😂💙	support	1
			funny	2
1600	004	VAMOS GANHAAAR PORCOOO!!!! 💙🐷💙🐷💙	support	5
1600		São Paulo jogando certinho, assim que tem que jogar contra	anger	1
	005	o palmeiras, mais tem que fazer a porra do gol né Luciano		
1600	006	palmeiras pelo amor de deus	support	1

Source: Authors

However, for each tweet, there should only be one predominant emotion. Therefore, within each row, emotions were ranked from the highest to the lowest count, and only the emotion with the highest count was kept. Finally, a recordset was created where the four different emotions are listed in the columns and the nine time periods are listed in the rows, and the recordset is populated with the total number of tweets that were classified in each emotion at that moment in time, as shown in Table 2.

Table 2 – Number of tweets classified in each hour

Hour	Anger	Support	Happiness	Funny
1900	28	32	0	0
1915	34	4	0	0
1930	30	12	1	0
1945	47	5	1	0
2000	33	4	0	0
2015	46	0	0	0
2030	25	2	0	0
2045	41	0	0	0
2100	40	2	0	0

Source: Authors

The last step was to manipulate this recordset so that the information could be outputted in a graphically and proper manner. To do this, a normalization was performed on the number of occurrences of each emotion, transforming all values into percentages.

4.5. Delivery

The output of the project was a graph with the percentage of each emotion at each moment of the soccer match. To generate this graph, a data visualization library from HPCC Systems called Visualizer was used. Two line charts were generated for each game, one showing the normalized values (in percentage) and the other without normalization, to give an idea of how many tweets, among the 100 tweets of each time period, had some emotion detected.

5. Results

The ideal way to work with NLP++ language is to write the code based on how humans perform the task, test it on a test corpus (set of tweets), and improve the code based on the outputs given by the analyzer. Nevertheless, for one soccer match, a comparison was made between the expected emotion (according to a human) and the emotion classified by the analyzer in the soccer match that occurred on October 16th, 2022. Table 3 shows the overall results, and Tables 4, 5, 6, and 7 break down the results for each of the emotions.

Based on this data, the accuracy was calculated, which represents how many tweets were correctly classified for each emotion among all the tweets. The accuracy of the "support" emotion was 93%, "funny" was 97%, "happiness" was 99%, and "anger" was 74%.

Table 3 – General confusion matrix

		Classified				
		Support	**Funny**	**Happiness**	**Anger**	**None**
Expected	**Support**	64	1	0	1	32
	Funny	0	44	0	6	7
	Happiness	1	0	0	0	1
	Anger	7	7	2	76	162
	None	16	8	6	44	392

