Multimedia Tools and Applications (2023) 82:12957–12976
https://doi.org/10.1007/s11042-022-14070-w

1224: NEW FRONTIERS IN MULTIMEDIA-BASED AND MULTIMODAL HCI

An opinion mining methodology to analyse games for health

Paula Alexandra Silva[1] · **Renato Santos**[1]

Received: 11 January 2022 / Revised: 26 September 2022 / Accepted: 6 October 2022 /
Published online: 5 November 2022

Abstract

Despite the positive impact of games for health on players' health, users tend to stop playing them after a short period of time, leading benefits to fade. It is therefore important to understand how to sustain interest and, in this way, preserve the health benefits of games for health. This could be achieved by continuously reviewing user feedback after product launch and using this information to inform (re)design and better address user needs. With the growth of social media, user opinions became widely available in public forums. This abundance of information affords us the possibility of, through the application of natural language processing and sentiment analysis techniques, tapping into user opinions and automatically analysing and extracting knowledge from them. This paper introduces a methodology that analyses user comments posted on YouTube about the Just Dance game, to automatically extract information about Usability, User Experience (UX), and Perceived Health Impacts related to Quality of Life (H-QoL). In doing so, the methodology uses a pre-established vocabulary, based on the English lexicon and its semantic relations, to annotate the presence of 38 concepts (five of Usability, 18 of UX, and 15 of H-QoL) and to analyse sentiment. The results of the information extraction and processing are displayed on a dashboard that allows for the exploration and browsing of the results, which can be useful to better understand the opinions and impacts perceived by users and to inform the (re)design of games for health. The methodology proposed builds upon over 500,000 user comments collected from over 32,000 videos.

Keywords Natural language processing · Opinion mining · Usability · User experience · Health · Quality of life

Paula Alexandra Silva and Renato Santos contributed equally to this work.

✉ Paula Alexandra Silva
paulasilva@dei.uc.pt

Renato Santos
renatojms@student.dei.uc.pt

[1] Department of Informatics Engineering, University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Coimbra, Portugal

1 Introduction

In recent years, serious games have attracted increasing attention and shown potential in promoting health, however research has shown that those games retain solely a short-term engagement [40]. It is therefore relevant to investigate how interest can be sustained, so that, in tandem, the health benefits that games for health can afford are preserved. A possible starting point could be to learn from successful commercial games, in particular those that have shown to produce a positive impact on players' health.

For a digital solution to be successful, it has to fit the specific needs of its users at the time of its construction and evolve to fit their changing needs later, which requires further refinements and iteration, for instance, when the solution is used in realistic conditions [13]. It is then important to identify ways in which user feedback could be continuously monitored beyond product launch. One such way could be to collect and analyse user comments from public fora, so that the knowledge and insights that can be extracted from those comments can be used to inform refinements.

In recent years, social media have made a huge source of free unlimited information available. This information can be used to explore user perceptions and inform future decisions [10, 21, 47, 53, 59], however, interpreting information from social media is not trivial [28, 30]. In addition, if we are to contribute to the increase in adherence to games for health, it is important to gauge user perceptions not only in terms of usability and user experience (UX), but also in terms of health-related aspects.

Recent approaches have shown that, in applying opinion mining techniques, it is possible to take stock from user comments posted in public fora to extract knowledge about usability and UX [14, 26, 62]. Other studies have applied similar approaches successfully in the domain of health [39]. However, to our knowledge, no approaches are available that combine both perspectives, nor that specifically address aspects related to health.

To contribute to resolving the challenges outlined above, this research poses the following research question: Can user comments from YouTube about Just Dance provide relevant insight into users' perceptions of the game's usability, UX, and health impacts? In addressing its research question, this research leverages a set of open source tools and proposes a social media opinion mining methodology that combines a usability and UX vocabulary [14] with a Health-related Quality of Life (H-QoL) vocabulary [45]. Since no curated dataset exists to address our research question, we planned the methodology accordingly, i.e., data collection, annotation, curation, and assessment.

To gauge the usefulness of the approach and vocabulary, the work analyses user comments retrieved from YouTube about Just Dance, a game that has shown to be beneficial for its users in promoting physical activity education, motivation to exercise, physical and psychological health as well as users social life [22]. In this way, albeit not intrinsically designed as a pure exergame, Just Dance holds the potential benefits of a health game. In addition, the amount of data (i.e.: user comments) that are publicly available about Just Dance makes it a good case study, upon which a curated data set can be created. Once created, this curated dataset can be made available and further explored, namely to effectively ascertain its usefulness in informing the design and development of games for health.

Making automatically processed information about usability, UX, and H-QoL available to design and development teams could constitute a valuable resource, as, by analysing user comments, teams would be able to tap into user preferences and perceptions and adapt the design to better suit user needs. This knowledge could likewise inform the identification of problems and areas in need of improvement.

After this introduction, this paper reviews related work on the potential of games to promote health and on opinion mining approaches applied to detect usability, UX, and health-related aspects. The paper then describes the methodology followed in developing this work, which is organised in two distinct sections, one presenting the scope of the work and another detailing the implementation approach followed to develop the opinion mining methodology we propose. After describing the implementation approach, the paper presents the results obtained and then proceeds to the discussion of the results and the conclusions.

2 Background

2.1 Serious games and health

With the advent and growth of serious games, one of the sectors which have seen the largest impact is the health sector [41]. According to Wattanasoontorn et al. [54], there are different types of serious games for health, including: serious games for professionals, which focus on the training of health professionals and the acquisition of health and medical skills; serious games for non-professionals, which are aimed at laypersons and at improving their own healthcare; and health and wellness games, which are aimed at the general population and at improving functional health.

One of the things that games for health and general entertainment games have in common is the capacity of immersing the user in the game environment for long periods, to the point that users eventually no longer considers themselves a *patient*, but rather a player, allowing them to fully enjoy the experience of the game while securing benefits for their health [51]. Among these type of games, dance games that require full-body movements have exhibited better results in terms of calorie consumption when compared to other games that only require the movement of the upper parts of the body [36]. Furthermore, dance games have been shown to promote motivation and pleasure [22], while also having positive effects on the social life of their users and on psychological and physical aspects of health, such as resistance, muscle strength, sense of rhythm, body image, sleep quality, and subjective well-being [16, 22, 27].

Despite their benefits, serious games continue to lag in adherence [40] and players' interest decreases over time [8]. Moreover, it is still challenging to quantify the effects of these games on the health and quality of life of their players [15]. It is therefore relevant to understand how to improve the design of games for health and what attracts and motivates players, so that design can increasingly meet the goals of their users. In addition, it is worth investigating whether players' perceived health impacts could be linked to usability and user experience factors, that could influence sustained interest.

