"Buon appetito!" - Analyzing Happiness in Italian Tweets

Pierpaolo Basile and **Nicole Novielli**
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.lastname}@uniba.it

Abstract

English. We report the results of an exploratory study aimed at investigating the language of happiness in Italian tweets. Specifically, we conduct a time-wise analysis of the happiness load of tweets by leveraging a lexicon of happiness extracted from 8.6M tweets. Furthermore, we report the results of a statistical linguistic analysis aimed at extracting the most frequent concepts associated with the happy and sad words in our lexicon.

Italiano. *Riportiamo i risultati dell'analisi esplorativa di un corpus di tweet in Italiano, al fine di individuare i concetti tipicamente associati alla felicità. Riportiamo inoltre i risultati di un'analisi time-wise dell'happiness load dei tweet nelle diverse ore della giornata e nei diversi giorni della settimana.*

1 Introduction

The widespread diffusion of social media has reshaped the way we interact and communicate. Among others, microblogging platforms as Twitter are becoming extremely popular and people constantly use them for sharing opinions about facts of public interest. Furthermore, its worldwide adoption and the fact that tweets are publicly available, makes Twitter an extremely appealing virtual place for researchers interested in language analysis as a mean to investigate social phenomena (Bollen et al., 2009; Garimella et al., 2016).

In addition, recent research showed how microblogging is also used for self-disclosure of individual feelings (Roberts et al., 2012; Andalibi et al., 2017). As such, microblogs constitute an invaluable wealth of data ready to be mined for discovering affective stereotypes (Joseph et al., 2017) using corpus-based approaches to linguistic ethnography (Mihalcea and Liu, 2006). Such analyses, can further enhance our understanding on how people conceptualize the experience of emotions and what are their more common triggers. Recent studies even envisaged the emergence of tools for monitoring the public mood [1] and health through the analysis of Twitter users' reaction to major social, political, economics events (Bollen et al., 2009).

In this study we report the results of an exploratory analysis of the language of happiness in Twitter. In particular, we perform a partial replication of the approach proposed by (Mihalcea and Liu, 2006) for mining sources of happiness in blog posts. The contributions of this paper are as follows. First, we extract a happiness dictionary from a sample of about 8.6M tweets from the TWITA corpus of Italian tweets (Basile and Nissim, 2013). For each word in the dictionary, we compute a *happiness factor* by adapting the approach proposed in the original study. Furthermore, we perform a qualitative investigation of the 100 happiest and saddest words by mapping them into psycholinguistic word categories (see Section 2). As a second step, we use our dictionary to perform a time-wise analysis of happiness as shared in different hours and days of the week (see Section 3). Third, we extract concepts most frequently associated with happy words in our dictionary, which we map into WordNet super-senses (see Section 4). We discuss limitations and provide suggestions for future work in Section 5.

2 The Happiness Dictionary

2.1 A Dataset of Happy and Sad Tweets

Our study is based on TWITA (Basile and Nissim, 2013), the largest available corpus of Ital-

[1] 'What Twitter tells us about our happiness' https://goo.gl/fmYBP3 - Last accessed: Oct. 2018

ian tweets. In particular, we analyze a subset of 400M tweets obtained by filtering-out re-tweets from all the 500M tweets collected from February 2012 to September 2015. Following the idea proposed in (Read, 2005; Go et al., 2009), we select positive and negative tweets based on the presence of positive or negative emoticons[2]. Since a tweet can contain multiple emoticons, we selected only tweets that contain a single emoticon appearing at the end of the tweet. Using this procedure we obtain a corpus C_{happy} of 8,648,476 tweets.

2.2 Happy/Sad Word Extraction and Scoring

From the C_{happy} corpus, we extract a subset of words and we assign them an happiness factor (hf) computed according to the log of the odds ratio between the probability that the word occurs in positive tweets $p_{happy}(w_i)$ and the probability that it occurs in negative tweets $p_{sad}(w_i)$ as in Eq. 1.

$$hf(w_i) = log\frac{p_{happy}(w_i)}{p_{sad}(w_i)} \quad (1)$$

We adopt additive smoothing (Laplace smoothing) for computing both p_{happy} and p_{sad} probabilities. In our lexicon, we include and compute the happiness factor only for words that occur at least 10,000 times, for a total of 718 words. We call this list "the happiness dictionary" (D_h)[3]. Table 1 reports the most happy/sad words with the corresponding happiness factor (score(hf)).

