Personal and Ubiquitous Computing manuscript No.
(will be inserted by the editor)

Sentiment Analysis and Twitter: A Game Proposal

Marco Furini · Manuela Montangero

Received: date / Accepted: date

Abstract Pervasive sensing of people's opinions is becoming critical in strategic decision processes, as it may be helpful in identifying problems and strengthening strategies. A recent research trend is to understand users' opinions through a sentiment analysis of contents published in the Twitter platform. This approach involves two challenges: the large volume of available data and the large variety of used languages combined with the brevity of texts. The former makes manual analysis unreasonable, whereas the latter complicates any type of automatic analysis. Since sentiment analysis is a difficult process for computers, but it is quite simple for humans, in this article we transform the sentiment analysis process into a game. Indeed, we consider the *game with a purpose* approach and we propose a game that involves users in classifying the polarity (e.g., positive, negative, neutral) and the sentiment (e.g., joy, surprise, sadness, etc.) of tweets. To evaluate the proposal, we used a dataset of 52,877 tweets, we developed a Web-based game, we invited people to play the game, and we validated the results through two different methods: ground-truth and manual assessment. The obtained results showed that the game approach is effective in measuring people' sentiments and also highlighted that participants liked to play the game.

M. Furini
Dipartimento di Comunicazione ed Economia
Universtà di Modena and Reggio Emilia, Viale Allegri 9, Reggio Emilia, 42121, Italy,
E-mail: marco.furini@unimore.it

M. Montangero
Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Universtà di Modena and Reggio Emilia, Via Campi 213, Modena, 42121, Italy,
E-mail: manuela.montangero@unimore.it

Keywords Pervasive Sentiment Analysis; Gamification; GWAP; Sentiment Analysis; Sentiment Classification; Twitter Analysis.

1 Introduction

The understanding of people' sentiments is a critical factor in strategic decision processes, as it may be helpful in identifying problems and strengthening strategies. For instance, politicians may gauge the public mood to improve their political decisions, enterprise managers may increase customers engagement by tracking what people think about products and services, city administrators may analyze citizens opinions to enhance the life quality of the city, advertisers can improve the effectiveness of their messages by analyzing what people think of the brand [1,2].

Traditionally, the process of understanding opinions and feelings of people involved market research companies that usually used opinion polls, interviews, questionnaires and forms [3–5]. Although effective in most cases, this methodology has some limits: it is expensive and time consuming. To overcome these limitations, nowadays researchers are proposing methods that focus on contents posted on social media platforms like Twitter [6,7]. Indeed, in the last few years, we have witnessed an exponential growth of data publicly available in Twitter, where users, consumers, voters, businesspersons, governments and organizations write and discuss about all sort of topics. Moreover, in social media platforms, users' generated contents are associated with metadata like OS language, device type, capture time and geographical location [8,9]. This means that Twitter contents provide a wealth of opportunities for

understanding people's opinions and feelings about services, brands, events, etc.

The change from users contents to users sentiment is not trivial. In the literature, there are three different approaches that aim to transform tweets into sentiments: manual, automatic, and hybrid. The manual approach requires human beings to read and categorize every tweet (e.g., positive, negative, happy, sad, etc.). If on the one side, this approach provides accurate results, on the other side it is time and cost consuming, which make it an impracticable approach when dealing with large data volumes. The automatic approach uses natural language processing, text analysis and computational linguistics techniques to identify the sentiment of the written text. If on the one side, this approach can analyze large data volumes, on the other side, it is often difficult to achieve accurate results due to the ambiguity of natural language, to the characteristics of the posted content (e.g., irony is very difficult to detect in an automatic way [10]), and to the presence of hashtags, cashtags, emoticons and links [11,12]. The hybrid approach manually annotate a dataset to train a technique that will automatically analyze messages. If on the one side, this approach can analyze large data volumes, on the other side the training phase requires a manual annotation of large datasets in order to achieve good results in the automatic phase. It is worth highlighting a recent alternative approach that relies on existing lexical databases to avoid the manual annotation of large datasets. These lexicons associate affective words with sentiment/polarity values. For instance, the AFINN lexicon [13] associates a value that ranges from -5 to +5 to a list of words and computes the polarity of a tweet consequently (e.g., the word *abandoned* contributes to the negative polarity score of a tweet with 2 points, the word *abuse* contributes with 3 negative points; conversely, the word *accomplish* contributes to the positive polarity with 2 points and the word *adorable* contributes with 3 points). If on the one side, the lexicon-based approach might speed-up the sentiment analysis process, on the other side it is to note that its effectiveness depends on the effectiveness of the lexicon database and, moreover, it might be difficult to access to lexical databases. Indeed, effective lexicons are widely available for English language [12, 14,15], but are poorly available for other languages [16–18].

Motivated by the scarcity of lexicon for non-English languages, by the success of social games and applications [19,20], by the Game With A Purpose methodology [21], and by a preliminary study in the area that showed how users liked to play with tweets and sentiments [22], in this paper we propose a game that transforms users' tweets into users' sentiments. The idea is to involve humans in the sentiment analysis process of tweets. Indeed, sentiment analysis is a typical task that is easy for humans, but difficult for computers. Our hypothesis is that the game approach overcomes the limitations of current methods, as it allows performing sentiment analysis of tweets written in any language with the accuracy typical of human beings. Moreover, the obtained sentiment classification might be helpful to create lexicons able to support automatic sentiment analysis of contents written in any language. Briefly speaking, we consider a discrete emotional space and we transform the sentiment analysis process into a GWAP game: players have to score more points than others do by correctly classifying the polarity (either positive, negative or neutral) and the sentiment (e.g., "Joy", "Surprise", "Anger", "Fear", etc.) of tweets. A correct classification occurs when two or more players (unknown to each other) agree on the same call.

To understand the potential of our proposal, we create a dataset by collecting and filtering tweets generated in Italy. The resulting dataset is composed of 52,877 tweets. Then, we sent a game invitation to ca. 950 students enrolled at our University (the University of Modena and Reggio Emilia). During the first 30 days, students played the game and classified the polarity of 10,253 tweets and the sentiment of 9,933 tweets. To evaluate our proposal, we performed an engagement and a validation analysis. The former shows that participants liked the game approach (72% of the players evaluated more than 10 tweets, and 32% of them evaluated more than 50 tweets) and played the game at any time of the day (i.e., particularly played in the early morning, in the late afternoon and in the evening) and on any week day (i.e., no significant difference between week-end or weekdays). The latter analysis is done with two different methods: ground-truth and manual assessment. The ground-truth method asks volunteers to classify tweets and then it compares the classification obtained with the game against the classification obtained with the ground-truth approach; the obtained results show that 88.8% of the polarity classifications and 83.2% of the sentiment classifications were consistent with those of the ground-truth, which is a very good result with respect to the 42% of inconsistencies found in the only other study, we are aware of, that checked the results of a manual classification against another manual classification [18]. The manual assessment method shows volunteers, different from those who performed the ground-truth, the game classifications and asks them if they agree with the classifications; the obtained results show that volunteers agree on 96.6% po-

larity classifications and on 88.3% sentiment classifications.

