Combining words, emoticons and emojis to measure sentiment in Italian tweet speeches

Livia Celardo[1], Domenica Fioredistella Iezzi[2]

[1]University of Rome Tor Vergata – livia.celardo@uniroma2.it

[2]University of Rome Tor Vergata – stella.iezzi@uniroma2.it

Abstract 1

Sentiment Analysis (SA) is a supervised classification technique used to measure the feelings and the opinions in a text. In SA there are two main approaches: the lexicon-based and the machine learning techniques. In the lexicon-based approach the SA is implemented starting from a dictionary of terms, where words are classified on the basis of their polarity. In this paper, we implement a SA constructing on the Italian language a dictionary of words, emojis and emoticons, classified accordingly to their polarity (negative or positive). To create this external resource, we started from two different dictionaries for sentiment analysis, including in addition to words also emojis and emoticons. The idea is to take into consideration when analysing the sentiment of texts both the semantic orientation of words and expressions like emoticons or emojis. We used this dictionary on a corpus composed of thousands of micro-messages from Twitter related to Greta Thunberg, in order to measure its sentiment.

Keywords: Sentiment Analysis; Twitter; Text Mining; Social Network Analysis.

Abstract 2

La Sentiment Analysis (SA) è un metodo di classificazione supervisionato usato per misurare il sentimento e le opinioni di un testo. Nella SA esistono due approcci: quello basato sul lessico e quello basato sull'apprendimento. Nell'approccio lessicale la SA è implementata a partire da un dizionario, dove le parole sono classificate sulla base del loro orientamento. In questo articolo, abbiamo implementato una SA costruendo sulla lingua italiana un dizionario composto da parole, emojis e emoticons, classificati secondo la loro polarità (positiva o negativa). La creazione di questa risorsa esterna è stata fatta a partire da due differenti dizionari creati per la SA, aggiungendo oltre alle parole anche le emojis e le emoticons. L'idea è quella di analizzare il sentimento prendendo in considerazione sia l'orientamento delle parole che quello delle emojis/emoticons. Abbiamo usato questo dizionario su un corpus composto da migliaia di micro-messaggi provenienti da Twitter riguardanti Greta Thunberg, al fine di misurare il loro orientamento.

Parole chiave: Sentiment Analysis; Twitter; Text Mining; Social Media Analytics.

1. Introduction

Over the course of last decade, the popularity of online communication through social media technology has evolved significantly (Best et al., 2014). In particular, microblogging – such as Twitter, Tumblr, Facebook and others – today has become an extremely popular communication tool among Internet users, who tend more and more to shift from traditional communication tools to microblogging services; users write on these media about their life, share opinions on variety of topics and discuss current issues, so microblogging web-sites has become valuable sources of people’s opinions and sentiments (Pak & Paroubek, 2010; Agarwal et al., 2011). According to Liu (2010), “textual information in the world can be broadly categorized into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people’s sentiments, appraisals or feelings toward entities, events and their properties”. In our text analytics, we focus attention on the “positive”, “negative” and “neutral” groups of words, emojis, and emoticons. The computational study of these opinions and emotions is known as *sentiment analysis* or *opinion mining*; with the huge growth of social media platforms on the Web, individuals and organizations are increasingly using public opinions in these media (Liu & Zhang, 2012) and Sentiment Analysis (SA) has become a broad and complex field of research. Its aim is to define automatic tools able to extract subjective information from texts in natural language, so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker (Pozzi et al, 2017). The first step of SA is usually involving in discriminating between objective and subjective sentences. When a sentence is categorized as objective, we do not proceed with SA, i.e. this is a pen; while if the sentence is classified as subjective, its polarity (positive, negative, or neutral) is measured, i.e. this is a great news (positive polarity). Social media, like Twitter, allow continuous and real-time monitoring of users' mood. Twitter users can write messages of up to 280 characters called “tweets”, with a more varied content, e.g. mood predicts the stock market, consumption, political addresses and much more (Bollen *et al*, 2010; Ceron *et al*, 2015). In any context, tweets are used, whether it is journalistic or scientific research or simply to communicate an opinion on a fact, the language could have a different linguistic register, and a different way of using emojis and emoticons. It currently ranks as one of the leading social networks worldwide based on active users. Twitter now has 126 million daily users, up from 115 million one year ago. Users communicate in a language made not only of words, but also extra-verbal components, like emoticons and emojis. The emoticons or smileys are stylized reproductions of those main human facial expressions that define an emotion made using combinations of characters while the emojis are pictographic symbols, very similar to emoticons that are used to express at best emotions and sentiments. In the age of the “electronic global village” where people of different national languages and cultures are in frequent contact through online interactions, the emoji code might well be the universal language that can help to solve problems of comprehension that international communications have always involved in the past (Danesi, 2017). In 2015, the emoji known as “Face with Tears of Joy” was chosen by the Oxford Dictionary as the “Word of the Year”, because it is flexible, and immediate. In the literature, two main approaches to the problem of extracting sentiment automatically can be recognized (Taboada et al., 2011). The first one is the lexicon-based approach, that involves calculating orientation for a document from the semantic orientation of words or phrases in the document. The other one implicates the construction of classifiers from labelled instances of texts or sentences, in the logic of a supervised classification task.

