FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis

Lucia C. Passaro, Alessandro Bondielli and **Alessandro Lenci**
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa (Italy)
lucia.passaro@for.unipi.it
alessandro.bondielli@gmail.com
alessandro.lenci@unipi.it

Abstract

English. In this paper we present the FB-NEWS15 corpus, a new Italian resource for sentiment analysis and emotion detection. The corpus has been built by crawling the Facebook pages of the most important newspapers in Italy and it has been organized into topics using LDA. In this work we provide a preliminary analysis of the corpus, including the most debated news in 2015.

Italiano. *In questo lavoro presentiamo il corpus FB- NEWS15, un corpus italiano creato per scopi di sentiment analysis ed emotion detection. Il corpus stato costruito scaricando le pagine Facebook delle maggiori testate giornalistiche in Italia e successivamente organizzato in topic utilizzando LDA. In questo articolo forniamo una analisi preliminare del corpus, e mostriamo le notizie pi discusse nel 2015.*

1 Introduction

The use of Social Networks (SN) platforms like Facebook and Twitter has developed overwhelmingly in recent years. SN are exploited for different purposes ranging from the sharing of contents among friends and useful contacts to the news-gathering about different domains such as politics and sports (Ahmad, 2010; Ahmad, 2013; Sheffer and Schultz, 2010). Many journalists indeed use SN platforms for professional reasons (Oriella, 2013; Hermida, 2013).

Several recent studies provide insights on how the popularity of blogs and other user generated content impacted the way in which news are consumed and reported. Picard (2009) states that SN platforms provide an easy and affordable way to take part in discussions with larger groups of people and, consequently, the bond between SN and information is becoming increasingly stronger.

Mass information is gradually moving towards general platforms, and official websites are losing their lead position in providing information. As noted by Newman et al. (2012), even though the use of internet in the years 2009-2012 has grown, the same is not reflected in the consumption of online newspapers, probably because of the increasing use of SN for news diffusion and gathering. If on the one hand this apparent decline of the traditional news platforms may lead to a decline in quality and news coverage (Chyi and Lasorsa, 2002), on the other hand the rise of SN as platforms to spread news promotes a more fervid debate between users (Shah et al., 2005). This issue is central for the present work. In fact, user's comments very often contain their own opinions about a certain issue. In addition, because of the colloquial style of the comments, they contain large amounts of words and collocations with a high subjective content, mostly concerning the author's emotive stance.

Facebook is one of the most popular online SN in the world with 1 billion active users per month and it offers the possibility to collect data from people of different ages, educational levels and cultures. From a linguistic point of view, previous studies (Lin and Qiu, 2013) demonstrated that the language in Facebook is more emotional and interpersonal compared for example to the language in Twitter. Probably, this is due to the fact that in Facebook there is a stronger psychological closeness between the author and audience because of the different structure (bidirectional vs. unidirectional graphs) of the SNs.

In this paper we present the FB-NEWS15 corpus, a new Italian resource for sentiment analysis and emotion detection. The FB-NEWS15 corpus can be freely downloaded at `colinglab.humnet.unipi.it/resources/` under the Creative Commons Attribution License `creativecommons.org/licenses/by/2.0`.[1]