2.2 Social media opinion mining and digital products

The subjective impressions of users can be collected in many ways, from conventional methods, such as surveys, questionnaires, field studies and lab studies, to more recent methods, based on user-generated content available in social media platforms [59]. With the advent of social media, approaches have leveraged user opinions published on public fora to gain knowledge about user preferences [14], in areas ranging from product use [58] to health [39], sports clothing [50] and house appliances [53]. Previous research also has shown that the information posted online allows for the identification of the user profiles

and preferences as well as the understanding of the characteristics of the context and product in use [59]. It is also possible to detect software limitations, such as errors and missing functionalities [58] or identify usability heuristic issues [10] and inform game design [47].

Hedegaard and Simonsen [14] analysed user experience reports on software products and games as posted on Amazon product reviews and concluded that 49% of the comments included information relevant to Usability and User Experience (UX) factors. Following the research of Hedegaard and Simonsen [14], who confirmed the usefulness of their approach extracting Usability and UX information, several authors have applied the same vocabulary, namely to extract knowledge about user satisfaction with the characteristics of software products [3] and to measure the impact of star ratings on the evaluations of mobile apps in the App Store [9]. It is then possible that similar approaches can be applied to improve our understanding of user perceptions about games for health.

2.3 Social media opinion mining and biopsychosocial health

The health domain is another in which text mining approaches have been applied. In this context, Prieto et al. [39] sought to detect incidences of health conditions, such as flu, depression, pregnancy, and eating disorders, by analysing user tweets. Hyde et al. [18] surveyed 286 people between 65 and 75 years of age, to develop a measure of quality of life satisfaction needs for early old age; their work covered four ontology-based domains: control, autonomy, pleasure, and self-realisation. Alqahtani and Orji [2] analysed comments on the App Store and Google Play with regards to user perceptions on mobile apps aimed at mental health to explore aspects such as medication, relaxation, education, advice, and social support, among others. Buzzy et al. [6] go as far as to argue that data from social networks can be used to gather health-related data at a lower cost and greater efficiency.

It is also possible to detect sentiment based on opinion and text mining approaches. Tuch et al. [50] analysed the content and structure of 691 user narratives on positive and negative experiences with technology, to describe and model user experiences. In addition to the positive or negative polarity of a comment, other authors [33, 49] also analysed the presence of the eight basic emotions — joy, trust, fear, surprise, sadness, anticipation, anger, and disgust — defined by Plutchik [37] in comments written on social media. Zhu and Fang [61] explored games' online reviews to characterise games and play-ability; their work used a lexical approach adapted from instruments which are generally used by psychologists in studying personality traits. Sirbu et al. [46] used opinion mining to better understand players' feelings from reviews and classify them into positive, negative and neutral reviews.

Regardless of the diversity of approaches elicited in this section and being these approaches based on comments from Twitter (e.g., [39]), Amazon (e.g., [14, 19]) or YouTube (e.g., [42]), we were not able to locate an approach focusing on the analysis of games for health or one that combines the analysis of usability and user experience perspective with health and quality of life aspects.

3 Data source and vocabulary

This study analyses user comments retrieved from YouTube to understand Usability, User Experience (UX), and Perceived Health Impacts related to Quality of Life (H-QoL). To scope the study, the analysis focuses on YouTube videos about a specific exergame: Just Dance. To automatically analyse user comments, this research uses previously validated vocabularies. Both these aspects are detailed next.

3.1 Source of user comments

The Just Dance saga, by Ubisoft, now with 25 editions launched [20], was first launched in 2009 and has since become a world reference in the category of physical exercise and dance videogames [48]. In this game, player movements are tracked in real-time and rewards are given according to the precision of the players' movements, as they follow the dance choreography displayed by an avatar on the screen (Fig. 1). Being one of the most popular dance games in recent years [48], the success of the video game over the years is attributed to the experience the game affords and in which people are continually encouraged to dance [25]. In addition, by making players feel at ease while dancing, both alone or with others, the game improves the sense of rhythm, fun, and well-being [16]. Previous studies have also demonstrated the positive influence of playing the game on the health and motivation to exercise of its players [22]. The popularity of the game as well as its positive impact on the health and well-being of its players [22] makes Just Dance a particularly interesting game to analyse.

To locate data about Just Dance, this study uses YouTube. YouTube is a social network used by billions of users, about 2.24 billion worldwide in 2021 [7]. YouTube content is also mostly open. This makes up for a great abundance of user comments readily available for analysis. Furthermore, comments on YouTube revolve around the theme of the video, creating topical discussions around that theme. Looking at YouTube videos about Just Dance, comments often refer to specific aspects of the game, such as the visual characteristics of the interface or the social context of gameplay. This makes YouTube a particularly interesting source of data for this work. A quick literature search on previous studies on user comments on social networks, also shows that studies that use YouTube as a source of information are less common than those that resort to other sources, such as Twitter or Amazon. This provides additional motivation for the use of YouTube as an alternative source of information for this kind of research.

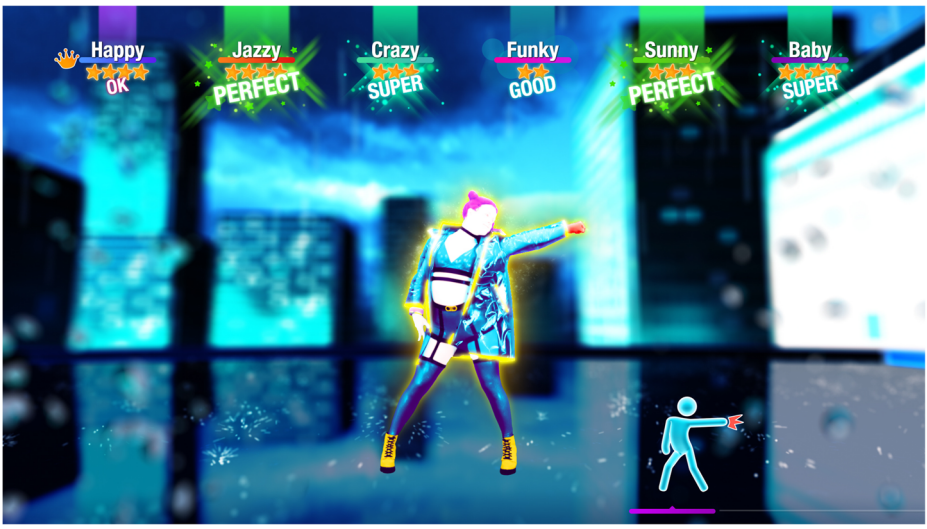

Fig. 1 Just Dance user interface screen [52]

3.2 Dimensions in study

Although previous studies have analysed user comments to extract information about Usability and UX, as far as we are aware, there are still no studies that combine these two dimensions with another that analyses H-QoL or that explores any potential interconnections among these three dimensions. To do so, this work builds upon the work of Hedegaard and Simonsen [14], that identified and validated a Usability and UX vocabulary, and the work of Silva and Santos [45], that proposed a validated H-QoL vocabulary. The vocabularies of these two studies, composed of five Usability concepts, 18 UX concepts, and 15 H-QoL concepts, are used in combination for the work described in this paper. Table 1 lists the 38 concepts of the vocabulary, where the first two columns show the concepts assembled and validated by Hedegaard and Simonsen [14] and the third column lists the 15 concepts proposed by Silva and Santos [45].