Source: Authors

Table 4 – Confusion matrix for "support" emotion

		Classified	
		Yes	**No**
Expected	**Yes**	64	34
	No	24	755

Source: Authors

Table 5 – Confusion matrix for "funny" emotion

		Classified	
		Yes	**No**
Expected	**Yes**	44	13
	No	16	804

Source: Authors

Table 6 – Confusion matrix for "happiness" emotion

		Classified	
		Yes	**No**
Expected	**Yes**	0	2
	No	8	867

Source: Authors

Table 7 – Confusion matrix for "anger" emotion

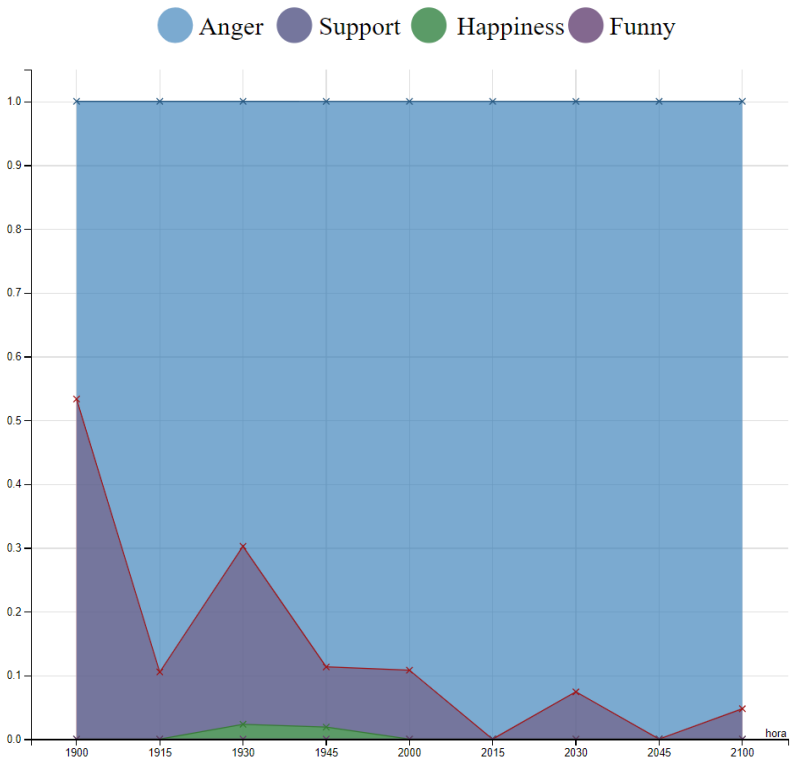
		Classified	
		Yes	**No**
Expected	**Yes**	76	178
	No	51	572

Source: Authors

In addition, for each soccer match whose tweets were collected, two graphs were generated to illustrate the dynamic of the emotions over time. In this report, Graphs 1 and 2 from the soccer match that occurred on January 14th, 2023, are presented.

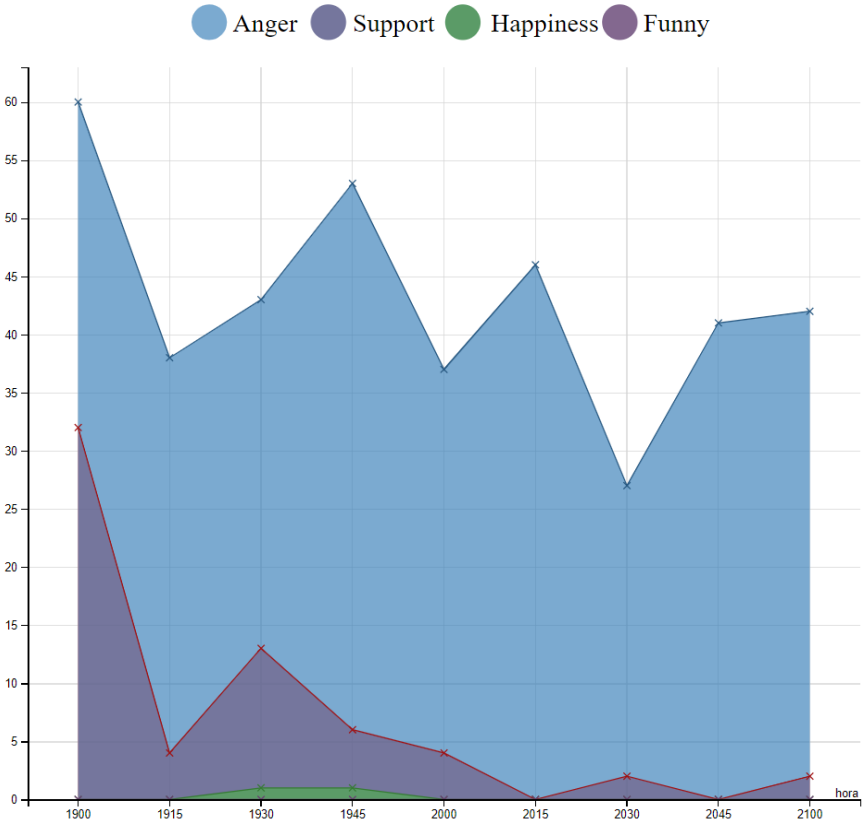
One of the goals of the graphical analysis was to understand if the results shown in the graph are associated with the events that occur during the soccer match. To verify this, the important moments in the match, such as goals, penalties, and cards, were marked on the graphs. Graph 3 is from the soccer match on October 3rd, 2022, and the red line shows the moment when Palmeiras received a red card. Graph 4 is from the match on January 19th, 2023, and the green line marks the moment when Palmeiras scored a goal.

