Long-term Social Media Data Collection at the University of Turin

Valerio Basile
University of Turin
basile@di.unito.it

Mirko Lai
University of Turin
mirko.lai@unito.it

Manuela Sanguinetti
University of Turin
msanguin@di.unito.it

Abstract

We report on the collection of social media messages — from Twitter in particular — in the Italian language that is continuously going on since 2012 at the University of Turin. A number of smaller datasets have been extracted from the main collection and enriched with different kinds of annotations for linguistic purposes. Moreover, a few extra datasets have been collected independently and are now in the process of being merged with the main collection. We aim at making the resource available to the community to the best of our possibility, in accordance with the Terms of Service provided by the platforms where data have been gathered from.

(Italian) In questo articolo descriviamo il lavoro di raccolta di messaggi — da Twitter in particolar modo — in lingua italiana che va avanti in maniera continuativa dal 2012 presso l'Università di Torino. Diversi dataset sono stati estratti dalla raccolta principale ed arricchiti con differenti tipi di annotazione per scopi linguistici. Inoltre, dataset ulteriori sono stati raccolti indipendentemente, e fanno ora parte della raccolta principale. Il nostro scopo è rendere questa risorsa disponibile alla comunità in maniera più completa possibile, considerati i termini d'uso imposti dalle piattaforme da cui i dati sono stati estratti.

1 Introduction

The online micro-blogging platform *Twitter*[1] has been a popular source for natural language data since the second half of the 2010's, due to the enormous quantity of public messages exchanged by its users, and the relative ease of collecting them through the official API.

Many researchers implemented systems to collect large datasets of tweets, and share them with the community. Among them, the Content-centered Computing group at the University of Turin[2] is maintaining a large, diversified collection of datasets of tweets in the Italian language[3]. However, although the Twitter datasets in Italian make the majority of our collection, over the years, and also in the recent past, several resources have been created in other languages and including data retrieved from other sources than Twitter.

In this paper, we report on the current status of the collection (Section 2) and we give an overview of several annotated datasets included in it (Section 3). Finally, we describe our current and future plans to make the data and annotations available to the research community (Section 4).

2 TWITA: Long-term Collection of Italian Tweets

The current effort to collect tweets in the Italian language started in 2012 at the University of Groningen (Basile and Nissim, 2013). Taking inspiration from the large collection of Dutch tweets by Tjong Kim Sang and van den Bosch (2013), Basile and Nissim (2013) implemented a pipeline to collect and automatically annotate a large set of tweets in Italian by leveraging the Twitter API. The process interrogates the *stream* API with a set of keywords designed to capture the Italian language and at the same time excluding other languages. At the time of its publishing, the resource contained about 100 million tweets in Italian in the first year (from February 2012 to February

[1] https://twitter.com/

[2] http://beta.di.unito.it/index.php/english/research/groups/content-centered-computing/people

[3] Some of the datasets included in this report and their methodology of annotation are described in Sanguinetti et al. (2014)

2013). The automatic collection, however, continued, and in 2015 was transferred from the University of Groningen to the University of Turin. From June 2018, a new filter based on the five Italian vowels has been added to the pipeline, along with the language filter provided by the Twitter API, which was not previously available, in order to limit the number of accidentally captured tweets in other languages. In the latest version of the data collection pipeline, a Python script employing the tweepy library[4] gathers JSON tweets using the following filter: track=["a","e","i","o","u"] and languages=["it"]. We stored the raw, complete JSON tweet structures in zipped files for backup. Meanwhile, we store the text and the most useful metadata (username , timestamp, geolocalization, retweet and reply status) in a relational database in order to perform efficient queries.

At the time of this writing, the collection comprises more than 500 million tweets in the Italian language, spanning 7 years (57 months) from February 2012 to July 2018. There are a few holes in the collection, sometimes spanning entire months, due to incidents involving the server infrastructure or changes in the Twitter API which required manual adjustment of the collection software. Figure 1 shows the percentage of days in each month for which the collection has data, at the time of this writing.