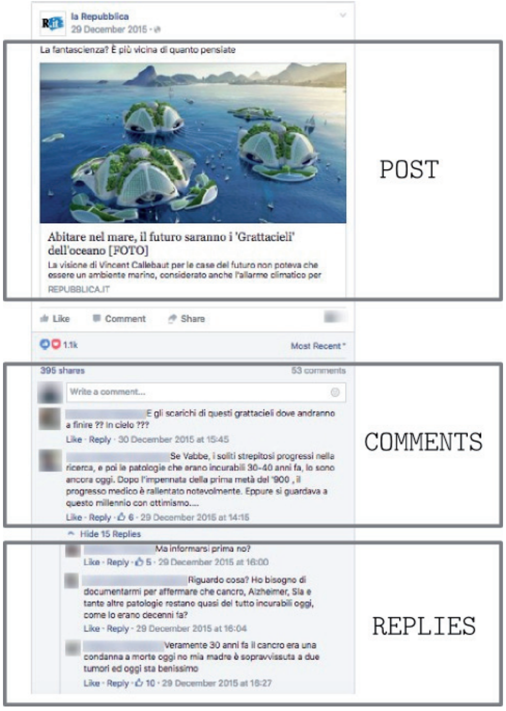
The debate among users in commenting news and posts on Facebook offers a lot of subjective material to study the way in which people express their own opinions and emotions about a target event. In fact, in FB-NEWS15 we find linguistic items expressing the whole range of positive and negative emotions. In analyzing a news corpus, however, it is not simple to aggregate the posts on the basis of a certain fact, since several posts relate to the same event. For this reason, we decided to organize the corpus into clusters of topically related news identified with Latent Dirichlet Allocation (LDA: Blei et al. (2003)). This approach allow us to infer the most debated news in the corpus, and, in a second step, to discover the readers' sentiment about a particular topic.

The paper is organized as follows: Section 2 describes the creation of the corpus, from crawling (2.1) to linguistic annotation (2.2), and finally provides basic corpus statistics (2.3). Section 3 reports on the automatic topic extraction with LDA.

2 FB-NEWS15

For the creation of the corpus we followed the most important Italian newspapers. Since we were interested in building a corpus as heterogeneous as possible, we decided to focus on major newspapers with different political orientations, and which have in general heterogeneous readers.

Facebook allow users to post states, links, photos and videos on their own wall. In general, users can be divided into two macro-categories: People and Pages. People are often individuals, and the interaction with them is usually bidirectional (user A can read what user B publishes if A and B have a *friendship relation*). Conversely, Pages are typically used to represent organizations, public figures (web stars), companies or, as in our case, newspapers. In this case, the relationship is unidirectional, in the sense that user A can access the timeline of the page P by putting a "Like" on P. Unlike a single-user, who usually publishes photos, videos and links about his private life, the timeline of a newspaper Facebook page, in general contains news titles with a link to the official website of the newspaper, where the user can read the full article. The corpus keeps tracks of the three-fold hierarchical structure of Facebook, which includes the news posts by the newspaper, the users' comments to the posts and the replies to the comments. In this context, it is clear that the emotive content of the post is often neutral, but this post can inspire long discussions among readers, which can become useful material for sentiment analysis and emotion detection. Figure 1 shows a post, with some of its comments and replies.



la Repubblica
29 December 2015

La fantascienza? È più vicina di quanto pensiate

Abitare nel mare, il futuro saranno i 'Grattacieli' dell'oceano [FOTO]

La visione di Vincent Callebaut per le case del futuro non poteva che essere un ambiente marino, considerato anche l'allarme climatico per

REPUBBLICA.IT

395 shares · 53 comments

[redacted] E gli scarichi di questi grattacieli dove andranno a finire ?? In cielo ???
Like · Reply · 30 December 2015 at 15:45

[redacted] Se Vabbe, i soliti strepitosi progressi nella ricerca, e poi le patologie che erano incurabili 30-40 anni fa, lo sono ancora oggi. Dopo l'impennata della prima metà del '900 , il progresso medico è rallentato notevolmente. Eppure si guardava a questo millennio con ottimismo....
Like · Reply · 6 · 29 December 2015 at 14:15

[redacted] Ma informarsi prima no?
Like · Reply · 5 · 29 December 2015 at 16:00

[redacted] Riguardo cosa? Ho bisogno di documentarmi per affermare che cancro, Alzheimer, Sla e tante altre patologie restano quasi del tutto incurabili oggi, come lo erano decenni fa?
Like · Reply · 29 December 2015 at 16:04

[redacted] Veramente 30 anni fa il cancro era una condanna a morte oggi no mia madre è sopravvissuta a due tumori ed oggi sta benissimo
Like · Reply · 10 · 29 December 2015 at 16:27

Figure 1: Example of post in Facebook with the relative comments and replies.