Table 1: The happiness factor of the most happy/sad words.

happy	score (hf)	sad	score (hf)
fback	4.04	triste	-2.37
ricambi	3.83	purtroppo	-1.91
benvenuta	3.17	dispiace	-1.68
grazie	2.32	brutto	-1.68
buon	2.14	peccato	-1.63
piacere	2.03	manca	-1.53
gentile	1.91	compiti	-1.35
auguro	1.86	paura	-1.33
dolcezza	1.74	studiare	-1.30

We observe that some happy words *(fback, ricambi, benvenuta)* are due to several positive tweets that users post when they establish new connections, i.e. when they start following a new user or when they ask sombebody to follow them back *(fback)* as in: *@usermention ciao sono nuova, fback? Grazie mille :)* Sad words refer to negative emotions or evaluations, such as *triste*, *dispiace*, *brutto*, *peccato*. Interestingly, several negative words emerge from the school domain (*compiti*, *studiare*) and the word *scuola* has a negative score of -0.93 itself.

2.3 Happiness by Psycholinguistic Categories

We are interested in understanding how happiness words map into psycholinguistic word classes. Hence, we check their distribution along the word categories in the Linguistic Inquiry and Word Count (LIWC) taxonomy (Pennebaker and Francis, 2001). To this aim, we perform a qualitative investigation on the 100 most happy and 100 most sad words, that are the words with the highest and lowest happiness scores, respectively. We map each word into LIWC word categories. LIWC organizes words into psychologically meaningful categories, based on the assumption that the language reflects the cognitive and emotional phenomena involved in communication. It has been used for a wide range of psycholinguistics experimental settings, including investigation on emotions, social relationships, and thinking styles (Tausczik and Pennebaker, 2010).

We perform a coding of the English translation of the happy/sad words into LIWC categories. When translating, we keep the information about the subject conveyed by the Italian verbs (e.g., 'penso' is translated as 'I think'). The coding is performed manually by the authors: in a first round, one rater associates each word with the corresponding LIWC category; then, the other revises the annotation, checking for consistency and verifying also the correctness of the translation. 22 words are discarded and replaced with others from the dictionary because we could not find a mapping with any of the categories. Furthermore, we add an *ad hoc* category to enable modeling of words from the social media domain (*retweet*, *follow*).

Figure 1 shows how the happy and sad words distribute along the dimensions associated with the most frequent categories. Sample words for each word category are reported in Table 2. We observe that happy words in the dictionary mainly refer to positive emotions as well as to the social and social media dimensions. Conversely, sad words mainly

[2] We use :-) and :) for happy and :-(and :(for sad.

[3] The dictionary is available on github `https://github.com/pippokill/happyFactor`

describe negative emotions with focus on the author. Words describing cognitive mechanisms are also associated with sadness.