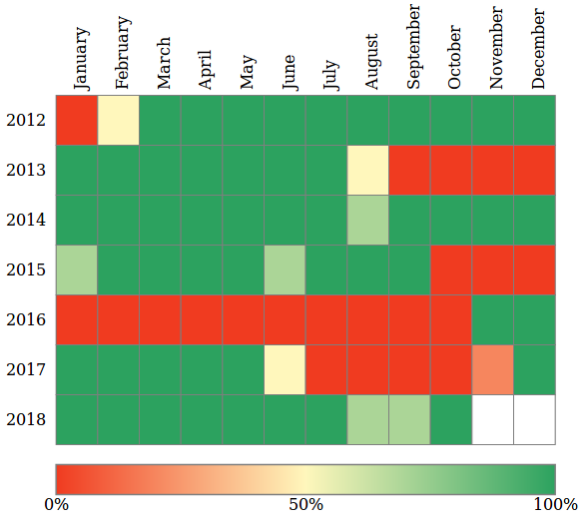

Figure 1: Percentage of days in each month for which tweets are available.

[4]http://www.tweepy.org/

3 Annotated Datasets

In the past years, the TWITA collection has been made available to many research teams interested in the study of social media in the Italian language with computational methods. Several such studies focused on creating new linguistic resources starting from the raw tweets and basic metadata provided by TWITA, including a number of datasets created for shared tasks of computational linguistics. In this section, we give an overview of such resources. Moreover, some datasets were created independently from TWITA, and are now managed under the same infrastructure, therefore we include them in this report.

For each dataset, we provide a summary infobox with basic information, including the type of annotation performed on the the dataset and how it was achieved, i.e., by means of expert annotators or a crowdsourcing platform.

3.1 Datasets From TWITA

The datasets described in this section are subsets of the main TWITA dataset, obtained by sampling the collection according to different criteria, and annotated for several purposes.

TWitterBuonaScuola (Stranisci et al., 2016) is a corpus of Italian tweets on the topic of the national educational and training systems. The tweets were extracted from a specific hashtag (#labuonascuola, the nickname of an education reform, translating to *the good school*) and a set of related keywords: "la buona scuola" (*the good school*), "buona scuola" (*good school*), "riforma scuola" (*school reform*), "riforma istruzione" (*education reform*).

Name: TWitterBuonaScuola
Size: 35,148 total tweets, 7,049 annotated tweets
Time period: February 22, 2014–December 31, 2014
Annotation: polarity, irony and topic
Annotation method: crowdsourcing
URL: http://twita.dipinfo.di.unito.it/tw-bs

TW-SWELLFER (Sulis et al., 2016) is a corpus of Italian tweets on subjective well-being, in particular regarding the topics of fertility and parenthood. The tweets were collected by searching for 11 hashtags — #papa (*father*), #mamma (*mother*), #babbo (*dad*), #incinta (*pregnant*), #primofiglio (*first child*), #secondofiglio (*second child*), #futuremamme (*future moms*), #maternita (*materhood*), #paternità (*fatherhood*), #allattamento (*nursing*), #gravi-

danza (*pregnancy*) — and 19 related keywords.

Name: TW-SWELLFER
Size: 2,760,416 total tweets, 1,508 annotated tweets
Time period: 2014
Annotation: polarity, irony and sub-topic
Annotation method: crowdsourcing
URL: http://twita.dipinfo.di.unito.it/tw-swellfer

Italian Hate Speech Corpus (Sanguinetti et al., 2018b; Poletto et al., 2017) is a corpus of hate speech on social media towards migrants and ethnic minorities, in the context of the Hate Speech Monitoring Program of the University of Turin[5]. The tweets were collected according to a set of keywords: invadere (*invade*), invasione (*invasion*), basta (*enough*), fuori (*out*), comunist* (*communist**), african* (*African*), barcon* (*migrants boat**).

Name: Italian Hate Speech Corpus
Size: 236,193 total tweets, 6,965 annotated tweets
Time period: October 1st, 2016–April 25th, 2017
Annotation: hate speech, aggressiveness, offensiveness, stereotype, irony, intensity
Annotation method: crowdsourcing and experts
URL: http://twita.dipinfo.di.unito.it/ihsc

TWITTIRÒ (Cignarella et al., 2017) is a dataset of tweets overlapping with other datasets included in the University of Turin collection, on which a finer-grained annotation of irony is superimposed. The TWITTIRÒ tweets are taken from TWitterBuonaScuola, SENTIPOLC (see Section 3.2), and TWSpino (see Section 3.3).