Graph 1 – Variation of emotions throughout the match (percentage)



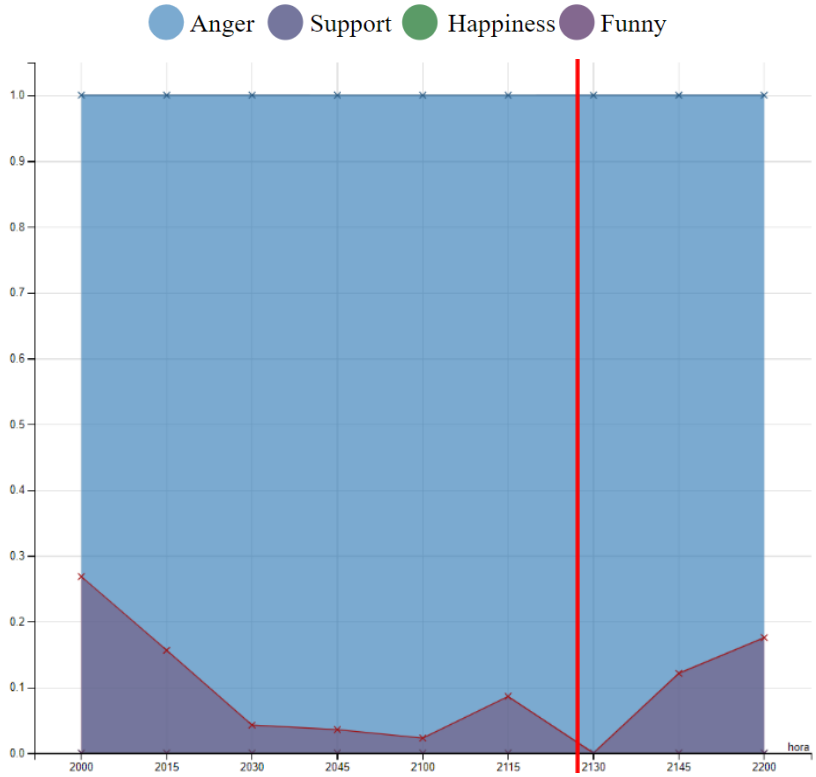
Source: HPCC Systems

Graph 2 – Variation of emotions throughout the match (absolute number of tweets)