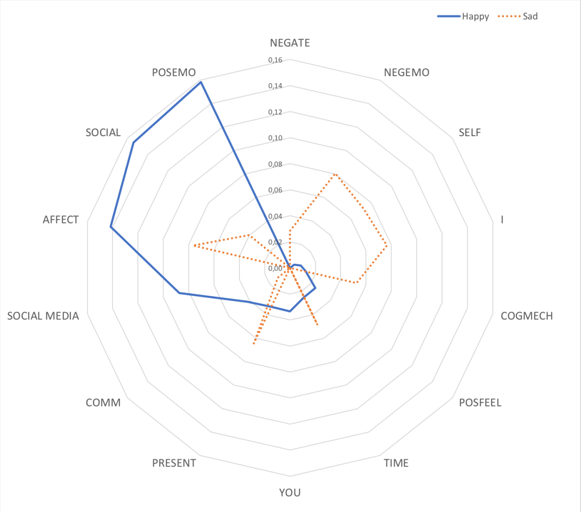


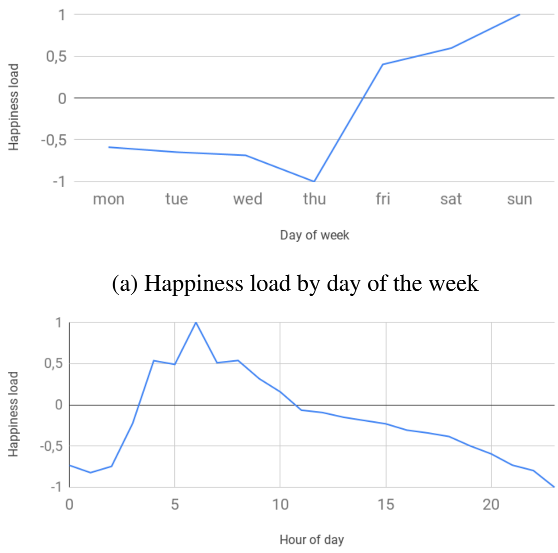
Figure 1: Comparing the most happy/sad words along dimensions associated with word categories.

Table 2: Mapping the happiness dictionary to word categories

Category	Sample words
Affect	buono/a, ottimo, triste, brutto
Cogmech	avrei, pensare, capisco, so, volevo
Comm	benvenut*, buonanotte, ciao
I	mi, io, *first person verbs*
Negate	mai, nulla, non
Negemo	difficile, peggio, sola
Posemo	benvenuta, piacere, sorriso, cara
Posfeel	cara, contenta, adoro, felice
Present	avermi, trovi, riesco
Self	mi, io, *first person verbs*
Social	ricambi, gruppo
S. media	fback, follow, seguire, Instagram
Time	serata, anticipo, periodo, ultima
You	te, tuo, *second person verbs*

3 Time-wise analysis

As observed in the original study, happiness is not constant in our life and different degrees of happiness might be observed at different moments in time. As such, we analyze how happiness changes over time. In particular we take into account the days of the week and the different hours in a day. For this analysis, we exploit the whole corpus of 400M tweets and we compute the distribution of words occurring in the happiness dictionary in each different time period. Using this strategy, in each time period the word has an happiness load obtained by multiplying its frequency in that period by its happiness factor. The happiness load of each time period is the average of all the happiness load in that period. The obtained values are mapped in the interval [-1, 1] and plotted in Figure 2a (for days) and in Figure 2b (for hours).



(a) Happiness load by day of the week

(b) Happiness load for a 24-hour day

Figure 2: Time-wise analysis.

Our time-wise analysis reveals a drop in happiness on Thurdsay, with a subsequent twist towards positive mood on Friday, before the weekend that is the happiest moment in the week. This is consistent with the findings of the original study reporting mid-week blues around Wednesday and a happiness peak on Saturday (Mihalcea and Liu, 2006). Regarding the hours, we observe the highest happiness load in the morning, with a peak around 6 AM, and it constantly decreases over the day, with the lowest value observed around 11 PM.

4 Concept analysis

We are interested in concepts related to words in the happiness dictionary. In the original study, the authors extract the 'ingredients' for their recipe of happiness by ranking the most relevant 2- and 3-grams from their corpus according to their happiness load. Such an approach is not easy to replicate as the number of 2- and 3-grams extracted from 400M tweets is potentially huge. Hence, starting from the words in our happiness dictio-

Table 3: The most happy and sad word pairs.

	word pair	score
happy	buon, appetito	9.74
	buon, auspicio	8.84
	dolcezza, infinita	6.94
	grazie, mille	5.23
	piacere, ciao	5.12
	grazie, esistere	4.50
sad	dispiacere, deludervi	-9.28
	brutto, presentimento	-8.45
	triste, arrabbiata	-8.10
	peccato, potevamo	-4.85
	triste, piangere	-3.68
	studiare, matematica	-3.55
	peccato, gola	-2.63
	manca, vederlo	-1.97

nary, we extract the most 50 co-occurring words in a window of two words. Then we rank all the word pairs (dictionary word, co-occurring word)[4] according to the Pointwise Mutual Information (PMI) multiplied by the happiness factor. Table 3 reports some of the most happy and sad pairs.