Name: TWITTIRÒ
Size: 1,600 total tweets: 400 tweets from TWSpino, 600 from SENTIPOLC tweets, 600 tweets from TWitterBuonaScuola
Time period: 2012–2016
Annotation: fine-grained irony
Annotation method: experts
URL: http://twita.dipinfo.di.unito.it/twittiro

3.2 Shared Task Datasets

The large collection of Italian tweets of the University of Turin has been exploited in different occasions to extract datasets to organize shared tasks for the Italian community, in particular under the umbrella of the EVALITA evaluation campaign[6]. In this section, we describe such datasets.

SENTIPOLC The SENTIment POLarity Classification task was proposed in two editions of the EVALITA campaign, namely in 2014 (Basile et al., 2014) and 2016 (Barbieri et al., 2016). Both editions were organized into three different sub-tasks: subjectivity and polarity classification, and irony detection. The data for SENTIPOLC 2014 were gathered from TWITA and Senti-TUT (see Section 3.3), while for the 2016 edition the dataset was further expanded by including other data sources, such as TWitterBuonaScuola (see Section 3.1) and a subset of TWITA overlapping with the dataset used for the shared task on Named Entity Recognition and Linking in Italian Tweets (Basile et al., 2016, NEEL-it).

Name: SENTIPOLC
Size: 6,448 (SENTIPOLC 2014), 9,410 (SENTIPOLC 2016) tweets
Time period: 2012 (SENTIPOLC 2014), 2014 (SENTIPOLC 2016)
Annotation: subjectivity, polarity, irony
Annotation method: experts (SENTIPOLC 2014), crowdsourcing and experts (SENTIPOLC 2016)
URL: http://twita.dipinfo.di.unito.it/sentipolc

PoSTWITA (Bosco et al., 2016b) is the shared task on Part-of-Speech tagging of Twitter posts held at EVALITA 2016. Its content was extracted from the SENTIPOLC corpus described above. The PoSTWITA dataset consists of Italian tweets tokenized and annotated at PoS level with a tagset inspired by the Universal Dependencies scheme[7].

Name: PoSTWITA
Size: 6,738 tweets
Time period: 2012
Annotation: part of speech
Annotation method: experts
URL: http://twita.dipinfo.di.unito.it/postwita

After the task took place, the PoSTWITA corpus has been used in a new independent project on the development of a Twitter-based Italian treebank fully compliant with the Universal Dependencies, thus becoming **PoSTWITA-UD** (Sanguinetti et al., 2018a). In particular, the first core of the resource was automatically annotated by out-of-domain parsing experiments using different parsers. The output with the best results was then revised by two annotators for the final version of the resource.
PoSTWITA-UD has been made available in the official UD repository[8] since v2.1 release.

Name: PoSTWITA-UD
Size: 6,712 tweets
Time period: 2012
Annotation: dependency-based syntactic annotation
Annotation method: experts
URL: http://twita.dipinfo.di.unito.it/postwita-ud

[5] http://hatespeech.di.unito.it/
[6] http://www.evalita.it/
[7] http://universaldependencies.org/
[8] https://github.com/UniversalDependencies/UD_Italian-PoSTWITA

IronITA The irony detection task proposed for EVALITA 2018[9] consists in automatically classifying tweets according to the presence of irony (sub-task A) and sarcasm (sub-task B). Given the array of situations and topics where ironic or sarcastic devices can be used, the corpus has been created by resorting to multiple annotated sources, such as the already mentioned TWITTIRÒ, SENTIPOLC, and the Italian Hate Speech Corpus.