In summary, the results obtained in the evaluation process show that the game approach might be useful to transform users' contents into users' sentiment. As mentioned, in addition to the direct understanding of the sentiment of tweets messages written in any language, the sentiment classification might be useful for the production of specific lexicon databases (e.g., for specific scenarios and/or for specific languages), which is the first step towards the development of a process able to automatically transform users' contents into users' sentiment.

The paper is organized as follows. Section 2 presents studies and proposals in the area of sentiment analysis applied to Twitter messages; Section 3 describes details of our proposal; Section 4 presents the experimental assessment; Section 5 describes the results validation process. Conclusions are drawn in Section 6.

2 Background and Related Work

In the past few years, many researchers focused their attention on the problem of detecting people's sentiments by using data publicly available in social media platforms. In the following, we first present studies that give classifications of the possible sentiments people may feel, then we focus on the recent approaches that aim to transform users' textual contents into users' sentiments and, finally, we overview some Game With A Purpose studies in which a gamification approach is used to solve problems that are easy for humans, but difficult for computers.

2.1 Sentiment Description

When dealing with sentiment analysis, it is mandatory to define what are the possible sentiments a human being may feel. Hence, it is necessary to understand what is meant by sentiment. According to the Merriam-Webster dictionary, a sentiment is defined either as "an attitude, thought, or judgment prompted by feeling" or as "emotion" and the word "emotion" is defined as "a conscious mental reaction subjectively experienced as strong feeling". Therefore, to understand the sentiments of people it is necessary to define which possible emotions a human being may feel. In literature, as shown in Table 1, there is no a unique model to categorize emotions: Parrott [23] defined six basic emotions (i.e., anger, fear, sadness, joy, love and surprise); Arnold [24] proposed a list of eleven basic emotions (i.e., aversion, anger, courage, dejection, desire, despair, fear, hate, hope, love and sadness); Ekman [25] delineated six basic emotions (i.e., anger, disgust, fear, joy, sadness and surprise), whereas Plutchik [26] defines eight basic emotions (i.e., joy, trust, fear, surprise, sadness, disgust, anger, anticipation) that may be extended to twenty-four emotions (i.e., serenity, acceptance, apprehension, distraction, pensiveness, boredom, annoyance, interest, joy, trust, fear, surprise, sadness, disgust, anger, anticipation, ecstasy, admiration, terror, amazement, grief, loathing, rage, vigilance).

Emotional Model	Categories of emotions
Arnold [24]	*aversion, anger, courage, dejection, desire, despair, fear, hate, hope love, sadness*
Ekman [25]	*anger, disgust, fear, joy, sadness and surprise*
Parrott [23]	*anger, fear, sadness, joy, love and surprise*
Plutchik [26]	*joy, trust, fear, surprise, sadness disgust, anger, anticipation*

Table 1: Categories of emotions.

There is not a criterion to define the quality or the effectiveness of the various models, and, usually, the choice falls on subjective criteria. For example, in image analysis the preferred model is Plutchik's, as the model provides a direct association between colors and emotions (e.g., yellow represents joy, red stands for anger, etc.).

2.2 Sentiment Detection

The most recent approaches that aim to transform users' textual contents into users' sentiments use either a crowdsourcing approach or involve the use of a lexicon (i.e., a database of affective words that are associated to sentiment values).

The crowdsourcing approach aims to use human beings in the sentiment evaluation process. For instance, Nakov et al. [12] used the crowdsourcing approach to call the sentiment of tweets and SMS messages. The millions of gathered tweets were filtered to identify messages that express sentiments towards specific topics by means of SentiWordNet [27], while SMS were taken from the NUS SMS Corpus[1]. Mitchell et al. [28] studied the correlation between social level of happiness and geographic location, both at state and urban level, across U.S.A. To measure happiness they use word frequency distributions, collected from a large corpus of geolocated tweets, with roughly 10,000 individual words scored for their happiness independently by users of

[1] http://wing.comp.nus.edu.sg/SMSCorpus/

Amazon's Mechanical Turk[2], a service that coordinates workers and requesters to perform tasks that are difficult for computers.

The lexicon approach might speed up the automatic analysis as it relies on a list of words that are associated to sentiment values. For instance, Lin [29] used the lexicon SentiSense [14] to extract the sentiments of tweets geolocated in the area of Pittsburgh. In particular, the lexicon associates English words to 14 different categories of sentiments (e.g., fear, joy, anticipation, disgust, etc.). Sahu et al. [7] relied on an existing lexicon to call the polarity of tweets. Fast et al. [30] focused on lexicon generation and designed Empath, a tool that uses a combination of deep learning and crowdsourcing to generate and validate new lexical categories on demand from a small set of seed terms. Hamilton et al. [31] proposed SentProp, a method to induce accurate domain-specific sentiment lexicons using small sets of seed words.

It is worth noting that the use of a lexicon in sentiment analysis is promising, but there are few burdens that may limit its effectiveness. For instance, the lexicon approach may have difficulties in capturing sentiment signals in details due to the language used to write tweets (i.e., emoticons, acronyms, links, etc.) [29]. Therefore, it is necessary to improve the pre-processing phase of tweets. For instance, Sahu et al. [7] suggested to remove URLs, mentions, stopwords, and symbols, to replace emoticons with words, to split joint words (i.e., the word "#germanydefeatbrazil" is split into "germany", "defeat" and "brazil"), and to perform a spell correction (i.e., the word "gooooal" is replaced with the word "goal"). Moreover, the availability of effective lexicons for the Twitter scenario is very limited for non English languages. Indeed, individual companies might have their own private lexicon for sentiment monitoring services, but these resources are not shared nor publicly available [18].

Due to the importance of having a lexicon for language analysis, researchers are beginning to produce domain-specific lexicons for non English languages. For instance, Bosco et al. [16] created a six-category lexicon (i.e., positive, negative, objective, mixed, ironic and unintelligible) by manually annotating 1500 Italian tweets. The lexicon is then used in Felicittà[3], a Web platform designed to estimate the happiness in Italian cities by means of sentiment analysis over geotagged tweets. Erik Tjong Kim Sang [17] created a Dutch sentiment lexicon by using tweets written in Dutch. Motivated by the lack of automatic mechanisms to perform Dutch sentiment analysis, the method proposed to manually create a lexicon using words belonging to tweets that contain either smiles or frownies. Words within tweets with smiles are positive, whereas words within tweets with frownies are negative. Aslan et al. [32] highlighted the importance of having a lexicon for language analysis and proposed a computational morphological lexicon for Turkish since there has been no study in the field. Mikolov et al. [33] developed a method that can automate the process of generating and extending dictionaries and phrase tables. Despite its simplicity, the method proved effective: authors achieved almost 90% precision for translation of words between English and Spanish.