Starting from word pairs, we perform another kind of analysis aiming at mapping the words occurring in each pair with super-senses in WordNet. A super-sense is a general semantic taxonomy defined by the WordNet lexicographer classes as a way for defining logical aggregation of senses in each syntactic category. We assign a happiness score to each super-sense by averaging the happiness factor associated with the dictionary word in the pair. Since each pair contains a dictionary word and a co-occurring word, we map the co-occurring word to its super-sense and increment the score of the super-sense by summing the happiness factor associated with the dictionary word. Finally, the score of each super-sense is divided by the number of the co-occurring words belonging to the super-sense. For ambiguous words, we select the super-sense associated with the most frequent sense. In this study, we do not rely on a Word Sense Disambiguation (WSD) algorithm since WSD is a critical task. We need to test the WSD performance on tweets before to use it. Generally, WSD algorithms give performance slightly above the most frequent sense. We plan to test WSD in a further study. As super-senses are defined in the English version of WordNet, we performed a mapping of Italian words to the English WordNet through the use of both Morph-it! (Zanchetta and Baroni, 2005) and MultiWordNet (Pianta et al., 2002), while sense occurrences are extracted from MultiSemCor (Bentivogli and Pianta, 2005).

In Table 4 we report the most happy and sad super-senses with the most frequent words extracted by our corpus. Consistently with the evidence provided by the analysis of the psycholinguistic word categories (see Section 2.3), we observe that socialness is associated with positive feelings, with concepts referring to people (*noun.person*) and communication (*verb.communication*, *noun.communication*) scoring high in happiness. Food (*noun.food*) also seems to be a major cause of positive mood, as well as money and gifts (*noun.possession*), sport achievements (*’vittoria* and *’gol’* in *noun.act*), and mundane locations and events (*’centro’, ’piazza’, ’concerto’, ’viaggio’* in *noun.location* and *noun.act*). This is consistent with suggestion by (Mihalcea and Liu, 2006) to enjoy food and drinks in an ’interesting social place’ as a recipe for happiness. People also report their desires and preferences (*voglio, amo, spero* in *verb.emotion*).

Also for sadness, results confirm findings emerging from the analysis of psycholinguistic categories in LIWC. In fact, we observe that people tend to report their own individual negative feelings (*rido, piango* in *verb.body*), thoughts (*verb.cognition*), perceptions (e.g., *’vedo’, ’sento’*), and personal needs (*’bisogno’* and *’sonno’* in *noun.state*). We observe also stereotypical complaints about weather (*piove*) as well as swear words (*noun.body*).

5 Discussion and Conclusions

We performed an exploratory analysis of the lexicon and concepts associated with happiness in Italian tweets. We leveraged a corpus of happy and sad tweets to extract a ”happiness dictionary’, which we use to perform a time-wise analysis of happiness on Twitter and to extract the most frequent concepts and psycholinguistic categories associated to positive and negative emotions.

This study is a partial replication of the previous one by (Mihalcea and Liu, 2006) on blog posts. The main differences with respect to the original study are in the size, language and source of the corpus used for extracting the happiness

[4] We do not take into account the word order in the pairs.

Table 4: The most happy and sad super-senses based in our corpus.

	super-sense	**most frequent concepts**
happy	noun.relation	resto, ricambio
	noun.food	cena, pranzo, colazione, caffé
	noun.attribute	coraggio, voce, numero, bellezza, splendore, silenzio
	noun.person	mamma, ragazz*, amic*, dio, tesoro, donna
	verb.communication	dico(no), parlare, prego, profilo, parla, chiedere
	noun.communication	film, scusa, merda, musica, buongiorno, canzone, concerto
	verb.possession	trov*, dare, perdere, perso, averti, comprato
	verb.emotion	voglio/vorrei, amo, piace, vuoi, spero, odio, auguri
	noun.location	sito, centro, post, piazza, scena, sud, nord, regione
	noun.possession	soldi, regalo, fondo
	noun.event	vittoria, gara, onda, campagna, scarica, fuoco, episodio, meraviglia
	noun.act	cose, partita, gol, colpa, ricerca, viaggio, tour, bacio, corso, sesso
sad	verb.consumption	bisogna, mangiare, usare, mangio/mangiato, usa/o, usato, mangio
	verb.body	piangere, dormire, ridere, sveglia, sorridere, piango, rido
	noun.body	*swear words*, testa, occhi, mano/i, capelli
	verb.change	inizio/inizia(re), cambiare, finito, morire/morte, successo, finisce
	verb.perception	vedere, vedo, sento, sentire, guarda, guardare, ascoltare, pare
	verb.cognition	so, sai, penso, letto, credo, sa, leggere, sapere, pensare, studiare
	noun.state	bisogno, punto, problemi/a, accordo, pace, crisi, situazione, sonno
	noun.substance	aria, acqua
	verb.weather	piove