4 Implementation approach

This paper contributes an opinion mining methodology and tool that analyses user comments, to automatically extract information about Usability, User Experience (UX), and Perceived Health Impacts related to Quality of Life (H-QoL). This section details the steps of the processing pipeline from the data collection up to the data visualisation on a dashboard created to explore the results (Fig. 2). The rationale and development decisions are described next. Only free and open-source tools were used.

Table 1 Vocabulary in use composed of five usability concepts, 18 UX concepts and 15 H-QoL concepts

Usability	UX	H-QoL
Efficiency	Aesthetics and Appeal	Bodily image and Appearance
Errors/Effectiveness	Affect and Emotion	Concentration
Learnability	Anticipation	Energy
Memorability	Comfort	Fatigue
Satisfaction	Detailed Usability	Learning
	Enchantment	Memory
	Engagement	Negative feelings
	Enjoyment and Fun	Pain and Discomfort
	Frustration	Personal relationships
	Hedonic	Positive feelings
	Impact	Self-esteem
	Likeability	Sexual activity
	Motivation	Sleep and Rest
	Overall Usability	Social Support
	Pleasure	Thinking
	Support	
	Trust	
	User Differences	

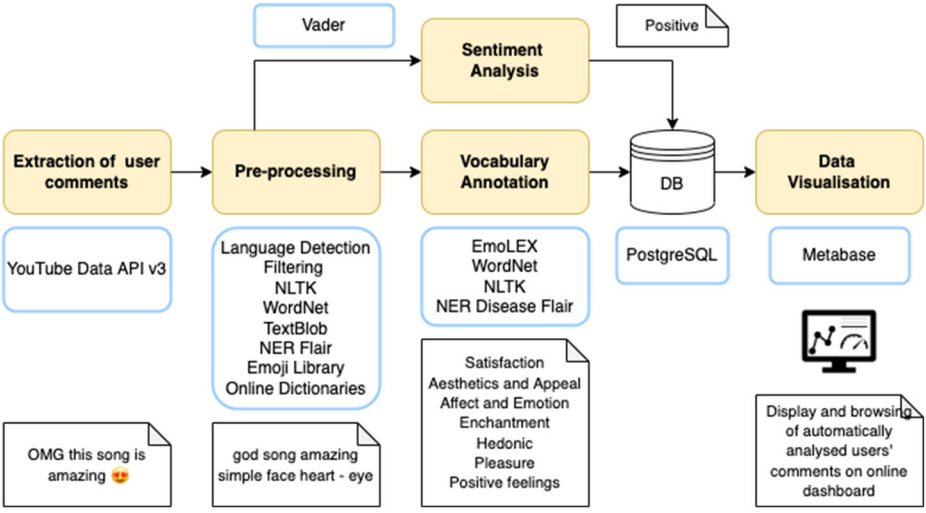

Fig. 2 Overall architecture of the processing tool pipeline from the moment when user comments are extracted from YouTube videos until the moment the results of the automatic processing of comments is displayed in the public interactive dashboard

4.1 Extraction of user comments

As mentioned before, the data set of this work is composed of user comments retrieved from YouTube. In our work, relevant videos were searched and retrieved whenever the title of a YouTube video contained the name of a Just Dance edition [20].

User comments were extracted using the YouTube Data API v3 [60] which sends a request with the search parameters and retrieves the results in JavaScript Object Notation (JSON). The results are retrieved according to the relevance assigned to the videos by the YouTube algorithm, which observes the specified language, the number of views, number of likes, average viewing time, video duration, and the presence of the search string in the title and description of the video. From these results, restrictions were applied to exclude videos which were not about the Just Dance game. For example, Lady Gaga has a song titled Just Dance and there are several YouTube videos featuring this song. Even though these videos were correctly retrieved, they were unrelated to the Just Dance game and therefore irrelevant for our work. For that reason, videos such as those of the Lady Gaga song were discarded. To locate potential irrelevant videos, a manual assessment was carried out to identify subjects, topics, and songs that could contain the words "Just Dance" in the title, but that were not related to the Just Dance game. Once located, sets of keyword-based restrictions were applied to the retrieved videos, to ensure that irrelevant videos were discarded.

The outcome of this process resulted in a set of 32,437 videos and 538,700 comments, collected over 12 years. From the retrieved videos, we extracted the title, description, channel name, publication date, the total number of comments, and the number of likes and dislikes and stored them in a PostgreSQL database (DB). Whereas the video contained a valid comment, the comment was extracted and stored in the DB. For language detection, we used TextBlob [24], a library for textual data processing that allows for linguistic detection and spelling correction, for short comments, and the Language detection library [43] for comments with more than four words.

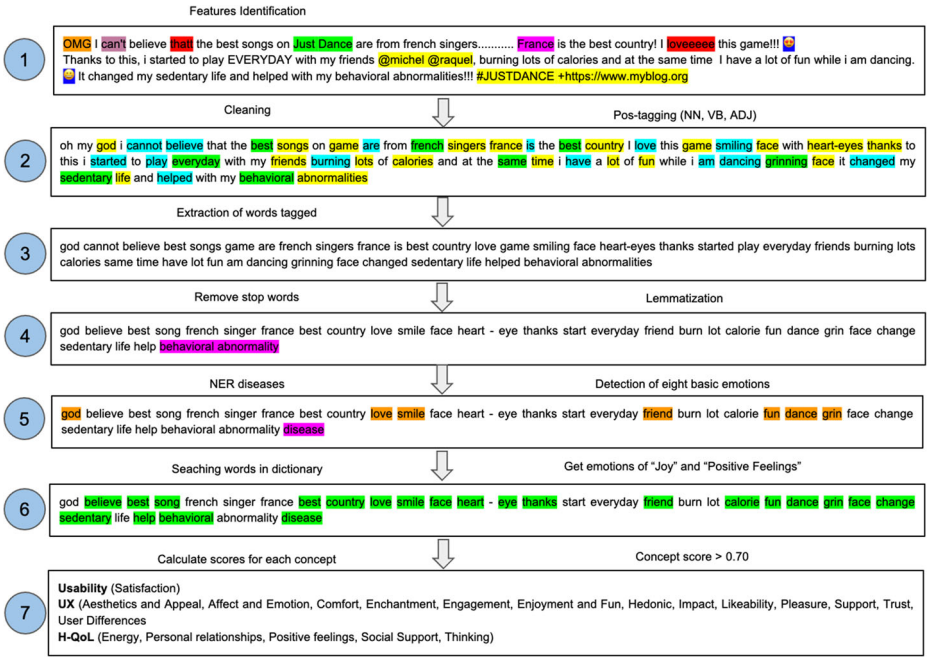

Fig. 3 Example of the processing of a user comment from extraction to annotation

4.2 Pre-processing of data

After obtaining the textual data, the pipeline proceeds with the cleaning and processing of data. For this, we used the Natural Language Toolkit (NLTK) [23], a library designed to work with human language data, with more than 50 corpora and lexical resources, together with a set of word processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. In addition, we used Flair [1] for name entity recognition and to identify diseases, WordNet [32] for concept expansion, and Emolex [33] to detect emotions. In short, the following steps were carried out in the pre-processing stage (for an example, see also Fig. 3, steps 1 to 5):