[2] http://www.mturk.com/mturk/

[3] http://www.felicitta.net

2.3 Games With A Purpose

The Games With A Purpose (GWAP) approach uses people to perform tasks that are difficult for computers and the particularity is that the involvement takes place with a game and in a fun way. For instance, typical actions that are easy for humans, but quite difficult for computers are: the understanding of the objects of a picture, the detection of a sentiment in a written sentence and the geolocalization of movie scenes. In general, humans are better than computers when the task involves creativity, reasoning, or emotions (e.g., [34–37]).

The first known GWAP example is ESP, a game designed to transform the labeling process of into an entertainment game [38]. Indeed, the labeling process is another example of process difficult for computers, but quite easy for humans. The assumptions of ESP is that if two people, who do not know each other, associate the same label to the same picture, then the label is considered appropriate for the image. For example, if two people associate the label "Paris" to the same picture, then it is very likely that the image relates to Paris. Over the years, the use of the GWAP approach involved different fields: Lux et al. [39] proposed to introduce game elements in the ranking images process; Kacorri et al. [40] proposed a video caption editing system based on crowdsourced work. Furini [35] proposed a gamification approach in the transcription process of digital video lectures.

It is worth noting that the GWAP approach must engage people: without participation, the game fails. Many different studies analyzed what motivates users in playing GWAP game (*e.g.* [21,41–43]), and the main reason is quite simple: they play because they desire to be entertained; they are not interested in the global task to be accomplish. Therefore, the most critical factor when designing a GWAP is to create a game structure (e.g., rules and winning conditions) that encourages and motivates people to play. GWAP can be seen

as a variant of more classical crowdsourcing strategies in which workers are often incentivized by monetary rewards.

3 The Sentiment Analysis Game

To transform the tweet sentiment analysis process into an entertainment activity, we design a GWAP game that asks players to classify the sentiment of tweets. The tweet to be classified is randomly selected from the available dataset and is presented to a player that wants to play the game. The player's classifications (polarity and sentiment) are attached to the tweet, but are not shown to anyone. When another player agrees on the same classification(s), both players gain points. The assumption is that if two people, who do not know each other, associate the same sentiment to the same tweet, then the sentiment is considered appropriate for the tweet and both players score points. For example, if two people associate "Joy" to the same tweet, then it is very likely that the tweet is a message of joy.

It is to note that tweets are randomly assigned to players and, therefore, the possibility of players' agreement occurring by chance is very low. Indeed, players do not know whether the tweet to classify has been already classified or not; if already classified, players do not know who classified the tweet. Therefore, the players classification is not affected by any other players classification and players cannot communicate with other players. Nevertheless, if somehow two players get to know each other and try to agree on the tweets classification, the chance that they are given the same (or more than one) tweet is really small. Indeed, suppose players A and B classified the same tweet. If the two players keep playing, the system will randomly select another tweet for player A and will randomly select another tweet for player B. Therefore, the possibility of players' agreement, either by chance or on purpose, is negligible.

As mentioned, when a process is turned into a GWAP, it is important that the game structure encourages people to play. Hence, in the following, after describing the emotional space (i.e., the categories of polarities and emotions we want to call), we provide details of the game structure by means of game rules, scoring example and game environment.

3.1 The Emotional Space

The sentiment analysis of a given text may classify messages according to polarity and/or emotion, and the classification process requires categories. For example, to classify the polarity, three categories (e.g., *positive, negative* and *neutral*) may be sufficient, but what are the categories that can be used to classify emotions? As mentioned in Section 2.2, different emotional spaces have been proposed in literature, each with specific characteristics, and that the choice of one or another depends on the particular application field.

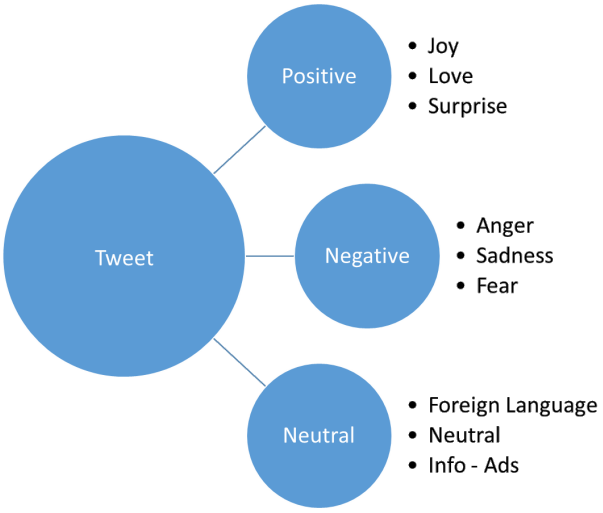

Fig. 1: The adopted sentiment classification: three levels of polarity, each one with three sub-levels.

In this paper, we consider the emotional space proposed by Parrot [23], which classifies emotions into six different categories: *joy, love, surprise, anger, sadness*, and *fear*. The main reason to select this emotional space is that we want the game to be simple, fast and fun and this can be done only with a limited set of emotions. Indeed, a large set likely makes the players think too much, risking to make the game to be complex, slow and boring. Therefore, as shown in Figure 1, our game considers three different categories for polarity (i.e., positive, neutral and negative) and six different categories for emotions. Note that, the Parrot emotional space is well balanced as it includes three positive (i.e., joy, love and surprise), and three negative (e.g., anger, sadness and fear) emotions. Finally, to investigate the reason of the neutrality of a tweet, we consider three possible options: foreign language, info-ads, and neutral. Indeed, in addition to tweets written in neutral terms, people may consider tweets written in a foreign incomprehensible language or written to advertise a product/service as neutral.

3.2 Rules of the Game

– **Object of the game**: become the "Sentiment-teller" by scoring more points than other players.

- **Playing**: once entered the game, the player is presented with a tweet and has first to evaluate its polarity, and then its sentiment. If a player leaves the game, the game engine shows the top ten high-scores.
- **Scoring**: the player gains one point for every played tweet. If the call matches the one of another player, the player gains a bonus of nine points for every match (e.g., 9 points if the call matches the polarity and 9 points if the call matches the sentiment).
- **End of the game**: the player exits the game or there are no tweets to call.
- **High-score**: when the player exits the game, the top ten scoring players list is shown.

3.3 Scoring Examples

Scenario 1: No Matches. Alice plays the game. Suppose the system asks her to judge t_a (tweet polarity and sentiment not defined yet). She calls both the polarity and the sentiment and she gains 2 points. Then, Alice wants to play again and the system asks her to classify t_b (a tweet that player Bob called as a "positive" and "joy"). She judges the tweet as neutral and "info/ads". Alice gains 2 points. Then the system asks her to judge t_c (tweet polarity and sentiment not defined yet). She calls just the polarity and then she leaves the game. She gains 1 more point.

Scenario 2: Full Match. Bob plays the game. The system asks him to judge t_a (defined as "negative" and "sadness" by Alice). Bob classifies the tweet as "negative" and "sadness". Bob gains 2 points and gets 18 extra bonus points because his calls (polarity and sentiment) matched the ones of another player. At the same time, Alice gains 18 points because her calls matched the ones of another player. Since two players submitted the same calls, t_a will never be presented to other players.