Name: IronITA
Size: 4,877 tweets
Time period: 2012–2016
Annotation: irony, sarcasm
Annotation method: crowdsourcing and experts
URL: http://twita.dipinfo.di.unito.it/ironita

HaSpeeDe The Hate Speech Detection task[10] at EVALITA 2018 consists in automatically annotating messages from Twitter and Facebook. The dataset proposed for the task is the result of a joint effort of two research groups on harmonizing the annotation previously applied to two different datasets: the first one is a collection of Facebook comments developed by the group from CNR-Pisa and created in 2016 (Del Vigna et al., 2017), while the other one is a subset of the Italian Hate Speech Corpus (described in Section 3.1). The annotation scheme has thus been simplified, and it only includes a binary value indicating whether hateful contents are present or not in a given tweet or Facebook comment. The task organizers created such harmonized scheme also in view of a cross-domain evaluation, with one dataset used for training and the other one for testing the system.

It is worth pointing out, however, that despite their joint use in the task, the resources are maintained separately, thus only the Twitter section of the dataset is part of TWITA.

Name: HaSpeeDe
Size: 4,000 tweets and 4,000 Facebook comments
Time period: 2016–2017 for the Twitter dataset, May 2016 for the Facebook dataset
Annotation: hate speech
Annotation method: crowdsourcing and experts for the Twitter dataset, experts for the Facebook dataset
URL: http://twita.dipinfo.di.unito.it/haspeede

3.3 Independently-collected Datasets

To complete the overview of the social media datasets, in this section we describe collections of tweets that have been compiled independently from TWITA. However, they are now hosted in the same infrastructure and therefore can be considered part of the same collection.

Senti-TUT (Bosco et al., 2013) is a dataset of Italian tweets with a focus on politics and irony. Senti-TUT includes two corpora: *TWNews* contains tweets retrieved by querying the Twitter search API with a series of hashtags related to Mario Monti (the Italian First Minister at the time); *TWSpino* contains tweets from Spinoza[11], a popular satirical Italian blog on politics.

Name: Senti-TUT
Size: 3,288 (TWNews), 1,159 tweets (TWSpino)
Time period: October 16th, 2011–February 3rd, 2012 (TWNews), July 2009–February 2012 (TWSpino)
Annotation: polarity, irony
Annotation method: experts
URL: http://twita.dipinfo.di.unito.it/senti-tut

Felicittà (Allisio et al., 2013) was a project on the development of a platform that aimed to estimate and interactively display the degree of happiness in Italian cities, based on the analysis of data from Twitter. For its evaluation, a gold corpus was created by Bosco et al. (2014), using the same annotation scheme provided for Senti-TUT.

Name: Felicittà
Size: 1,500 tweets
Time period: November 1st, 2013–July 7th, 2014
Annotation: polarity, irony
Annotation method: experts
URL: http://twita.dipinfo.di.unito.it/felicitta

ConRef-STANCE-ita (Lai et al., 2018) is a collection of tweets on the topic of the Referendum held in Italy on December 4, 2016, about a reform of the Italian Constitution. This is supposedly a highly controversial topic, chosen to highlight language features useful for the study of stance detection. The tweets were collected by searching for specific hashtags: #referendumcostituzionale (*constitutional referendum*), #iovotosi (*I vote yes*), #iovotono (*I vote no*). Subsequently, the collection was enriched by recovering the conversation chain from each retrieved tweet to its source, annotating triplets consisting in one tweet, one retweet, and one reply posted by the same user in a specific temporal window. The aim of the collection is to monitor the evolution of the stance of 248 users during the debate in four different temporal windows and also inspecting their social network.

[9] http://www.di.unito.it/~tutreeb/ironita-evalita18

[10] http://www.di.unito.it/~tutreeb/haspeede-evalita18

[11] http://www.spinoza.it

Name: ConRef-STANCE-ita
Size: 2,976 tweets (963 triplets)
Time period: November 24th, 2016–December 7th, 2016
Annotation: stance
Annotation method: crowdsourcing and experts
URL: http://twita.dipinfo.di.unito.it/conref-stance-ita

3.4 Work in Progress and Other Datasets

Finally, there are a number of additional datasets hosted in our infrastructure that are being actively developed at the time of this writing. Some of those datasets include a collection of geo-localized tweets on the 2016 edition of the "giro d'Italia" cycling competition, a dataset of tweets concerning the 2016 local elections in 10 major Italian cities, and an addendum to the ConRef-STANCE-ita dataset described in Section 3.3.