In order to create the FB-NEWS15, we decided to download the timeline of the following newspapers, from 1 January 1 to 31 December 2015: La Repubblica, Il Giornale, L'Avvenire, Libero, Il Fatto Quotidiano, Rainews24, Corriere Della Sera, Huffington Post Italia.

2.1 Crawling

Facebook offers developers Application Programming Interfaces (APIs) for creating apps with Facebook's native functionalities. In order to develop the crawler, we exploited the Graph API, which provides a simple view of the Facebook social graph by showing the objects in the graph and the connections between them. The Graph API allows us to navigate through the graph of the social network, which is organized into nodes

[1] All data collected have been processed anonymously for scientific purposes, without storing personal information.

```
<doc user="<newspaper(string)>"
 id="<id_post(string)>"
 type="post"
 parent_post=""
 parent_comment=""
 date="AAAA-MM-DD HH:MM:SS"
 location=""
 likes="662"
 comments="54"
 shares="322">

Un business truffaldino [E ora
finitela con l'eco-balla dei
controlli sulle emissioni]

</doc>
```

Figure 2: Example of crawled text.

(Users, Pages, Photos and Comments) and Edges (Connections such as Friendship or Likes). The graph is navigated by exploiting HTTP requests, that may be implemented using any programming language. The native APIs offered by Facebook has some drawbacks: i) the maintenance of the app, since the APIs change over time, making it necessary to update the code of the crawler; ii) only public data can be accessed without requiring the user's consent; iii) Facebook places limitations on the number of requests through a given period of time. For each post, comment and reply, we stored the message (text), the story (presence of photos and links tags), its timestamp, the type (post, comment, reply), the parent post/comment, the number of likes, shares and replies (Figure 2).

2.2 Linguistic annotation

A very basic preprocessing phase has been applied to the corpus before linguistic annotation, to replace urls with the tag _URL_. The text has been subsequently feed to a pipeline of general-purpose NLP tools. In particular, it has been POS-tagged with the Part-Of-Speech tagger described in (Dell'Orletta, 2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009). In addition, complex terms like *forze dell'ordine* (security force) or *toccare il fondo* (hit rock bottom) have been identified using the EXTra term extraction tool (Passaro and Lenci, 2015).

2.3 Corpus Analysis

Except for Avvenire and Rainews24, for which we downloaded very few data, the other newspapers are attested in the corpus in a balanced way. In general, the number of posts is very low compared to the number of comments and replies. The average number of posts for each newspaper is 27,341.25, while for comments and replies is respectively 2,016,243.38 and 576,498.5. Table 1shows the number of texts (including posts, comments and replies) in FB-NEWS15 for each Newspaper and Figure 2.3 shows their cumulative distribution for each Newspaper.

Newspaper	N. of Texts
La Repubblica	4558,829
Avvenire	91,824
Il Giornale	3,497,610
Libero	2,436,246
Il Fatto Quotidiano	4,900,314
Rainews24	369,834
Huffington Post	1,552,042
Corriere della Sera	3,553,966
Overall	20,960,665

Table 1: Number of texts aggregated by Newspaper in FB-NEWS.

Table 2 shows the total number of tokens for each page and the average number of texts, produced for each post for each page. We can notice that the most followed newspapers on Facebook are Il Fatto Quotidiano and La Repubblica.

Newspaper	Tokens	Texts/Posts
La Repubblica	96,059,756	182.61
Avvenire	2,611,899	12.65
Il Giornale	64,345,260	77.93
Libero	41,166,457	81.87
Il Fatto Quotidiano	99,025,541	193.33
Rainews24	7,735,908	10.21
Huffington Post	32,587,065	84.06
Corriere della Sera	64,197,579	95.01
Overall	407,729,465	94.83

Table 2: Tokens and Texts/Posts ratio for page.

3 Topics in FB-NEWS15

FB-NEWS15 contains texts referring to a large variety of events. In order to organize the corpus into clusters of thematically related news, we used LDA (Blei et al., 2003). LDA represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. These random mixtures express a document semantic content, and document similarity can be estimated by looking at how similar the corresponding topic mixtures are. For the topic identification we used the software Mallet (McCallum, 2002).