Scenario 3: Partial Match. Camilla plays the game. The system asks her to classify t_b (defined as "positive" and "joy" by Bob and as "neutral" and "info/ads" by Alice). Camilla judges the tweet as "positive" and "surprise". Camilla gains 2 points and gets 9 extra bonus points because her polarity call matches the one of another player. At the same time, Bob gains 9 extra points because his call matched the one of another player. Suppose now that David plays the game. The system asks him to judge the sentiment of t_b (the polarity has been frozen as "positive" since both Camilla and Bob). He calls the sentiment as "surprise". He gains 1 point and 9 extra points since his call matches the one of another player. Thanks to the David's call, also Camilla gains 9 extra points.

3.4 Game Environment

The game engine uses three different sets of data:

- $T = \{t_1, t_2, t_3, ..., t_n\}$ stores the tweets to classify;
- $P = \{p_1, p_2, p_3, ..., p_n\}$ stores the tweets classified by polarity;
- $S = \{s_1, s_2, s_3, ..., s_n\}$ stores the tweets classified by sentiment.

Each t_i is a triplet $(ID, text, state)$, where ID is the tweet ID, *text* is the content of the tweet and *state* is a flag defined/undefined. Note that, at the beginning, T contains all tweets with "undefined" state, but when two players agree on the same tweet classification, the state is turned to "defined".

Each p_i is a triplet $(ID, state, polarity)$, where ID is the tweet ID, *state* is a flag defined/undefined, and *polarity* is a set of pairs $(player, polarity)$, where $polarity \in \{positive, negative, neutral\}$ is the call of *player* on the tweet. Note that, at the beginning, P is empty and the default value for state is *undefined*. This value changes to *defined* when two players agree on the same polarity.

Each s_i is a triplet $(ID, state, sentiment)$, where ID is the tweet ID, *state* is a flag defined/undefined, and *sentiment* is a set of pairs $(player, emotion)$, where $emotion \in \{$ *joy, love, surprise, anger, sadness, fear, foreign language, neutral, Info/Ads* $\}$ is the call of *player* on the tweet. Note that, at the beginning, S is empty and the default value for state is *undefined*. This value changes to *defined* when two players agree on the same sentiment.

The game flow is shown in Figure 2. First, the game server randomly selects a tweet t_i among the undefined ones. Then, the game asks the polarity (if undefined) and the sentiment. Finally, it shows the high score list.

- **Polarity evaluation:** The game engine checks whether the ID of tweet t_i exists in P and if its status is set to defined. If so, the game skips the polarity investigation and moves to the sentiment evaluation. Otherwise, the player is asked to call the tweet polarity and, if not existing, a new entry is created in P with the tweet ID. Then, the game engine checks whether the call matches the polarity of a previous call (if it exists) by checking the polarity filed in P. If so, players that provided the same call both gain 9 points, and the state of tweet t_i in P is set to defined. Otherwise, the player gains one point and the called polarity is stored in P.

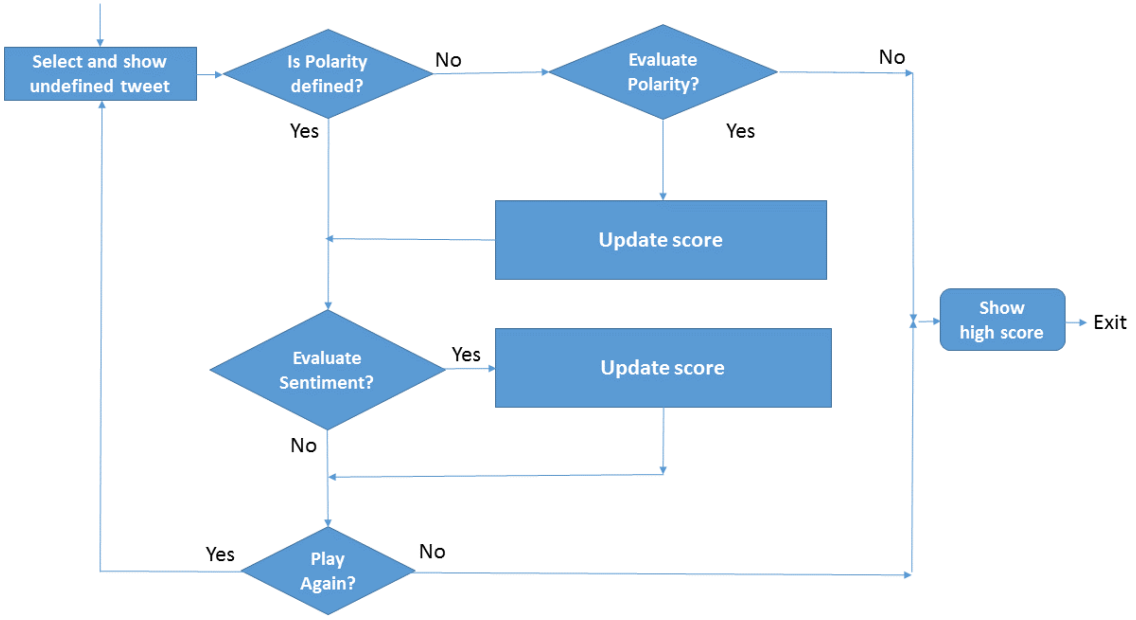

Fig. 2: The Game Flow.

– **Sentiment evaluation:** The player is asked to call sentiment for t_i, and, if not existing, a new entry is created in S with the tweet ID. Then, the game engine checks whether the call matches the sentiment of a previous call (if it exists) by checking the sentiment filed in S. If so, players that provided the same call both gain 9 points, and the state of tweet t_i in S and in T is set to defined. Otherwise, the player gains one point and the called sentiment is stored in S.

4 Experimental Assessment

To evaluate the effectiveness of the proposed approach, we setup an experimental assessment composed of five different steps, as shown in Figure 3: tweets harvesting, tweets filtering, tweets evaluation, analysis of players' behavior and results validation.

4.1 Tweets Harvesting

The first step of the experimental assessment is data collection. Since our goal is to investigate whether the game approach can be effective in understanding the sentiment of people through the usage of Twitter data, we used the Twitter streaming APIs to harvest tweets. In particular, we collected tweets generated in Italy. The number of collected tweets is equal to 65,514 tweets.

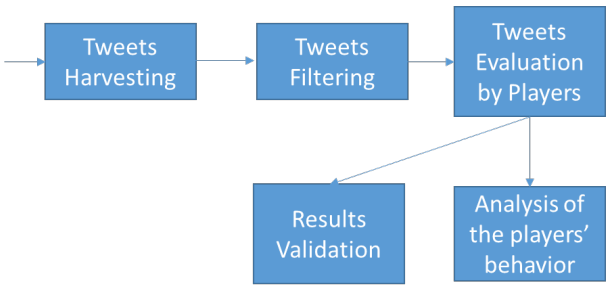

Fig. 3: Five different steps compose the experimental assessment: tweets harvesting, tweets filtering, tweets evaluation by players, results validation and analysis of the players' behavior.