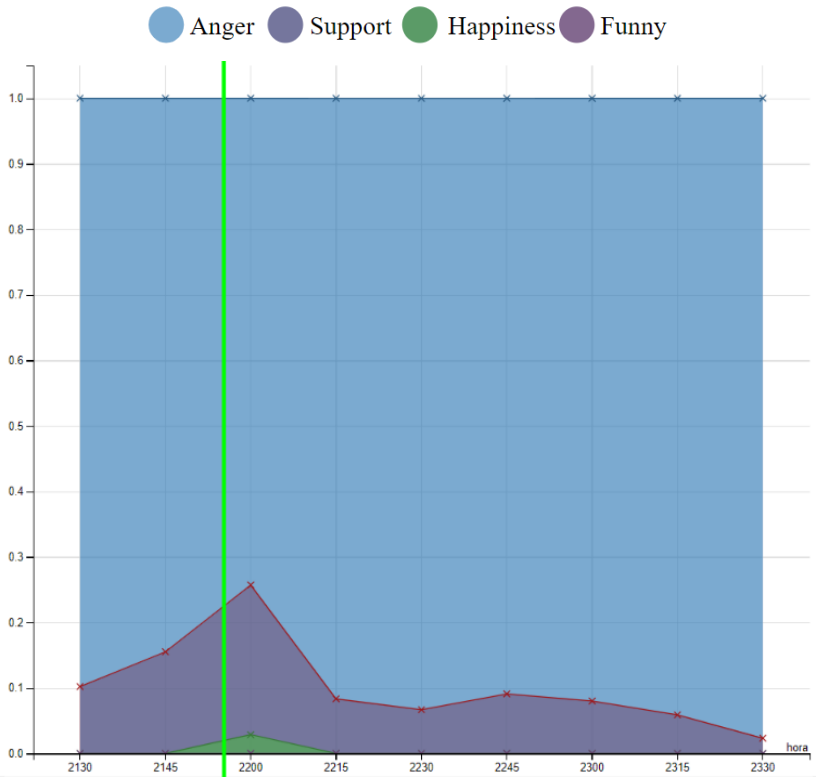
Source: HPCC Systems

Graph 3 - Variation of emotions in the match on October 3rd, 2022



Source: HPCC Systems (Note: the red line shows the moment when Palmeiras received a red card)

Graph 4 - Variation of emotions in the match on January 19th, 2023



Source: HPCC Systems (Note: the green line marks the moment when Palmeiras scored a goal)

6. Conclusions

Firstly, in Table 2, there is a large number of tweets in the "none" row and column. These tweets were not classified into any emotions for many reasons. Most of them are informative texts, therefore, they do not represent an emotion, but rather state a fact that occurred in the game.

Overall, the accuracy values are good, indicating that the analyzer had a good performance. The only caveat is in relation to the "anger" emotion. In this case, it is interesting to analyze the recall metric, which is calculated by dividing the number of expected tweets classified as "anger" by the total number of expected tweets as "anger". The recall is 30%, which is a low value and indicates that the analyzer is not performing well in correctly classifying tweets with "anger", thus making mistakes most of the time.

This happens because there are many ways to show anger through text, therefore, the challenge of encompassing all possible forms of anger expressions in the code increases. This issue could be improved by enhancing the terms and linguistic constructions that characterize the "anger" emotion.

Through the analysis of the graphs generated from the collected tweets, a predominance of the "anger" emotion is noted, while the "funny" and "happiness" emotions were less present. In other words, fans of the team express much more "anger" in their posts than other emotions.

In most soccer matches analyzed, the "support" emotion is very present at the beginning of the match, which can be considered reasonable since at the beginning of the match fans are encouraging and supporting their team. In some cases, it was noted that variations in emotions are related to the events throughout the soccer match, but in many cases, this does not happen. This may be related to the interval between tweet collections, which is done every 15 minutes, and the timespan between the event and the fans' reaction.

For example, if there is a goal in minute 15 of the match, fans will take some moments to write tweets of happiness celebrating the goal, and consequently, the increase in the "happiness" emotion cannot be perceived by a tweet extraction performed exactly at the minute 15. The increase in happiness cannot be perceived either by a tweet extraction performed at the minute 30, given that there will be no more fans celebrating a goal that happened 15 minutes ago. This limitation could potentially be minimized by reducing the interval for data extraction, as long as the social media platform supports a real time data extraction.

One of the contributions of the project with NLP++ was the creation of functions involving words with repeated letters. These functions are essential because they avoid to list all possible combinations of letters that result in laughter or a celebration shout, for example.

It is worth noting that the study focused on a specific context, posts about a soccer team and therefore utilizes specific terms related to this context in its code. Focusing on a specific context or entity is a common approach in natural language processing scenarios, so although the analyzer developed in this study may not be suited to analyze all possible texts in Brazilian Portuguese, it is important to highlight that it could be used to detect emotions in posts about other sports teams as well.

Nevertheless, the same approach can be extended to other topics, suggesting that the findings of this study can contribute to the research effort of reproducing human thinking in emotion detection on a larger scale.