1. Lowercase transformation.
2. Text expansion of contractions, e.g.,: "I'm" to "i am", and acronyms, e.g.,: "OMG" to "oh my god". Expansion of contractions was made based on a table following [56] and expansion of acronyms was based on online dictionaries [4, 35, 44].
3. Mapping from emojis to Unicode Common Locale Data Repository short name [11].
4. Removal of repeated characters (when repeated three times or more), mentions, hashtags, URLs, punctuation, alphanumeric and special characters, and white spaces.
5. Mapping the words "just dance" to "game", to prevent the word "dance" from being considered as a dance activity in the annotation process.
6. Short comments are excluded, i.e.: those with less than four characters, or not in English.
7. Named-entity recognition (NER), using Flair [1].
8. Spell check, in comments without named entities, as these correspond to names of people, places, institutions, or others and thus should not be modified.

9. Identification and extraction of grammatical classes, namely names, verbs and adjectives [12].
10. Removal of stopwords, and lemmatization of the remaining words of the comment.

Once these steps were concluded, the processed comment was stored in the DB to confirm in subsequent phases whether it has the necessary characteristics to be annotated.

4.3 Sentiment analysis

Sentiment analysis was performed on a comment level and was based on three polarities: positive, neutral or negative. To perform this operation, a combination of lexicon and rules-based approach was applied using VADER [17], a tool that was specially designed to perform sentiment analysis of social media data and, for this reason, works with the raw comments, as displayed on social media.

Each comment entered for analysis is the same as originally displayed on YouTube, after verifying the minimum number of characters (i.e.: 3), the language (i.e.: English). To keep the comment as close as possible to its original form is important to capture subjective details of the user's message, such as expressions in capital letters, or repeated punctuation to emphasise an opinion (e.g. "!!!!!").

After performing the basic pre-processing tasks elicited above, sentiment analysis was performed by VADER applying the following rules:

1. Word negations, e.g., "don't/not/isn't".
2. Repetition of punctuation, e.g., "!!!!!!!!!".
3. Words or phrases in capital letters, e.g., "AMAZING".
4. Changes in the intensity of the sentiment, e.g., "very", "little".
5. Perception of the feeling present in emojis.
6. Perception of the present feeling when using acronyms, e.g., "lol".

Since the processing was performed at the comment level, we did not investigate multiple sentiments in the same post/comment. Instead, we classified the overall comment sentiment according to three general polarities.

4.4 Vocabulary annotation

The annotation process used the previously described vocabulary (Table 1) that includes concepts of Usability, User Experience (UX), and Perceived Health Impacts related to Quality of Life (H-QoL). Some of those concepts link to emotions, therefore the eight basic emotions – joy, trust, fear, surprise, sadness, anticipation, anger, and disgust [37] – were detected using the lexicon provided by Emolex [33]. Additionally, the H-QoL dimension may be related to specific diseases, so the named entity recognition model of Flair [55] was used to detect the presence of diseases.

Building upon the results of the pre-processing phase, the annotation process checked whether the words were included in the pre-established dictionary to map and annotate relevant concepts. In this process, the WordNet knowledge base [32] was used to expand the meaning of words through synonyms and antonyms, the latter being applied only to unambiguous concepts. The search for lexical relations was also expanded, namely by hyponyms, meronyms and hyperonyms, repeating the process. To avoid missing similar concepts when

the words do not match, stemming was performed by resorting to the Snowball algorithm [38]. For that, further to the concepts in the vocabulary, we also built a dictionary of lemmatized words with different weights.

The annotation process proceeded to assign a numeric score to each concept annotation, which ranks its strong/weak presence in the comment. If a concept gathered a high score and the polarity of the comment's sentiment was consistent with the concept's characteristics, that concept was annotated. The final score was calculated for each concept and resulted from the accumulation of the number of words detected in each concept and its respective weight, as assigned in the dictionary. In the score calculation, there is a distinction that depends on the context of the words, in which a greater weight is assigned to the exact match of the words, followed by a lower weight to the match of synonyms and antonyms, and finally, for lexical relations. Figure 3, steps 5 to 7, shows an example of the annotation of one comment.

4.5 Data visualisation

To allow for the browsing and exploration of results, an interactive web-based dashboard[1] was developed. The dashboard is based on the Metabase engine, an open-source tool [31]. The information displayed in the dashboard is retrieved from queries to a DB that follows a "Star Schema" model [34] in PostgreSQL, to ensure higher efficiency in all read operations. That model, does not require complex joins when querying the data, which enables faster access and facilitates the generation of reports [34].

The tool runs on-premises, in a Java Virtual Machine, inside a web application, served by Jetty [29]. The configuration of the Ubuntu Server 20.02 operating system was configured from a Java ARchive (JAR) file, for Metabase version 0.38.2, in an instance of the Java Development Kit. The standard data storage integration was also changed to a new instance of PostgreSQL, serving as a database for Metabase information.

Metabase allows for the definition of automatic cache storage for long query results, where the minimum average query duration and the Time To Live (TTL) cache multiplier can be specified. For this work, the minimum duration of the configured query was set to two seconds, which means that queries that take more than two seconds are kept in the cache. The cache lifetime configured for a TTL multiplier was set to 500, which means that a query that takes 10 seconds will have the result cached for 5000 seconds (about 83 minutes); in this way, the result of the cached query is always proportional to its execution time. It is also possible to set a maximum cache entry size for each query; for this work, this parameter was set to 5MB.

5 Results

This section presents the outcomes of the work elicited in the previous section. Where appropriate it describes, analyses and discusses the results.

[1] https://www.bit.ly/dashboard-just-dance

5.1 Data characteristics and processing

Following the procedures described in Sections 4.1 and 4.2, this work collected about 538,700 comments. Comments were both fetched from main comments and replies to main comments. This being said, most comments come from the main comments, because replies to main comments were rare and, when existent, they were almost always discarded at the pre-processing phase.

The analysis of the characteristics of the comments shows that, even though YouTube allows users to enter comments of up to 9,999 characters, the majority of the comments have a maximum of 100 char / 20 words and an average of 57 char / 11 words. Looking at the characteristics of annotated comments only, on average they are 75 characters and 15 words long.

The processing time of a comment is directly related to the number of words in the comment and the hardware characteristics (a regular laptop in our case). The pre-processing phase of a typical 20-word comment takes an average of 0.7 seconds. The annotation stage of a similar-sized comment takes an average of 2.54 seconds.