4.2 Tweets Filtering

Before submitting tweets to players, it is necessary to filter these tweets in order to eliminate spam and meaningless tweets, thus avoiding the game to become boring very quickly. Indeed, as shown in Figure 4, we observed that several tweets contained a considerable number of hashtags (e.g., up to 15 hashtags in the same tweet), a considerable number of tweet addresses (up to 13 Twitter addresses in the same tweet), and/or a considerable number of links (e.g., up to 5 links in the same tweets). Therefore, we considered as "spam" all the tweets with more than 4 hashtags (3,843 tweets), more than 2 Twit-

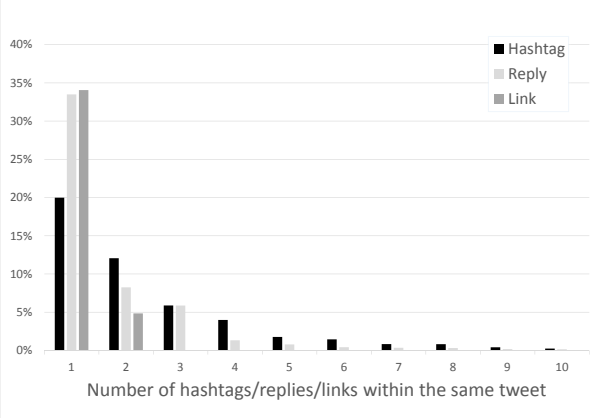

Fig. 4: Characteristics of the tweets with respect to the number of hashtags, replies and links.

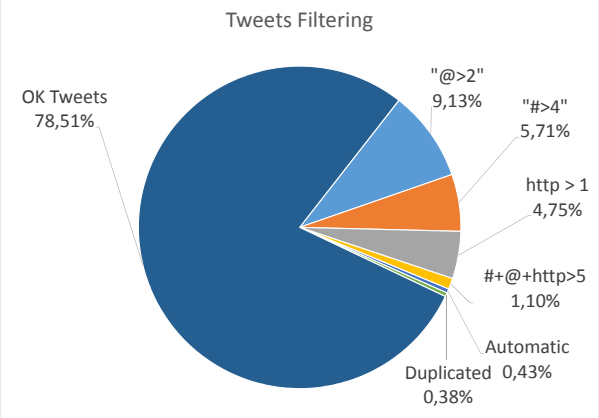

Fig. 5: Dataset analysis: 22.49% of the original messages were considered as "spam" or as "meaningless" tweets.

ter addresses (6,149 tweets) and with more than one link (3,198 tweets).

Furthermore, we observed the presence of meaningless tweets from an emotional point of view. As shown in Figure 5, we considered as "meaningless" all the tweets composed of just a link (735 tweets), all the tweets with a total number of hashtags, Twitter addresses and links more than 5 (740 tweets), all the tweets automatically generated by applications (e.g., "I'm at this place") (290 tweets) and all the duplicated posts (255 tweets).

In summary, we excluded from the dataset 12,637 tweets, which means that the dataset that we use in the game is composed of 52,877 tweets. Note that, although a non negligible percentage of tweets are not written in Italian, we decided to keep them in the dataset because these messages might be worth analyzing to detect the people's sentiment. However, it s worth mentioning that one might decide to easily filter them out if the final goal of the game is to evaluate tweets written in a particular language (e..g, the case of lexicon generation).

(a) (b)

Fig. 6: Game interface: participants are asked to call the polarity of a tweet (a) and then to call its sentiment (b). The tweet can be translated as "#Finland will stop operations in #Kosovo and #Afghanistan".

4.3 Players Tweets Evaluation

To make the system playable, we developed a Web-based and mobile friendly game. Instead of focusing on implementation details, which go outside the scope of this paper, in the following we describe the obtained results. To get users to play the game, we sent invitations through the Department social media forum, to ca. 950 students (note that we are unable to check how many of them read the message). It is worth mentioning that the invitation simply asked to participate in an experiment through a game. No monetary or other incentive has been promised or given in exchange for participation in the game. The developed Web interface is shown in Figure 6. During the analyzed period of 30 days, 521 students participated to the game and they evaluated the polarity of 10,253 tweets (30% positive, 22% negative and 48% neutral) and the sentiment of 9,933 tweets. Note that, the total number of tweets classified by sentiment is slightly smaller than the number of tweets classified by polarity: this is due to two main reasons: i) sometimes players quit after polarity classification, and ii) players agreed on the same polarity, but not on the same sentiment. Figure 7 summarized the obtained tweets classification.

4.4 Players Engagement Analysis

The main goal of this analysis is to measure the participants' engagement in relation to: i) the time of the day in which participants play, ii) the days when par-

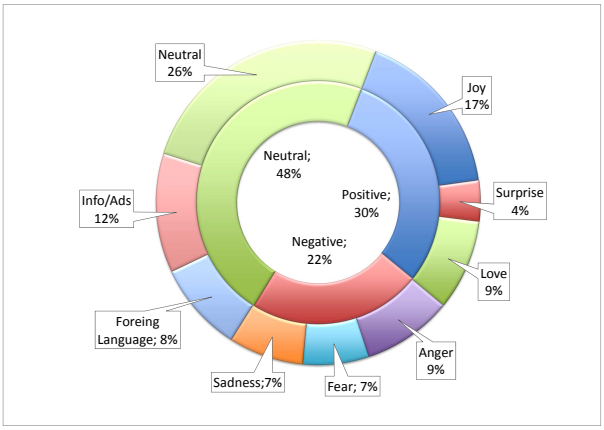

Fig. 7: Players tweets classification.

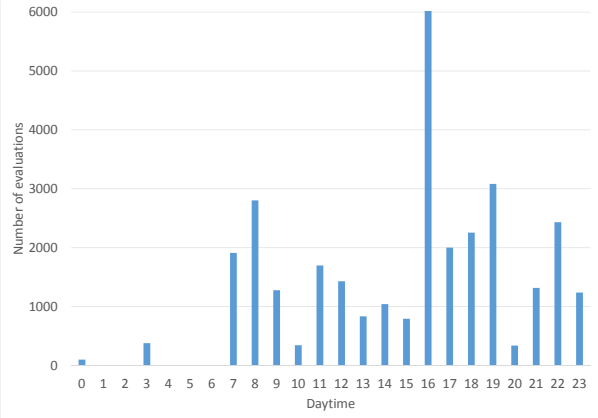

Fig. 8: Time of the day in which participants played the game.

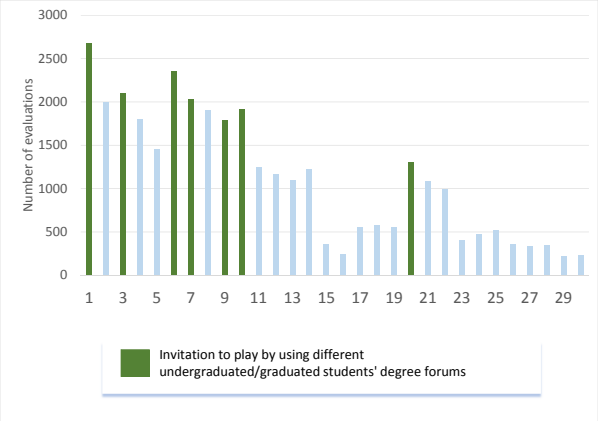

Fig. 9: Tweets played by participants during the 30 days. Day#5, #12, #19 and #26 were Sundays.