In this paper, we present a new method to classify users’ sentiment combining words, emoticons and emojis. Emojis and emoticons are used in SA to identify intensity and ironies in tweets, by way of the relationship between words and emojis/emoticons. We tested it on a set of Italian tweets to measure moods in communication and evaluate how much takes place with words or symbols, and how these can help us to disambiguate language, identifying ironies and puns. This manuscript is structured as follows. In section no. 2, the data and method are introduced; in section no. 3, the main results are drawn and conclusions are discussed.

2. Data and methods

For this study, we collected 30.894 posts from Twitter, in the period from November the 9th to December the 17th, 2019 related to Greta Thunberg[1] in the Italian language. To extract the data, we used the R package TwitteR (Gentry, 2016). The final dataset contained the 31% of original posts and the 69% of retweets.

To the corpus we implemented SA taking into consideration, at the same time, words, emojis and emoticons. As well known, there are two common approaches to SA: (a) lexicon, and (b) learning. The lexicon approach assigns a polarity to words from a previously created dictionary. This dictionary defines a word and its polarity. The learning-based method builds an automatic sentiment classifier for a document set manually annotated, that constitutes a classifier trained to measure users' mood (Ignatow and Mihalcea 2016).

In this paper, we adopt a new lexicon-based approach, calculating for each tweet the semantic orientation and taking into consideration both words and emojis/emoticons. We started constructing a dictionary of words, emojis and emoticons based on the Italian language, classified accordingly to their polarity (negative or positive). The dictionary contains 3.205 positive forms and 4.127 negative lexical forms. To create this external resource, we used the *Italian Sentiment lexicon*, developed by the Institute for Computational Linguistics "Antonio Zampolli" together with the Italian National Research Council (CNR), and the *Madda dictionaries* from the R package TextWiller (Solari, Sciandra and Finos, 2019). Moreover, we included in this dictionary another four lists, related to negative (64) and positive (288) emojis and negative (38) and positive (50) emoticons[2]. For the polarity of emojis and emoticons, as for the words, we used a value of +1 for positive and a value of -1 for negative expressions.

Because on the Web often users use lexical forms coming from the spoken language, tweets could contain words which are not in the dictionary we constructed; so, before implementing SA we investigated the vocabulary of our corpus, paying a particular attention to those forms that are not recognized by the Italian dictionary. We found in particular two forms (*afrontosa* and *maluco*) with a high frequency that do not belong to the Italian language but to the Portuguese one, so our dictionary wasn't able to classify them. Actually, these two forms are very negative in the Portuguese language, so we decided to create a supplementary dictionary for negative forms in which we put these two words.