5.2 Comments annotation

From the total volume of 538,700 comments, only 277,019 were annotated (52%) after going through the pre-processing phase. Looking at the results of the annotation of these comments, we observe that 39% have been annotated with Usability concepts, 90% with User Experience (UX) concepts, and 76% with concepts of Perceived Health Impacts related to Quality of Life (H-QoL). Figure 4 illustrates the interrelations among the three classification dimensions. The overlapping of comments associated with more than one dimension becomes apparent, namely the one that shows that a portion of the comments marked with UX concepts has also been marked with usability concepts. This was expected, as usability is part of what UX entails. By exploring and analysing the annotated data, namely through the dashboard mentioned in Section 4.5, it is possible to make inferences about the aspects that are (not) working well in the game. However interesting, to examine these possibilities and how they can be exploited by designers and developers is beyond the scope of this paper, which has the methodology as its main focus.

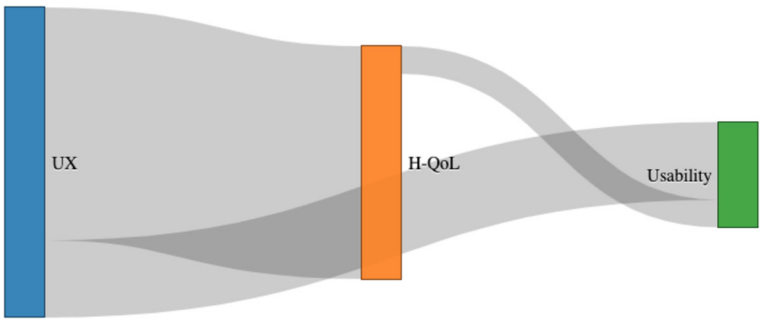

Fig. 4 Sankey diagram with interconnections among the concepts annotations of the Usability, UX and H-QoL dimensions

5.3 Annotation performance

The methodology proposed in this paper combined two vocabularies, which have not been used in combination before and which had not been used to annotate user comments about games, therefore it was important to analyse the annotation performance. In doing so, the researchers manually validated the annotations of 11,261 comments[2], as annotated by the system following the approach we propose in this paper. This was done through a table that listed each of the 38 concepts in different columns and the 11,261 in different rows. The analysis was recorded in the form of 1/0, where "1" confirmed the correct presence and annotation of the concept in the comment and "0" stated that the concept was not present in the comment. The validation was performed by the two authors who collaboratively verified the dataset and recorded the result in the table, solving conflicts as they appear. The annotation precision was calculated as follows:

$$Precision = \frac{\text{True Positive}}{\text{True Positive + False Positive}} = \frac{\text{Correct annotations}}{\text{Total of annotations}}$$

Table 2 shows the results of the manual validation per vocabulary dimension, before and after human intervention.

The overall accuracy of annotation was of 80%. Annotation lacked accuracy in face of user comments that:

- Referred to general aspects of the YouTube channel the comment had been sourced from and not the user experience with Just Dance;
- Mentioned the titles of songs or the names of artists of songs that are used by Just Dance;
- Exhibited irony, sarcasm, and humour; and
- Words in the comments were used in contexts different to the ones anticipated in the lexical dictionary developed.

It is worth mentioning that the concept of sexual activity, which had only been annotated in 161 comments, appears to have little interest, and for this reason, it could be discarded in the future.

In a second phase, the authors selected a random sample of 300 original comments that included both annotated and non-annotated comments to check the overall annotation performance. Both authors made manual annotations of each of the concepts on the 300 comments, which were then compared with the automatic ones. This was again done collaboratively on a shared spreadsheet. Acquired these results, besides precision, we computed the recall and F-1 scores, as follows:

$$Recall = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$F1\text{-}score = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

The results of the global annotation performance, including annotated and non-annotated concepts, are presented in Table 3. Overall the results are aligned with the previous discussion.

[2]Comments were selected from each of the 38 concepts of the vocabulary, from which the researchers randomly picked 300 comments, when available. The concept of sexual activity was only annotated in 161 comments.

Table 2 Precision of the annotated concepts per vocabulary dimension

Dimension	Total of annotations	Correct annotations	Precision
Usability	7,016	4,224	0.60
UX	48,557	41,522	0.86
H-QoL	20,562	15,345	0.75
Overall	**76,135**	**61,091**	**0.80**

5.4 Dashboard

A public interactive dashboard was developed to allow for the browsing and exploration of results. The dashboard displays 24 different charts, that result from different queries to the database. It is also possible to apply filters to the data.

To provide a few examples of what can be browsed and explored in the dashboard, the dashboard displays general information, such as the total number of comments annotated (Fig. 5) and their distribution per game console (Fig. 6), as detected in the title of the video from which annotated comments were sourced.

The dashboard also shows the results of the automatic annotation organised by dimensions (Usability, UX, and H-QoL) and each of the terms of the vocabulary (Fig. 8) as well as the sentiment (positive, neutral, negative) detected in the annotated comments (Fig. 7). It is also possible to filter data, for example, by date. Among others, the dashboard allows the user to select the dimensions to view (e.g., all or any subgroup of H-QoL, Usability, UX), to see the evolution of the concepts along a period, or to compare the different editions of the game. In exploring the dashboard, the users, let's say a designer, developer, or else, can interact with the tool and the information displayed on it, and shape the results in a way that could potentially help them improve or gain a better understanding of user perceptions about the game.

The goal of this paper is not to analyse the specific results obtained about the Just Dance game and what they might mean, however, for illustrative purposes of the kind of plots that can be seen in the dashboard and the insights that can emerge from analysing them, one can observe Fig. 5, where a steep increase in the number of comments annotated is noticeable around the beginning of 2020, when the COVID-19 pandemic was at its peak worldwide. Likewise, Fig. 6 shows the prevalence of the Wii console and Fig. 7 exposes the overall positive sentiment detected in the user comments.

To conclude this set of illustrative examples, Fig. 8 shows annotations according to the 38 concepts in the vocabulary, where Usability is represented in purple, UX in yellow, and H-QoL in green. In Fig. 8, the size of the circles increases with the number of times a concept is annotated. For example, in the plot displayed in Fig. 8, the most annotated concept,

Table 3 Global annotation performance, including annotated and non-annotated concepts

Dimension	Precision	Recall	F1-score
Usability	0.64	0.97	0.77
UX	0.87	0.96	0.91
H-QoL	0.74	0.91	0.82
Overall	**0.82**	**0.95**	**0.88**