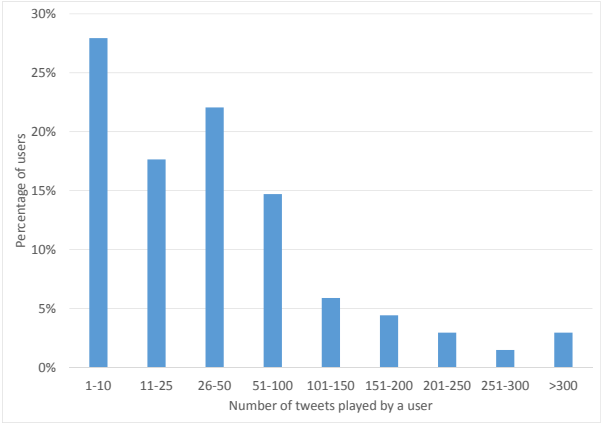

Fig. 10: Users' activity: percentage of participants who played a certain number of tweets.

ticipants play, and iii) the number of tweets played by participants. These measures provide a good indication of if and how the game encourages and motivates people to play.

Figure 8 shows the time of the day in which participants play. It is interesting to observe that, with high probability, students played the game in the morning, likely after checking their personal profile in the students' forum or when commuting; they played the game in the mid afternoon, likely after their lecture time; they played the game in the early evening, likely on their way home or when arriving at home; they played the game after dinner and before going to bed. It is interesting to notice that there are peaks that roughly correspond to lectures breaks. These results highlight the interest towards the game.

Figure 9 shows, day per day, the number of evaluations done during the observed 30 days. It can be noted that the number of evaluations resembles the long-tail shape, which is a typical behavior in video game sales (i.e., sales are very high in the days of the release and then tend to decrease in the next days). However, we can also observe that there are days where the number of evaluations increases with respect to the day before. This is likely due to the timing of the invitations we sent: at day #1 we sent the invitation to 250 ca. students, at day #3 to other 70 ca. students, at day #6 to other 70 ca. students, at day #7 to other 200 ca. students, at day #9 to other 200 ca. students, at day #10 to other 80 ca. students and at day #20 to other 80 ca. students. In total, we sent the invitation to 950 ca. students. Note that t days #5, #12, #19 and #26 were Sundays. As mentioned, these results show the typical trend of newly released games and also show that an invitation to play (in our case, sent through the department social forum) is fueling interest in the game.

Once again, these results highlight the interest towards the game.

Figure 10 shows the percentage of users who played a certain number of tweets. It can be noted that a large number of them (i.e., 32%) played more than 51 tweets, and 22% played a number of tweets that varies between 26 and 50. On the one side, this data highlights the interest towards our game, and that, once started, a player tends to continue playing. On the other side, data also shows a potential problem: 28% of the participants played a number of tweets not greater than 10. Likely, these participants did not find the game attractive, but were nevertheless curios to see how the game was.

5 Results Validation

To fully access the effectiveness of the proposed approach, it is necessary to evaluate the quality of the classification. To this aim, we perform an evaluation similar to the one done in the ESP GWAP game [38] and we consider two different distinct evaluations: *ground-truth* and *manual assessment*. The former involves humans to call polarity and sentiment of tweets and compares results against the ones obtained during the game, whereas the latter involves humans to simply check the classifications performed during the game. It is worth noting, that it is not possible to automatically evaluate the quality of the classification because there are no effective automatic mechanisms to call the sentiment of tweets written in Italian (i.e., this is one of the reasons that urged us to propose the gamification of the sentiment analysis for tweets). In addition, we also provide an analysis of what happens if we increase the number of people who need to agree for the call of the tweet (i.e., in our game this number is equal to two).

5.1 Ground-truth Evaluation

To perform a ground-truth analysis, we called for 15 voluntary participants (different from the players) among Department' students (seniors, graduated and PhDs, aged between 20 and 25 year-old, 10 females and 5 males). We grouped participants into five different groups, each one composed of three evaluators and we split the dataset (i.e., 10.253 tweets) into five subsets (i.e., four composed of 2050 tweets and one composed of 2053 tweets). Each group of evaluators had to classify one subset. A tweet is considered classified when at least two evaluators agree on the same classification. At the end of the process, we compare the ground-truth classification against the game classification. If results differ greatly, it is likely that players classified tweets without paying particular attention to the contents of the tweets; but if results do not differ greatly, then the game approach can be considered effective. It is worth noting that, although the two sets of results are produced by humans, they may differ. Indeed, there are tweets that are difficult to classify, even for a human being. In literature, when using a manual classification approach, it is usually assumed that the produced classification is correct and unambiguous, without any further check. Unfortunately, when dealing with tweets, this is not true and it might be possible to have inconsistent results. For instance, a simple message like "It's snowing" can be labeled as *positive* by young people, and as *negative* by drivers. To the best of our knowledge, the only study that checked the results of a manual classification with another manual classification is [18], where authors aimed to classify tweets according to polarity and irony: they found 42% of inconsistencies between the two manual classifications.

The ground-truth procedure is done as follows: to mitigate the "click here and there" phenomenon (i.e., a random evaluation of the tweets), tweets were listed on paper spreadsheets (each row reported, in addition to the tweet, the check-boxes that evaluators had to check); to avoid distractions, after evaluating 150 tweets, a break of 5 minutes was forced; to avoid any kind of possible influence, evaluators were not aware of the classification obtained during the game and can not communicate among themselves.

The comparison shows that 2,869 classifications were not consistent with those obtained during the game. In particular, 1,149 calls (11.2%) were not consistent with the polarity and 1,720 calls (16.8%) were not consistent with the sentiment. It is interesting to note that these percentages are much lower than the one found in [18] (42% of inconsistencies observed in a manual vs. manual classification).

To deeply understand these inconsistencies, we group them into three different categories: objective (opposite polarity like negative vs. positive), subjective (neutral polarity vs. positive/negative), light (same polarity but different sentiment within the polarity).