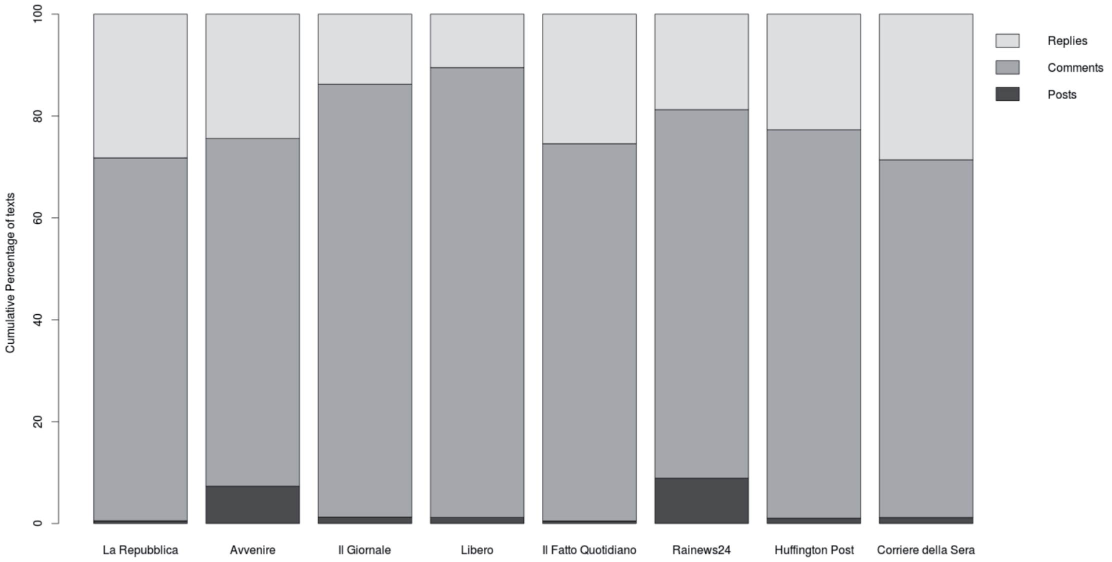


Figure 3: Cumulative distribution of posts, comments and replies in FB-NEWS15 for each Newspaper.

3.1 Selecting the vocabulary

Since we were interested in extracting the topics from the news articles, we have built the model on the portion of FB-NEWS15 containing the posts (FB-NEWS15_posts) published by the newspaper. In particular, we used entropy (Dumais, 1990) as a global term weighting and we selected for training the terms (nouns, adjectives, verbs and complex terms) with a high informative value (threshold fixed to 0.3), while using the remaining words as stopwords in Mallet (McCallum, 2002).

3.2 Extracting topics from posts

In order to determine the most debated topics in 2015, we used LDA to assign 50 topics to the posts in FB-NEWS15_posts and we navigated the graph to assign the topics to the comments and the replies. Later, we restricted the topics associated to a post P to the topics T having a probability higher than the 90^{th} percentile of the topic distribution of P. In this way, each post has been assigned, on average, to 3.06 topics. Finally, comments and replies have inherited the probability of belonging to the topic T from their parent post. Among the extracted topics ranked according to the sum of these probabilities we can find national and foreign politics, terrorism and church but also food, football, cinema and weather forecast. We report some topics below, with the number of texts and the relative ranking (i.e., rank 1 is given to the topic with the higher number of texts).