Fig. 5 Comments annotated over time

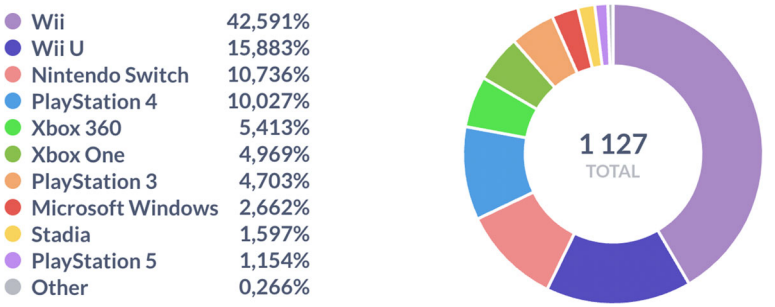

Fig. 6 Platform used to play the game

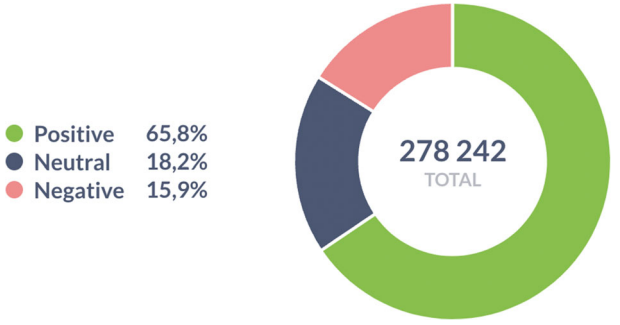

Fig. 7 Sentiment of annotated comments

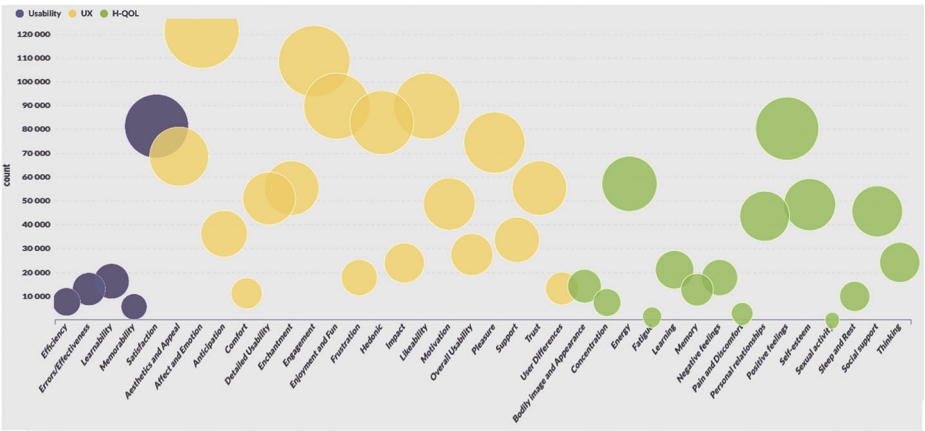

Fig. 8 Distribution of vocabulary concepts of all three dimensions in study (Usability, UX, and H-QoL), where the size of the circle associated with each concept increases as that concept is annotated on the user comments

displayed by the largest yellow circle, is the UX concept Affect and Emotion. Conversely, the smallest circle in green towards the right edge of the image refers to the H-QoL concept Sexual Activity.

In the Usability dimension, the most annotated concept is Satisfaction, while the remaining ones are much less mentioned by users. UX is highly represented by concepts related with Affect and Emotion, Engagement, Enjoyment and fun, and Likeability, while concepts such as Comfort and User Differences are the less detected. In the H-QoL dimension, the most prevalent concepts are related to dance activity and physical exercise, highlighting Positive feelings, Energy, Personal relationships, Self-esteem, and Social support. The least annotated comments are Sexual activity and Fatigue. Overall, considering the three dimensions, the most annotated concepts are Affect and Emotion and Engagement.

6 Discussion and future work

This research explored whether social media, in particular user comments posted on YouTube about Just Dance, contain relevant information about the user-perceived impacts of games for health in three dimensions: Usability, User Experience (UX) and Perceived Health Impacts related to Quality of Life (H-QoL). To scope the study, we focused the data pipeline and analysis on Just Dance. The obtained results confirm that user comments posted on YouTube about Just Dance are indeed valuable and rich enough to assess user sentiment and perceptions about the three dimensions of the study.

Regarding the prevalence of the three dimensions under analysis, UX is the predominant one, reflecting players' emotions and satisfaction while engaged in a game. Usability is the dimension with the smallest number of annotated concepts.

Our proposal of adding a new dimension, H-QoL, to those of Usability and UX, seems relevant, as it is the second dimension with the largest number of annotated concepts. Furthermore, the extracted information now affords us the analysis of specific aspects of games that go beyond entertainment. For the specific case of Just Dance, the analysis of the extracted information shows that the game is associated with Positive feelings,

Energy, Personal relationships, Self-esteem, and Social support, confirming the purpose of the interactive dance activity of the game. This illustrates the potential and usefulness of the approach we propose to analyse health-related feedback.

This being said, some aspects, such as Fatigue, are more challenging to assess through social media comments and may require that further research is conducted in the presence, physical or remote, of the user. In these cases, a more complete assessment may be necessary, in which the results obtained through opinion mining analysis need to be combined with those of a more traditional approach. Still, the results of the automatically processed comments could provide important insights on the specific focus of future testing and research with users.

In analysing the results displayed on the Dashboard, it is possible to observe a steep increase in the number of comments in the year 2020, strongly marked by the COVID-19 pandemic. This increase may be related to the quarantine periods of millions of people, who were forced to spend most of their time inside their homes. With gyms, discos and other recreational spaces closed, people looked for alternatives to remain active during the lockdown, and in this way preserve their physical, psychological and social well-being. Results such as this show that it is also possible to track the popularity of a product.

All in all, the approach adopted to explore user opinions is effective in identifying the different perceptions of users, whether positive, neutral, or negative. In exploring these user narratives, it is possible to tap into the perceived impacts of the game and act upon them in future (re)design iterations.

The annotation approach followed, based on a controlled vocabulary, also has limitations. For example, words which are not included in the dictionary are overlooked, thus discarding the annotation of some comments that could be associated with concepts of the vocabulary. As mentioned in the previous section, it is also difficult to identify subjective aspects such as irony, sarcasm or humour, namely when the comment is absent of emojis.

Another shortcoming relates to the comments themselves that, in being entered by users, may provide inaccurate information (e.g., title entered without explicitly mentioning the version of the game) or malicious or repetitive comments (e.g., comment repeating the same opinion over and over again). Comments could also be entered by bots. It is, therefore, necessary to work towards a strategy that could identify and exclude such comments. It is also worth noting that Just Dance is not a pure exergame, nor are its users necessarily identical to those of the average user who plays games for health. Nevertheless, this game that we chose as a case study has been studied for its impacts on health and has shown to be beneficial in promoting the health and well-being of its players. This warrants its health-related interest, thus making Just Dance a suitable starting point for this work. As far as the users beyond the comments and the implications that different types of users could imply in the type of user narratives produced, we consider it a strength that our methodology uses raw, untidy and unstructured comments, as these are possibly the closest, most natural, and plain form of externalisation of what the user impressions of their experience might be, and thus the most relevant when considering product use in realistic conditions. This being said, this aspect of the research deserves to be further investigated.

One of the initial challenges of this work concerned the lack of a public annotated and validated data set that could be used to carry out the analysis using a large set of reliable data. The backdrop of that resource has now been created with this work and all information

collected and analysed has been made publicly available[3]. Now that this curated dataset has been created, it is possible to further investigate its usefulness in understanding user perceptions of specific games for health and in using natural language processing on social media data to extract knowledge about the quality-of-life implications in using a game for health. A possible evolution of this research could be applied to a more health-oriented game such as Nintendo's Ring Fit Adventure. Eventually, the methodology should be ready to be applied to a dedicated serious game for health, addressing, for example, stroke rehabilitation or fall-prevention.