Furthermore, we limited this report to the datasets of tweets in the Italian language, which make for the majority of our collection. However, we curate several datasets in other languages, often as a result of collaborations with international research teams and projects, such as, for instance, **TwitterMariagePourTous** (Bosco et al., 2016a), a corpus of 2,872 French tweets extracted in the period 16th December 2010 - 20th July 2013 on the topic of same-sex marriage. In addition, several new corpora have been developed within the Hate Speech Monitoring program (see Section 3.1), aiming at studying hate speech phenomenon against different targets such as women and the LGBTQ community, and resorting to other data sources than Twitter (Facebook and online newspapers in particular). Although such resources are still under construction - therefore it is not possible to provide any corpus statistics yet - our goal is to include them in our resource infrastructure, thus making a step forward and ensuring its improvement also in terms of diversity of data sources.

4 Data Availability

The main goal of collecting and organizing datasets such as the ones described in this paper is, generally speaking, to provide the NLP research community with powerful tools to enhance the state of the art of language technologies. Therefore, our default policy is to share as much data as possible, as freely as possible. Twitter has proven to behave cooperatively towards the scientific community, relaxing the limits imposed to data sharing for non-commercial use over time[12]. However, there are considerations about the privacy of the users that must be accounted for in releasing Twitter data. In particular, the EU General Data Protection Regulation from 2018 (GDPR)[13] strictly regulates data and user privacy. For instance, if a tweet has been deleted by a user, it should not be published in other forms (Article 17), although it can still be used for scientific purposes.

Technically, we follow these consideration by implementing an interface to download the ID of the tweets in our collection, and tools to retrieve the original tweets (if still available). The annotated datasets can instead be shared in their entirety, given their limited size, thus we provide links to download them in tabular format. Finally, we are developing interactive interfaces to select and download samples of the collection based on the time period and sets of keywords and hashtags.

[12] https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

[13] https://gdpr-info.eu/

Acknowledgments

Valerio Basile and Manuela Sanguinetti are partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

Mirko Lai is partially supported by Italian Ministry of Labor (*Contro l'odio: tecnologie informatiche, percorsi formativi e story telling partecipativo per combattere l'intolleranza*, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del d.l. 3 luglio 2017, n. 117 - anno 2017).

References

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)*, pages 95–106, Turin, Italy.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

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.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2014)*, Pisa, Italy.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian tweets (NEEL-IT) task. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016) & the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti, and Emilio Sulis. 2014. Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In *Proceedings of the 5th International Workshop on EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA*, pages 56 – 63.

Cristina Bosco, Mirko Lai, Viviana Patti, and Daniela Virone. 2016a. Tweeting and being ironic in the debate about a political reform: the French annotated corpus Twitter-MariagePourTous. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, Portorož, Slovenia.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016b. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian task. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016) & the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

Alessandra Teresa Cignarella, Cristina Bosco, and Viviana Patti. 2017. Twittirò: a social media corpus with a multi-layered annotation for irony. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),*, pages 86–95, Venice, Italy.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *NLDB*, volume 10859 of *Lecture Notes in Computer Science*, pages 15–27. Springer.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an Italian Twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy.

Manuela Sanguinetti, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, Leonardo Allisio, Valeria Mussa, and Cristina Bosco. 2014. Developing corpora and tools for sentiment analysis: the experience of the University of Turin group. In *First Italian Conference on Computational Linguistics (CLiC-it 2014)*, pages 322–327, Pisa, Italy.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018a. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the 11th Language Resources and Evaluation Conference LREC 2018)*, pages 1768–1775, Miyazaki, Japan.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018b. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marco Stranisci, Cristina Bosco, Delia Iraz Hernndez Faras, and Viviana Patti. 2016. Annotating sentiment and irony in the online italian political debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Emilio Sulis, Cristina Bosco, Viviana Patti, Mirko Lai, Delia Irazú Hernández Farías, Letizia Mencarini, Michele Mozzachiodi, and Daniele Vignoli. 2016. Subjective well-being and social media. A semantically annotated Twitter corpus on fertility and parenthood. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016) & the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

E. Tjong Kim Sang and A. van den Bosch. 2013. Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3(12/2013):121–134. Reporting year: 2013.