NATIONAL POLITICS (2,516,640 TEXTS, RANK 1): {*Renzi, presidente, premier, Mattarella, riforma, Alfano, senato, camera, Boschi, aula*} (Renzi, president, Mattarella, reform, Alfano, senate, chamber, Boschi, hall)

SCHOOL (1,707,145 TEXTS, RANK 2):{*scuola, giovane, studente, protesta, corso, mancare, sospendere, inglese, spiegare, lezione*} (school, young, protest, class, lack, suspend, English, explain, lesson)

CRIME (1,543,735 TEXTS, RANK 7): {*uccidere, polizia, arrestare, fermare, sparare, uomo, poliziotto, colpo, ferire, agente*} (kill, police, detain, stop, open fire, man, policeman, bump, wound, police officer)

ISIS (1,267,749 TEXTS, RANK 16): {*Isis, guerra, siria, minaccia, U.S.A., Libia, colpire, islamico, usare, jihadisti* } (Isis, war, Syria, threat, U.S.A., Libya, damage, islamic, use, jihadist)

FOOD (949,520 TEXTS, RANK 40): {*mangiare, ricetta, cibo, preparare, consiglio, evitare, perfetto, trucco, salute, semplice*} (eat, recipe, food, prepare, advice, avoid, perfect, trick, health, simple)

FOOTBALL (606,560 TEXTS, RANK 50): {*seguire la diretta, guardare il video, campo, calcio, serie, Napoli, Milan, segnare, battere, partita*} (follow the live, look at the video, football field, football, league, Naples, Milan)

4 Conclusions and ongoing work

As one of the most widespread social networks, Facebook offers the possibility to collect opinionated pieces of texts from people of different ages, cultures and education. The composition of FB-NEWS15, in which each comment is explicitly associated with a particular post, allows us to study the differences in terms of readers' perceptions about a particular topic. Differently from other social media like Twitter, Facebook contains larger texts including lot of subjective expressions that are very useful for the construction of sentiment and emotive lexicons.

Starting from previous works (Passaro et al., 2015; Passaro and Lenci, 2016), we plan to use this corpus to build lexical resources for sentiment analysis and emotion detection, which will include both words and complex terms. In addition, we plan to optimize the topic modeling phase and to investigate the possibility of using the extracted topics as a prior for inferring the sentiment orientation of a particular comment.

References

A. Ahmad. 2010. Is twitter a useful tool for journalists? *Journal of Media Practice*, 11(2):145–155.

A. Ahmad. 2013. Whats in a tweet? foreign correspondents use of social media. *Journalism Practice*, 7(1):33–46.

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *EVALITA 2009 Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

H. Chyi and D. L. Lasorsa. 2002. An explorative study on the market relation between online and print newspapers. *Journal of Media Economics*, 15(2):91–106.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *EVALITA 2009 Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

S. T. Dumais. 1990. Enhancing performance in latent semantic indexing (lsi) retrieval. Technical Report TM-ARH-017527.

A. Hermida. 2013. #journalism. reconfiguring journalism research about twitter, one tweet at a time. *Digital Journalism*.

H. Lin and L. Qiu. 2013. Two sites, two voices: Linguistic differences between facebook status updates and tweets. In P. L. Patrick Rau, editor, *Cross-Cultural Design. Cultural Differences in Everyday Life: 5th International Conference, CCD 2013, Held as Part of HCI International 2013*, volume 2, pages 432–440, Las Vegas (USA). Springer Berlin Heidelberg.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

N. Newman, W. H. Dutton, and G. Blank. 2012. Social media in the changing ecology of news: The fourth and fifth estate in britain. *Internet Science*, 7(1):6–22.

Oriella. 2013. The new normal for news. have global media changed forever? *The 6th Annual Oriella Digital Journalism Survey*.

L. C. Passaro and A. Lenci. 2015. Extracting terms with extra. In *Proceedings of the EUROPHRAS 2015 Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196, Malaga (Spain).

Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), may.

L. C. Passaro, L. Pollacci, and A. Lenci. 2015. Item: A vector space model to bootstrap an italian emotive lexicon. In *Proceedings of the second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 215–220, Trento (Italy).

R. Picard. 2009. Blogs, tweets, social media, and the news business. *Nieman Reports*, 63(3):10–12.

D. V. Shah, J. Cho, W. P. Eveland, and N. Kwak. 2005. Information and expression in a digital age. *Communication Research*, 32(10):531–565.

M. L. Sheffer and B. Schultz. 2010. Paradigm shift or passing fad? twitter and sports journalism. *International journal of Sport Communication*, 3(4):472–484.