Although the system generally performs well, there are situations in which it may misidentify concepts, as highlighted by the manual validation that was carried out. This issue should be addressed in future iterations of the system and could be further improved by resorting to the manual annotation by experts of a larger data set, for example by using Amazon Mechanical Turk [5], as previous studies have done [57]. This type of asset would be key to training deep learning models that could further explore comments' hidden concepts.

7 Conclusions

This research explored the potential of user opinions available on YouTube to automatically extract information about Usability, User Experience (UX) and Perceived Health Impacts related to Quality of Life (H-QoL). In doing so, this work takes stock of the increasing amount of information that is available on public fora and paves the way for using natural language processing on social media data to extract knowledge about quality-of-life implications in using a game for health.

Specifically, this work aimed to address the following research question: Can user comments from YouTube about Just Dance provide relevant insight into users' perceptions of the game's usability, user experience, and health impacts? Having developed this research, we can confirm that it is possible to retrieve valuable information from YouTube comments and that these can improve our understanding of the user's perceptions. In analysing YouTube comments about Just Dance, 90% of the user comments provided relevant information about User Experience (90%), 76% about Perceived Health Impacts related to Quality of Life (H-QoL), and a smaller percentage, of 39%, about usability.

The work makes two important contributions, in which it extends existing research, contributing:

1. A games for health opinion mining methodology, that furthers previous research, by adding a dimension of health-related quality-of-life aspects (H-QoL).
2. A public annotated and validated data set of user comments collected from comments retrieved from YouTube videos about Just Dance.
3. A web-based dashboard that allows for the visual exploration of the automatically extracted and analysed data.

In working towards these contributions, the work uses Just Dance, a world-reference in exergames, and the comments posted by its players on YouTube over the past 12 years, in an effort that, at present, amounts for over 500,000 comments, 32,000 videos, from 15,000

[3]Available at https://www.kaggle.com/datasets/renatojmsantos/just-dance-on-youtube; all related materials at https://linktr.ee/justdanceproject

channels. A byproduct of this work is then the analysis of the users' experience and perceived impacts of the game. This information can be browsed and explored in detail by accessing a public dashboard developed and deployed using open source tools.

The analysis of the automatically extracted information about the Just Dance game displayed in the dashboard shows, for example, that there was a peak in the volume of comments when the COVID-19 pandemic surged. It also shows that the perceptions of users about the game are positive, not only in general, but also in each of the dimensions analysed, where *Satisfaction* is the most detected concept in the usability dimension, *Affect and Emotion* is the most detected concept for UX, and *Positive feelings* is the most detected concept within the H-QoL dimension. Having this information available, it is now possible for those interested in improving the game, to delve into the comments and find out exactly what is in the user comments that could inform redesign and further improvements or iterations. This is however not without limitations, as, for example, sarcastic comments are difficult to recognise by humans let alone machines.

Furthermore, in using the Just Dance game as an example, it was possible to confirm the usefulness of the H-QoL vocabulary, which is the second dimension with the most annotated concepts. With the work developed, we hope to have contributed a new approach to improving games for health and that this research can now be the stepping stone to similar analyses of other applications and games for health.

Acknowledgements The authors are grateful to Joel Arrais, Sérgio Matos, and José Luís Oliveira for the insightful suggestions and discussions throughout the development of this work.

Supplementary information All data related to this work is available at https://linktr.ee/justdanceproject and at https://www.bit.ly/dashboard-just-dance where links can be found to the project github, and Kaggle, among other resources.

Funding This work was supported by project the FCT - Foundation for Science and Technology in the context of the project CISUC - UID/CEC/00326/2020 and by project CENTRO-01-0247-FEDER-047148 INPACT - "Intelligent Platform for Autonomous Collaborative Telerehabilitation" financed by the Portugal2020 program and European Union's structural funds.

Data Availability All data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Conflict of Interests The authors declare that they have no conflicts of interest.

References

1. Akbik A et al (2019) Flair: an easy-to-use framework for state-of-the-art nlp. :54–59
2. Alqahtani F, Orji R (2020) Insights from user reviews to improve mental health apps. Health Inf J 26:45821989649. https://doi.org/10.1177/1460458219896492
3. Bakiu E, Guzman E (2017) Which feature is unusable? detecting usability and user experience issues from user reviews. IEEE, Portugal, pp 182–187
4. Bravo-Marquez F (2021) felipebravom/StaticTwitterSent. https://github.com/felipebravom/StaticTwitterSent/blob/f27b5f ee1aedc9a2eb3241d81911cc27879f521a/extra/SentiStrength/SlangLookupTable.txt
5. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5. https://doi.org/10.1177/1745691610393980
6. Buzzi MC et al (2017) Facebook: a new tool for collecting health data? Multimed Tools Appl 76(8):7–10700. https://doi.org/10.1007/s11042-015-3190-4
7. Ceci L (2022) YouTube - statistics & facts. https://www.statista.com/topics/2019/youtube/