Then we implemented SA; according to Liu (2012), sentiment is the underlying feeling, attitude, evaluation or emotion associated with an opinion. It is represented as a triple, (*y*, *o*, *i*), where *y* is the type of the sentiment, *o* is the orientation of the sentiment, and *i* is the intensity of the sentiment. Sentiment can be classified into several types, e.g. linguistic-based, psychology-based, and consumer research–based classifications. Sentiment orientation can be positive, negative, or neutral. In this paper we implemented a sentiment analysis based on a

[1] The keywords used for the selection of the tweets were: #gretathunberg and greta+thunberg

[2] We classified the emoji (in positive or negative forms) accordingly to a sentiment analysis computed from 70,000 tweets, labeled by 83 human annotators in 13 European languages (Kralj Novak, Smailović, Sluban and Mozetič, 2015).

linguistic-based approach, and instead of using just the words we used also emojis and emoticons to express intensity of feelings or to catch several forms of irony in text. In a text, when a feeling (positive or negative) is emphasized by using emojis or emoticons (which have the same polarity of the words), then the sentiment score of the text is higher. Moreover, when there are some ironies in the messages, a way to identify them is to find whether or not words and emojis/emoticons have opposite polarity.

3. Results and conclusions

We pre-processed the corpus, obtaining a Term-Document matrix of size (15.079 × 30.894). In the pre-processing phase we lemmatized the corpus, we removed stop-words (*articles, conjunctions, prepositions, articles, pronouns, auxiliary verb*) and hapax forms. The most frequent forms[3] in the corpus – deleted the theme terms like *Greta*, *Thunberg*, and *GretaThumberg* – are the emoji *clapping hands* (4.763 times), followed by the words *person of the year*, *photo*, *Turin*, *hahaha*, *miserable*, and *crazy* (see Table 1).

Table 1 – Most frequent words

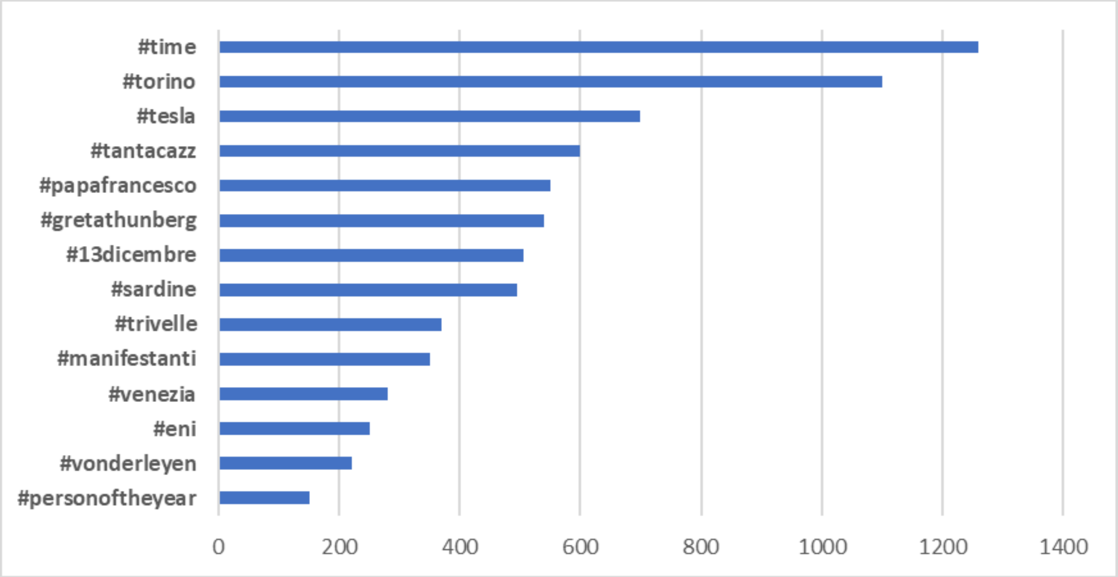
Terms	Frequency
Clapping hands (EMOJI)	4.763
Persona (PERSON)	2.444
Dell’anno (OF THE YEAR)	2.402
Foto (PHOTOGRAPH)	1.621
Torino (TURIN)	1.616
ahahah	1.521
Afrontosa (MISERABLE)	1.518
Maluco (CRAZY)	1.516

Greta Thumberg was named, in fact, for the Time magazine’s the person of the year for 2019. In Turin, December 14th, 2019 the Swedish activist gave them a speech, and probably in the summer could host the international meeting of *Fridays for Future*. An ironic expression *hahaha* follows which invokes laughter. It is evident that tweets represent a way of communicating information in real time, but facts are told not in neutral way. Anger, indignation and happiness accompany the news. Many foreign terms are used, but not only in English language; in this case, i.e. in Portuguese, the features *afrontosa* and *maluco* are retweeted many times. A language that is more similar to the spoken language than to the written one, where words, slang or emoji highlight the speaker's mood. In this study, the tweets that at least include an emoji represent about 9% of the total (74 negative, and 2666 positive). The people who choice emojis to communicate often use more than one.