Figure 11 shows the different groups of inconsistencies: 9% for the objective group (i.e. 272 inconsistencies), 31% for the subjective group (i.e., 877 inconsistencies), and 60% for the light group (i.e., 1,720 inconsistencies). With respect to the entire dataset, the percentages of inconsistencies are: 2.7% (objective), 8.6% (subjective) and 16.8% (light). It is worth noting that most of the inconsistencies are within the light category, meaning that evaluators agreed on the polarity, but disagreed on the sentiment. This is likely due to

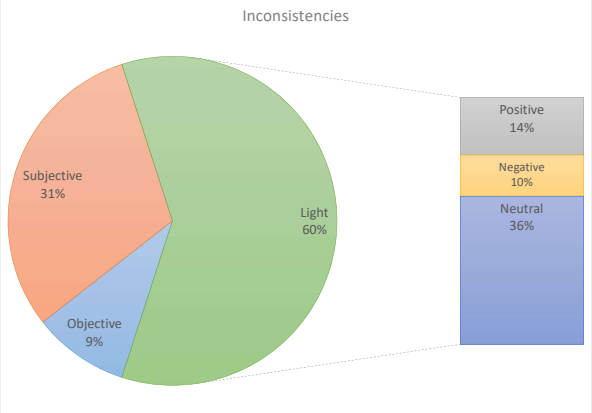

Fig. 11: Different groups of inconsistencies: most of them (60%) are related to sentiment classification (e.g., joy, love and surprise for the positive polarity, anger, sadness and fear for the negative polarity and foreign language, neutral and info/ads for the neutral polarity).

the perception or definition one has of the specific sentiment. Indeed, the between joy, love are feelings that have overlaps and surprise and their distance might be very subtile.

To understand why the light category is so inconsistent, we investigate it deeper to undestand if some sentiments are more difficult to call than others. To this aim, we observed that 402 inconsistencies were related to the positive polarity (3.9% of the entire dataset), 286 were related to the negative polarity (2.8% of the dataset), and 1,032 were related to the neutral polarity (10.1% of the dataset).

Clearly, sentiments within the neutral polarity are more difficult to call. Within neutral polarity, we observed 631 inconsistencies (22% of the inconsistencies and 6.2% of the entire dataset) were between the *neutral* and *information* classification. Clearly, participants did not have clear the difference between the sublevels of the neutral polarity (i.e., foreign language, neutral, info/ads). This data highlights a problem in the definition of the sublevels and not a problem of the game.

In summary, the ground-truth evaluation provides evidences that players input appropriate classification for the tweets.

5.2 Manual Assessment of the classification

To perform a manual assessment of the classification, we called for 15 voluntary participants (different from the players and from ground-truth evaluators) among Department' students (graduating, graduated and PhDs, aged between 20 and 25 year-old, 12 females and 3 males), and we randomly selected 1,500 tweets among the ones classified during the game.

We showed each participant 100 different tweets with the following procedure: (i) show the tweet and the polarity classification and ask participant if he/she agrees on the classification, (ii) show the sentiment classification and ask participant if he/she agrees on the classification.

Results show that participants agreed on 96.6% polarity classification (i.e., 1,449 positive checks vs. 51 negative checks) and on 88.3% sentiment classification (i.e., 1,324 positive checks vs. 176 negative checks), with no significant statistical difference among the 15 participants.

We can observe that these results differ from the ground-truth ones. Probably, the main reason is that evaluators' calls are affected by the given classification. Indeed, if an evaluator is undecided between the neutral and info/ads classification, and if the game classification called for one of the two classifications, the evaluator likely reports the call as correct. Again, this indicates that there is a certain degree of ambiguity in the nature of some tweets and that the classification might not be rigid and unique.

In summary, the manual assessment evaluation provide evidences that players input appropriate classification for the tweets.

5.3 On Improving the quality of the classification

Although, results provide evidences that players input reasonable classifications for the tweets, in this section we analyze possible approaches to improve the quality of the classification. In particular, we focus on the number of players who have to agree on the same classification and on the automatic identification and removal of tweets difficult to classify. In the rest of this section, such number will be addressed as *call-threshold*.

5.3.1 Changing the number of players who need to agree on the same classification

When two players agree on the same tweet classification, the status of a tweet changes from undefined to defined. In the previous subsections, we analyzed the quality of the obtained classifications and found it reasonably good. On the other side, we also highlighted the fact that some tweets seem more difficult than others to classify and some sentiments more difficult to distinguish. Thus, one might think to enhance the confidence on the classification by increasing the call-threshold (for

polarity or sentiment or both). There is no technical difficulty in implementing this variant, nevertheless, it is to notice that it has an impact on the time necessary for the classification.

Suppose we increase the call-threshold for polarity from two to $k > 2$. Then, to classify the polarity of n tweets we need kn evaluations instead of $2n$.

Referring to our 3o days long experiment, on average, players evaluated 845 tweets per day with respect to their polarity (a total of 25,350 evaluations that led to 10,253 polarity classification). If we set $k = 3$, to classify the polarity of 10,253 tweets, we need, at least, 3x10,253=30,759 evaluations. With 854 evaluation per day, it means that 36 days are necessary to make 30,759 evaluations. If we set $k = 4$, then 47 days are necessary.

However, even if one might be willing to wait more, the resulting classification might not be sensibly better. Indeed, we observed that there are tweets that are very difficult to classify because they are intrinsically ambiguous and asking a larger number of people to evaluate them does not really solve the problem. For instance, one of the tweets that received opposite polarity classification, is "Starnuti a gogo?", which can be translated into "Sneezing galore?". This message was called as "positive" and "surprise" by the game, but as "negative" and "anger" by the ground-truth evaluation. It is almost impossible to have a definitive classification on this type of messages. The next section is devoted to discuss how to possibly identify and eliminate such tweets from the dataset.

In summary, by increasing the call-threshold, one might achieve better quality, but the analysis phase takes longer. However, it is worth noting that the tuning of the call-threshold depends on different factors, like the type of analysis one wants to do, the size of the dataset and the number of potential players that can be reached. Indeed, in some specific domains, like enterprise and society, one might prefer the speed of assessment to accuracy, whereas in other specific domains, like in the production of a lexicon reference corpus, a precise identification of people' sentiments would be preferred. There might also be cases (e.g., a social event) in which it is not really important to classify all tweets of the dataset, but just a large enough subset. Therefore, once the number of tweets that received at least one evaluation is over a given threshold (or percentage), one might think to replace the original dataset with the set of the these tweets and, thus, conclude the classification with larger call-threshold in a shorter time.

5.3.2 Automatic identification and removal of tweets difficult to classify

Given a dataset composed by tweets, it is realistic to assume that there are tweets that are difficult to classify or that are not worth classifying (e.g., any type of classification can be questionable for the tweet "Sneezing galore?"). Therefore, it may be worth removing these tweets from the dataset to improve the quality of the classification. Needless to say, this operation cannot be done manually, but it must be done in an automatic way by the system (i.e., the system has to understand when a tweet is difficult to classify). This identification and removal procedure can be based on inconsistent calls. Indeed, currently the classification is based on the agreement of a number of players (2 in our case) and does not consider inconsistencies. For example, it may happen that a tweet is classified as "positive" by player A, as "negative" by player B, and as "positive" by player C. Since two players agree on the same call, then the tweet is classified. The call of player B is not considered and, through the experimental phase, we observed that this procedure can lead to bad classifications.