Finally, the NLP++ computational language proved to be a powerful tool for dealing with the challenges of natural language processing and emotion analysis. This characteristic is due to the fact that humans can encode their linguistic and world knowledge to perform a task without depending on a large amount of training data.

REFERENCES

DE HILSTER, David. Understanding Natural Language. HPCC Systems, c2023. Available in: <https://hpccsystems.com/blog/Understanding-Natural-Language>. Accessed on: September 02, 2022.

DOSCIATTI, Mariza; FERREIRA, Lohann; PARAISO, Emerson. Identificando Emoções em Textos em Português do Brasil usando Máquina de Vetores de Suporte em Solução Multiclasse. Oct., 2013.

DRUMMOND, Tom. Vocabulary of Emotions [Online]. **North Seattle Community College**, 2004.

EKMAN, Paul. All emotions are basic. **The nature of emotion: Fundamental questions**, 1994.

GONSALVES, Elisa. Emoção x Sentimento. **Núcleo de Educação Emocional**, 2020. Available in: <http://www.ce.ufpb.br/neemoc/contents/videos/emocao-x-sentimento#:~:text=A%20emo%C3%A7%C3%A3o%20%C3%A9%20uma%20rea%C3%A7%C3%A3o,enquanto%20que%20sentimento%20%C3%A9%20constru%C3%A7%C3%A3o.>. Accessed on: April 23, 2023.

HPCC SYSTEMS. Taming the Data Lake: The HPCC Systems® Open Source Big Data Platform. **HPCC Systems**, c2022. Available in: <https://cdn.hpccsystems.com/whitepapers/wp_introduction_HPCC.pdf.>. Accessed on: April 23, 2023.

IZARD, Carroll. Emotion theory and research: Highlights, unanswered questions, and emerging issues. **Annual Review of Psychology**, 2009.

KEMP, Simon. Digital 2022: Brazil. **Datareportal**, 2022. Available in: <https://datareportal.com/reports/digital-2022-brazil.>. Accessed on: April 23, 2023.

KEMP, Simon. Digital 2022: July Global Statshot Report. **Datareportal**, 2022. Available in: <https://datareportal.com/reports/digital-2022-july-global-statshot.>. Accessed on: April 23, 2023.

LOUGHBOROUGH UNIVERSITY. Emotive Systems, 2022. Página inicial. Available in: <https://emotive.systems/>. Accessed on: April 04, 2023.

LOUGHBOROUGH UNIVERSITY. Loughborough research explains emotions behind the US presidential election. **Loughborough University London**, 2016. Available in: <https://www.lborolondon.ac.uk/news-events/news/2016/presidential-election-emotions/>. Accessed on: April 04, 2023.

PLUTCHIK, Robert. Emotion: A Psychoevolutionary Synthesis. **Longman Higher Education**, 1980.

RODRIGUES, Pedro. EmotionsDetection Repository. GitHub, c2023. Available in: <https://github.com/pedro-lima-rodrigues/EmotionsDetection>. Accessed on: May 03, 2023.

TWITTER. Developer Platform, c2023. Documentation. Available in: <https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction>. Accessed on: September 08, 2022.

APPENDIX

Appendix A - Rule for the team's name

```
_frasePalmeiras <-
        _artigo [s optional]                                    ### (1)
        _pronomePossessivo [s optional]                         ### (2)
        _palmeiras                                              ### (3)
```

Appendix B - Rule for interaction with "vamos"

```
_emocao <-
        _vamos                                                  ### (1)
        _xWILD [s optional one matches=(ganhar ganha virar vira)]  ### (2)
        \, [s optional]                                         ### (3)
        _frasePalmeiras                                         ### (4)
```

Appendix C - Rule for insult

```
_emocao <-
        _xingamento             ### (1)
        \, [optional]           ### (2)
        _frasePalmeiras         ### (3)
```