8. Clark EM et al (2018) A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter
9. da Silva THO, Freitas LM, Mendes MS (2017) Beyond traditional evaluations: user's view in app stores. ACM, Brazil, pp 1–10
10. Diniz LdN, de Souza Filho JC, Carvalho RM (2022) Can user reviews indicate usability heuristic issues? CHI EA '22. Association for Computing Machinery, New York. https://doi.org/10.1145/3491101.3519659
11. Full Emoji List v13. 1 (2022). https://unicode.org/emoji/charts/full-emoji-list.html
12. Guzman E, Maalej W (2014) How do users like this feature? A fine grained sentiment analysis of app reviews. In: 2014 IEEE 22nd international requirements engineering conference, RE 2014 - proceedings, pp 153–162. https://doi.org/10.1109/RE.2014.6912257
13. Hallewell Haslwanter JD, Fitzpatrick G, Miesenberger K (2018) Key factors in the engineering process for systems for aging in place contributing to low usability and success. J Enabling Technol 12(4):186–196. https://doi.org/10.1108/JET-12-2017-0053, publisher: Emerald Publishing Limited.
14. Hedegaard S, Simonsen JG (2013) Extracting usability and user experience information from online user reviews. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 2089–2098, https://doi.org/10.1145/2470654.2481286
15. Hori Y, Baba A (2010) Evaluation of digital games using QOL measurements. In: Proceedings of the 18th ACM international conference on multimedia. ACM, New York, pp 1039–1042, https://doi.org/10.1145/1873951.1874144
16. Hoysniemi J (2006) International survey on the dance dance revolution game. Comput Entertain 4(2):8. https://doi.org/10.1145/1129006.1129019
17. Hutto CJ (2021) cjhutto/vaderSentiment. https://github.com/cjhutto/vaderSentiment
18. Hyde M, Wiggins RD, Higgs P, Blane DB (2003) A measure of quality of life in early old age: the theory, development and properties of a needs satisfaction model (CASP-19). Aging Mental Health 7(3):186–194. https://doi.org/10.1080/1360786031000101157
19. Jin J, Ji P, Kwong CK (2016) What makes consumers unsatisfied with your products: review analysis at a fine-grained level. Eng Appl Artif Intell 47:38–48. https://doi.org/10.1016/j.engappai.2015.05.006
20. Just Dance (video game series) (2021). https://en.wikipedia.org/w/index.php?title=Just_Dance_(video_game_series)&oldid=998789672
21. Li Z, Fan Y, Jiang B, Lei T, Liu W (2019) A survey on sentiment analysis and opinion mining for social multimedia. Multimed Tools Appl 78(6):6939–6967
22. Lin J-H (2015) "Just Dance": the effects of exergame feedback and controller use on physical activity and psychological outcomes. Games Health J 4(3):183–189. https://doi.org/10.1089/g4h.2014.0092
23. Loper E, Bird S (2002) Nltk: the natural language toolkit
24. Loria S (2018) Textblob documentation Release 0.15 2
25. MacDonald K (2014) How just dance conquered the world - IGN. https://www.ign.com/articles/2014/01/17/how-just-dance-conquered-the-world
26. Magge A et al (2021) Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. 21–32
27. Mansfield L et al (2018) Sport and dance interventions for healthy young people (15–24 years) to promote subjective well-being: a systematic review. BMJ Open 8(7):e020959
28. Maynard D, Roberts I, Greenwood MA, Rout D, Bontcheva K (2017) A framework for real-time semantic social media analysis. J Web Semant 44:75–88. https://doi.org/10.1016/j.websem.2017.05.002
29. McConnell J (2021) Eclipse Jetty | the eclipse foundation. https://www.eclipse.org/jetty/
30. Messaoudi C, Guessoum Z, Ben Romdhane L (2022) Opinion mining in online social media:, a survey. Soc Netw Anal Min 12(1):1–18
31. Metabase (2021). https://www.metabase.com/
32. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: an on-line lexical database *. Int J Lexicography 3(4):235–244. https://doi.org/10.1093/ijl/3.4.235
33. Mohammad S, Turney P (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29:436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x
34. O'Neil P, O'Neil E, Chen X, Revilak S, Nambiar R, Poess M (2009) Berlin. In: Nambiar R, Poess M (eds) Performance evaluation and benchmarking. Springer, pp 237–252
35. Oyebode O, Alqahtani F, Orji R (2020) Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. IEEE Access 8:41–111158. https://doi.org/10.1109/ACCESS.2020.3002176
36. Peng W, Lin J-HT, Crouse Waddell J (2011) Is playing exergames really exercising? a meta-analysis of energy expenditure in active video games. Cyberpsychol Behav Soc Netw 14:681–8. https://doi.org/10.1089/cyber.2010.0578

37. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
38. Porter MF (2001) Snowball: a language for stemming algorithms
39. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL (2014) Twitter: a good place to detect health conditions. PLoS ONE 9(1):e86191. https://doi.org/10.1371/journal.pone.0086191
40. Sardi L, Idri A, Fernández-Alemán JL (2017) A systematic review of gamification in e-Health. J Biomed Inf 71:31–48. https://doi.org/10.1016/j.jbi.2017.05.011
41. Sawyer B (2008) From cells to cell processors: the integration of health and video games. IEEE Comput Graph Appl 28(6):83–85. https://doi.org/10.1109/MCG.2008.114
42. Severyn A, Moschitti A, Uryupina O, Plank B, Filippova K (2016) Multi-lingual opinion mining on YouTube. Inf Process Manag 52(1):46–60. https://doi.org/10.1016/j.ipm.2015.03.002
43. Shuyo N (2010) Language detection library for java. http://code.google.com/p/language-detection/
44. Sifei (2021) Sifei/Dictionary-for-Sentiment-Analysis. https://github.com/sifei/Dictionary-for-Sentiment-Analysis
45. Silva PA, Santos R (2021) Setting up a health-related quality of life vocabulary, ICMI '21 Companion. Association for Computing Machinery, New York, pp 169–175. https://doi.org/10.1145/3461615.3485401
46. Sirbu D et al (2016) Extracting gamers' opinions from reviews. IEEE, Romania, pp 227–232
47. Stråat B, Verhagen HH (2017) Probing user opinions in an indirect way: an aspect based sentiment analysis of game reviews, AcademicMindtrek '17. Association for Computing Machinery, New York, pp 1–7
48. The 15 Best Dancing (And Rhythm) Video Games Ever Made Ranked (2019). https://www.thegamer.com/best-dancing-rhythm-video-games-ever-made-ranked/
49. Tromp E, Pechenizkiy M (2014) Rule-based emotion detection on social media: putting tweets on plutchik's wheel. arXiv:1412.4682
50. Tuch AN, Trusell R, Hornbæk K (2013) Analyzing users' narratives to understand experience with interactive products. ACM, France, pp 2079–2088
51. Ushaw G, Davison R, Eyre J, Morgan G (2015) Adopting best practices from the games industry in development of serious games for health. ACM, Italy, pp 1–8
52. WASD (2020) Just dance 2020. https://www.wasd.pt/wp-content/uploads/2019/11/JD20_SCREENSHO T_RAIN_OVER_ME_02_350557.jpg
53. Wang Y, Lu X, Tan Y (2018) Impact of product attributes on customer satisfaction: an analysis of online reviews for washing machines. Electron Commer Res Appl 29:1–11. https://doi.org/10.1016/j.elerap.2018.03.003
54. Wattanasoontorn V, Boada I, García R, Sbert M (2013) Serious games for health. Entertainment Computing 4(4):231–247. https://doi.org/10.1016/j.entcom.2013.09.002
55. Weber L et al (2020) Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. arXiv:2008.07347
56. Wikipedia (2021) List of English contractions. https://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_English_contractions&oldid=1025517835
57. Woldemariam Y (2016) Sentiment analysis in a cross-media analysis framework. IEEE, China, pp 1–5
58. Yadav A, Fard FH (2020) Semantic analysis of issues on Google play and twitter. ACM, South Korea, pp 308–309
59. Yang B, Liu Y, Liang Y, Tang M (2019) Exploiting user experience from online customer reviews for product design. Int J Inf Manag 46:173–186. https://doi.org/10.1016/j.ijinfomgt.2018.12.006
60. YouTube (2021) YouTube Data API. https://developers.google.com/youtube/v3
61. Zhu M, Fang X (2015) A lexical approach to study computer games and game play experience via online reviews. Int J Hum–Comput Interact 31(6):413–426. https://doi.org/10.1080/10447318.2015.1036228, publisher: Taylor & Francis _eprint:
62. Zucco C, Calabrese B, Agapito G, Guzzi PH, Cannataro M (2020) Sentiment analysis for mining texts and social networks data: methods and tools. Wiley Interdiscip Rev Data Min Knowl Disc 10(1):e1333