lexicon. Specifically, (Mihalcea and Liu, 2006) rely on a collection of 10,000 blog posts in English from LiveJournal.com to extract a list of happy/sad words with their associated happiness scores, while we leverage a bigger corpus consisting of 8.6M Italian tweets. Furthermore, the blog posts were labeled as happy or sad by their authors. Conversely, for tweets we relied on silver labeling based on the presence of emoticons as a proxy the author self-reporting of her own positive or negative emotions.

Our analysis of psycholinguistic categories and the extraction of concepts and WordNet supersenses associated with them reveals interesting findings. Happiness appears related to the social aspects of life while sad tweets mainly revolves around self-centered negative feelings and thoughts. In addition, our-time wise analysis reveals a mid-week drop in happiness also observed in the original study. We also observe that happiness is high in the morning and decreases over the day. As a future work, it would be interesting to investigate if time-wise analysis based on hours produces consistent results if a weekday or the weekend is considered and if emotion-triggering concepts associated with happiness also vary over time.

We are aware of the main limitations of this study. First of all, by relying on microblogs we are probably able to mine emotion triggers that do not necessarily coincide with those shared in daily face-to-face conversations or reported in private logs. Furthermore, we do not attempt to make any categorization of the authors of tweets. Indeed, different target user groups could be studied to fulfill specific research goals and enable perspective applications, i.e. for supporting creative writing or for providing personalized recommendations based on moods. Finally, we consider only Twitter as a source of data. The same methodology could produce different results if applied to other social media. Indeed, recent research (Andalibi et al., 2017) showed that other media, such as Instagram, are also used for sharing extremely private emotions, such as feelings linked to depression. Based on these observations, further replications could focus on finer-grained emotions, also leveraging corpora from different platforms and including consideration of demographics and geographical information (Mitchell et al., 2013; Allisio et al., 2013) as additional dimensions of analysis.

References

[Allisio et al.2013] Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, and Giancarlo Ruffo. 2013. Felicittà: Visualizing and estimating happiness in italian cities from geotagged tweets. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 3, 2013.*, pages 95–106.

[Andalibi et al.2017] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive self-disclosures, responses, and social support on instagram: The case of #depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1485–1500, New York, NY, USA. ACM.

[Basile and Nissim2013] Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

[Bentivogli and Pianta2005] Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

[Bollen et al.2009] Johan Bollen, Alberto Pepe, and Huina Mao. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583.

[Garimella et al.2016] Kiran Garimella, Michael Mathioudakis, Gianmarco De Francisci Morales, and Aristides Gionis. 2016. Exploring controversy in twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW '16 Companion, pages 33–36, New York, NY, USA. ACM.

[Go et al.2009] Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. *Entropy*, 17:252.

[Joseph et al.2017] Kenneth Joseph, Wei Wei, and Kathleen M. Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from twitter. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 1362–1374.

[Mihalcea and Liu2006] Rada Mihalcea and Hugo Liu. 2006. A corpus-based approach to finding happiness. In *Proc. AAAI Spring Symposium and Computational Approaches to Weblogs*, page 6 pages.

[Mitchell et al.2013] Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5):1–15, 05.

[Pennebaker and Francis2001] J. Pennebaker and M. Francis. 2001. Linguistic inquiry and word count: Liwc. *Mahway: Lawrence Erlbaum Associates*, 71.

[Pianta et al.2002] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. 1st gwc. In *Proceedings of the First International Conference on Global WordNet*.

[Read2005] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43–48. Association for Computational Linguistics.

[Roberts et al.2012] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and Detecting Emotions on Twitter. In Nicoletta C. Chair, Khalid Choukri, Thierry Declerck, Mehmet U. Dou gan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Tausczik and Pennebaker2010] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

[Zanchetta and Baroni2005] Eros Zanchetta and Marco Baroni. 2005. Morph-it!: a free corpus-based morphological resource for the italian language.