To improve the quality of the classification, the system might consider using the number of inconsistent calls that a tweet receives. For example, for each inconsistency the call-threshold for that specific tweet might be increased by one. Therefore, if there is an inconsistency, the agreement of two players is not sufficient, but it is necessary to have the agreement of three players. If there are two inconsistencies, then the call-threshold rises to four. Moreover, if the number of inconsistencies goes beyond a predefined threshold, then the tweet might be removed from the dataset because it is considered too difficult to classify.

6 Conclusions

In this paper, we proposed to gamify the sentiment analysis of tweets. In particular, the game aimed to entertain people by asking them to classifying the polarity (e.g., positive, negative, neutral) and the sentiment (e.g., joy, surprise, sadness, etc.) of tweets. To evaluate the proposal, we collected and filtered a dataset of tweets written in Italian language. Then, we developed a Web-based game and we invited people to play the game. After 30 days, we performed two different analysis in order to understand the effectiveness of our proposal. The engagement analysis showed that participants liked the game since they played it at every hour of the day, on normal weekdays and also on weekends. The validation analysis showed the effectiveness of the

game approach since the players' classification was consistent for both the ground-truth (88.8% for the polarity and 83.2% for the sentiment) and for the manual assessment (96.6% for the polarity and 88.3% for the sentiment). These numbers show that the game approach is an interesting methodology to use when dealing with tweets sentiment analysis. Indeed, it might provide interesting insights about the sentiment of people when automatic techniques are not available or when these techniques do not achieve interesting results, and it can facilitate the creation of a lexicon reference corpus for any language, that, in turn, might be used in several frameworks for automatically detecting sentiments from big data sources.

Acknowledgments

Authors would like to thank the students who played the game, the ones who performed the ground-truth and the manual evaluation. Authors wish also to thank the anonymous referees that, with their comments, helped us to improve the readability and the quality of the paper.

References

1. Ben A. Amaba. Industrial and business systems for smart cities. In *Proceedings of the International Workshop on Emerging Multimedia Applications and Services for Smart Cities*, pages 21–22, 2014.
2. Y. Seki. Use of twitter for analysis of public sentiment for improvement of local government service. In *Proceedings of IEEE International Conference on Smart Computing*, pages 1–3, May 2016.
3. James W Ainsworth. Why does it take a village? the mediation of neighborhood effects on educational achievement. *Social Forces*, 81(1):117–152, 2002.
4. George Galster. The mechanism(s) of neighbourhood effects: Theory, evidence and policy implications. In *Neighbourhood effects research: New perspectives*, pages 23–56. Springer, 2012.
5. Gallup-Healthways. Well-being index. Technical report.
6. M. Montangero and M. Furini. Trank: Ranking twitter users according to specific topics. In *Proceedings of the IEEE Consumer Communications and Networking Conference*, pages 767–772, Jan 2015.
7. S. Sahu, S. K. Rout, and D. Mohanty. Twitter sentiment analysis – a more enhanced way of classification and scoring. In *Proceedings of IEEE International Symposium on Nanoelectronic and Information Systems*, pages 67–72, Dec 2015.
8. Marco Furini and Valentina Tamanini. Location privacy and public metadata in social media platforms: attitudes, behaviors and opinions. *Multimedia Tools and Applications*, 74(21):9795–9825, 2015.
9. Marco Furini. Users behavior in location-aware services: Digital natives vs digital immigrants. *Advances in Human-Computer Interaction*, 2014, 2014.
10. Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. Irony detection in twitter: The role of affective content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24, July 2016.
11. Armir Bujari, Marco Furini, and Nicolas Lainà. On Using Cashtags to Predict Companies Stock Trends. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, Jan 2017.
12. Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 312–320, June 2013.
13. Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the Extended Semantic Web Conference*, pages 93–98, 2011.
14. Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervs. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, may 2012.
15. Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. In *Proceedings of the Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI*, December 2013.
16. C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, and E. Sulis. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicittá. In *Proc. of the 5th Workshop on Emotion, Social, Signals, Sentiment & Linked Open Data*, May 2014.
17. Erik Tjong Kim Sang. Using tweets for assigning sentiments to regions. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*, May 2014.
18. Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63, March 2013.
19. Marco Furini. Mobile games: What to expext in the near future. In *Proceedings of GAMEON Conference on Simulation and AI in Computer Games*. EuroSis Society, November 2007.
20. Ben Kirman, Staffan Björk, Sebastian Deterding, Janne Paavilainen, and Valentina Rao. Social game studies at CHI 2011. In *Proceedings of Human Factors in Computing Systems*, pages 17–20, 2011.
21. Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, August 2008.
22. M. Furini and M. Montangero. Tsentiment: On gamifying twitter sentiment analysis. In *Proceedings of IEEE Symposium on Computers and Communication*, pages 91–96, June 2016.
23. W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
24. Magda B Arnold. *Emotion and personality*. Columbia University Press, 1960.
25. Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, May 1992.
26. R. Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper and Row, New York, 1980 1980.
27. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of*

the International Language Resources and Evaluation Conference, volume 10, pages 2200–2204, 2010.
28. Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):1–15, 05 2013.
29. Yu-Ru Lin. Assessing sentiment segregation in urban communities. In *Proceedings of the International Conference on Social Computing*, pages 9:1–9:8, 2014.
30. Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4647–4657, New York, NY, USA, 2016. ACM.
31. William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. *CoRR*, abs/1606.02820, 2016.
32. Ozkan Aslan, Serkan Gunal, and B. Taner Dincer. A computational morphological lexicon for turkish: Trlex. *Lingua*, pages –, 2018.
33. Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.
34. Marco Furini. Vimood: using social emotions to improve video indexing. In *Proceedings of the 12th International IEEE Consumer Communications and Networking Conference*, Jan 2015.
35. Marco Furini. On gamifying the transcription of digital video lectures. *Entertainment Computing*, 14:23 – 31, 2016.
36. C. E. Palazzi, M. Roccetti, and G. Marfia. Realizing the unexploited potential of games on serious challenges. *Comput. Entertain.*, 8(4):23:1–23:4, December 2010.
37. Roberta De Michele and Marco Furini. Understanding the city to make it smart. In *Internet of Things*, volume 169 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer International Publishing, 2016.
38. Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004.
39. Mathias Lux, Mario Guggenberger, and Michael Riegler. Picturesort: Gamification of image ranking. In *Proceedings of the International Workshop on Gamification for Information Retrieval*, pages 57–60, 2014.
40. Hernisa Kacorri, Kaoru Shinkawa, and Shin Saito. Introducing game elements in crowdsourced video captioning by non-experts. In *Proceedings of the Web for All Conference*, pages 29:1–29:4, New York, NY, USA, 2014. ACM.
41. Nicolas Kaufmann, Thimo Schulze, and Daniel J. Veit. More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk. In *Proceedings of the Americas Conference on Information Systems*, 2011.
42. Haichao Zheng, Dahui Li, and Wenhua Hou. Task design, motivation, and participation in crowdsourcing contests. *Int. Journal of Electronic Commerce*, 15(4):57–88, 2011.
43. Mokter Hossain. Users' motivation to participate in online crowdsourcing platforms. In *Proceedings of the International Conference on Innovation Management and Technology Research*, pages 310–315, 2012.