[3] Both words and emojis/emoticons.

3.1 Top Hashtags

Figure 1 – The top of the hashtags



We selected the top of hashtags (#) to filter the keywords of the texts, and to compare them with the top of features, based on the frequency. We construct a document-feature matrix of 50 hashtags top tag. Figure 1 shows the top hashtags, based on threshold of frequency.

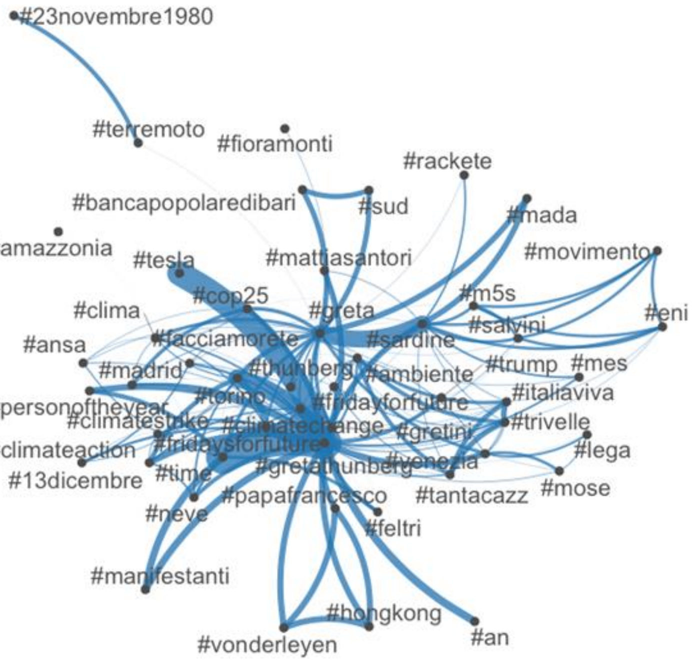
We decided to use the hashtag because they label the message based on keyword and facilitates the grouping of similar messages within the platform. The hashtag was created by the Twitter community to overcome the problem of immediately grouping users' tweets, and to follow all the conversations and messages organized within a certain topic: the hashtag makes the word a link linking all the other tweets containing that word. By clicking on the hashtag, it is possible to list all messages that mention the same keyword, which in this way "labels" tweets, categorizing and grouping them in a dedicated timeline.

Figure 2 shows the network of the 50 top hashtags; we calculated the global clustering of coefficient (transitivity, C), that is equals to 0.455 (1). The clustering coefficient of a node is the number of actual connections across the neighbors of a particular node, as a percentage of possible connections. The clustering coefficient for the entire system is the average of the clustering coefficient for each node. In this case, 45% of the top of the hashtags are linked.

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i . \qquad (1)$$

where *n*= number of nodes, *C*=clustering coefficient for each node *i*.

Figure 2 – Network of the 50 top hashtags



We calculated degree, and betweenness centralities to evaluate the popularity of the top of the hashtags (Iezzi, 2012). For degree index is #gretathunberg, #greta, #sardine, #time, #torino; for betweenness centrality is #gretathunber, #time, #cop25, #venezia, #ambiente. The correlation between degree and betweenness is not very high (0.51). Betweenness measure is better to grasp the different clusters of the debate on twitter; while lying with the degree it is not possible to identify the differences between the different classes of tweets.

3.2 Sentiment Analysis of tweets, emojis and emoticons

From SA results we found that 35% of the tweets presents a positive score considering words, emojis and emoticons, against a bigger percentage (37%) of negative posts. 9% of tweets contains emojis or emoticons in the text. We look also at the consistency between words and emojis/emoticons polarity; we found that only 6% of the tweets presents discordance between them.

On the tweets that presented respectively the highest and the lowest score in terms of sentiment – i.e. the most positive and the most negative posts – we analyzed the most frequent words (Table 2). It is interesting to look at the top words in positive and negative posts; in fact, what happened was that not always in positive tweets we found positive words – and vice-versa. For example, in negative posts we found with a high frequency the expression of having a safe trip, that is used ironically in the texts. The issue of irony would appear very frequent in Twitter communication (Carvalho et *al*., 2009; Reyes *et al*., 2012); related to our topic, we also found this feature in the discordance between words and emoji/emoticon lists.

Table 2 – Most frequent words in positive and negative tweets

Most frequent words in positive posts	Most frequent words in negative posts
Credere (TO BELIEVE)	Lavoro (WORK)
Mondo (WORLD)	Emissioni (EMISSIONS)
Migliore (BEST)	Buon viaggio (TO HAVE A SAFE TRIP)
Combattere (TO FIGHT)	Tesla (*Car brand*)
Bella ciao (*Famous Italian song*)	Protesta (PROTEST)
Sardine (*Italian movement*)	Torino (TURIN)
Cantare (TO SING)	Flop
Fenomeno (PHENOMENON)	Semivuota (HALF-EMPTY)

From the results we found that, related to the subject we chose, the sentiment in the communication is predominant; there are few facts and many opinions, that in most of the case are politicized. It is well-known that Twitter has become more and more the environment for sharing political debates. Over the past 10 years, Twitter has indeed played a crucial role, e.g., in the 2009 Iranian political elections, in the 2011 North African revolutions. You cannot forget one of the retweeted messages: "Four more years", with which Barak Obama announces his victory over Mitt Romney, accompanied by a photo of the President's embrace with his wife Michelle (Valeriani, 2013; Vasterman, 2018). Twitter has become a platform suitable for the dissemination of public information, with a great penetration in the world for the dissemination of different kinds of news. In fact, even if we expected to find themes related to climate change or global warming, we found that most of the posts were connected to the discussion about the Italian political parties.

Related to the Twitter language, our analyses gave us many causes for reflection. First of all, the continuous evolution in the vocabularies and the systematic creation of new words doesn't allow the use of a static dictionary; we overcome this issue by implementing an exploratory analysis before the SA in which we identified specific words not included in our vocabulary. So, we suggest to customize and update the vocabulary every time an analysis is implemented on specific themes. Moreover, emojis and emoticons represent a significant element of the language; we showed that they could be used to identify ironies and emphases in texts, which are very important in speeches and then they should not be ignored.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. Children and Youth Services Review, 41, 27-36.

Bollen J., Mao H., Xiao-Jun Z. (2010). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*.

Carvalho C., Sarmento L., Silva M. J., and de Oliveira E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's so easy" ;-). In *1st CIKM WS on Topic-sentiment Analysis for Mass Opinion*, pages 53–56. ACM.

Ceron A, Curini L, Iacus SM (2015) Using social media to forecast electoral results: A review of state-of-the-art. *Statistica Applicata - Italian Journal of Applied Statistics* 25(3):239{261

Danesi M. (2017) *The Semiotics of Emoji. The Rise of Visual Language in the Age of the Internet.* Bloomsbury Academic.

Gentry J. (2016). R Based Twitter Client. R package version 1.1.9.

Ignatow, G., & Mihalcea, R. (2016). Text mining: A guidebook for the social sciences. Los Angeles: Sage Publications.

Iezzi D. F. (2012). Centrality measures for text clustering. Communications in Statistics-Theory and Methods, 41(16-17), 3179-3197.

Kralj Novak P., Smailović J., Sluban B., Mozetič I. (2015). Sentiment of Emojis. PLoS ONE 10(12): e0144296.

Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627-666.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

Pozzi F.A., Fersini E., Messina E., Liu B. (2017) Chapter 1 - Challenges of Sentiment Analysis in Social Networks: An Overview, Editor(s): Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu, Sentiment Analysis in Social Networks, Morgan Kaufmann: 1-11.

Reyes A., Rosso P., and Buscaldi D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*, 74(0):1 – 12.

Solari, D., Sciandra, A., & Finos, L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the italian language. Journal of Open Source Software, 4(41), 1256.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Valeriani A. (2013). *Twitter factor: come I nuovi media cambiano la politica internazionale*. Laterza & figli. Bari

Vasterman P. (2018). *From Media Hype to Twitter Storm. News Explosions and Their Impact on issues, Crises, and Public Opinion.* Amsterdam University Press. Amsterdam.