Detection of Privacy-Harming Social Media Posts in Italian

Detection of Privacy-harming Social Media Posts in Italian

Federico Peiretti[1] and Ruggero G. Pensa[1][0000−0001−5145−3438]

University of Turin, Dept. of Computer Science, I-10149, Turin, Italy
{federico.peiretti,ruggero.pensa}@unito.it

Abstract. As many psychological and sociological study reveal, many people disclose too much privacy-harming information in social media in the form of text and multimedia posts, thus exposing themselves and other persons to several security risks. Consequently, many researchers have addressed this problem by investigating on the detection and analysis of the so-called self-disclosure behavior in social media and blogging platforms. Among the others, content sensitivity analysis has emerged as a promising research direction, but, so far, it has only focused on English text posts, although it is well-known that people tend to disclose mostly in their own native languages. Therefore, in this paper, we address this limitation by proposing a new text corpus of Italian posts that we have annotated following to the anonymity assumption. We then apply several language models based on transformers to classify them according to their sensitivity. Moreover, since Italian is a lower-resource language compared to English, we also apply some multilingual zero-shot transfer learning architectures trained on a rich and manually annotated English corpus and tested on the Italian one. We show experimentally that the approaches trained directly on the Italian corpus, still outperform multilingual ones trained on the English data and tested on Italian, although some of them exhibit promising prediction performances.

Keywords: Privacy · Neural language models · Social media

1 Introduction

Online social media are valuable and somehow irreplaceable content sharing and networking platforms, but are often subject of criticism, for many reasons. Sometimes such reasons are unjustified and the results of prejudices or lack of knowledge about social media and their enabling technologies (e.g, smartphones), but those regarding privacy are real, as proven by the many studies [68, 39, 7, 30, 66, 18, 42, 2]. The concerns about the risks of privacy violation in social media also inspired documentaries and movies, including several episodes of the award-winning TV show "Black Mirror" [13]. However, when referring to personal data in social media platforms, there are two complementary aspects that should be considered. The first one concerns the usage social media companies do with personal data, which is often the object of their terms of service and can be

partly customized by registered users (e.g., the users may decide not to allow the access to their geolocation data). The second aspect regards the way people communicate and interact with other users and how much personal information they expose about themselves, a phenomenon that, in psychology, is referred to as self-disclosure [32]. Self-disclosure has been studied in relation to different contexts, including online forums [5], online support groups [65] and social media [39], although it has often been investigated for discussion boards dealing with intrinsically sensitive topics, such as health issues, intimate relationships or sex life, and where the identity of the users is masked by pseudonyms or entirely anonymous. Instead, in most social media platforms, social profiles usually carry the real identities of their owners, and yet this does not prevent their users from disclosing very private information [7, 18, 48] thus harming their own security.

In a very recent work [12], the authors have addressed the analysis of what they call "content sensitivity" (a more general problem than self-disclosure) of social media posts, and has drawn interesting insights about the possibility of automatically detecting the sensitivity of short texts by using natural language processing (NLP) techniques and also proposing a new annotated text corpus. However, as in most previous closely-related works, their study was focused on contents written in English only, although it is a well-known fact that most people mainly interact on social media using their own native language and, consequently, they tend to self-disclose more in their native language, than in English [56]. Unfortunately, with the exception of English, Chinese, Japanese and some European languages, the majority of national or regional idioms are considered low-resource languages, due to the lack of large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications. As a result, all major existing works on the automated characterization or recognition of sensitive contents focus on English texts only [45, 46, 67, 65, 31, 12].

In this work, we address this limitation by presenting a new annotated corpus of Italian posts and applying several monolingual and multilingual approaches for classifying them according to their privacy-sensitivity. We compare several pre-trained language models including the main transformer-based models and two alternative approaches, LASER [3] and MultiFiT [26], in two different experimental settings: in the first one the models are trained on the Italian corpus only; in the second one, the models are trained on the English corpus and the knowledge is transferred on the Italian one. Our experiments show that, although promising, multi-lingual approaches can not replace fully monolingual models in such task, where short social media posts are considered.

2 Related work

Since modern online social networking platforms have gained popularity and success, the characterization and measurement of the exposure of user privacy in the Web has attracted the scientific interest of many research groups [42, 41]. To assess the risk of privacy disclosure in social networks, many different

approaches have been proposed [60, 2], mostly focusing on measures based on the privacy settings of the users [36, 47], or on their position within the network [48]. On the other hand, very few studies have investigated the problem of detecting the sensitivity of the contents posted by social network users, also because there is no consensus on how to define privacy-sensitivity [28, 57].

A common solution to this problem is to consider all contents posted anonymously as sensitive, thus simplifying the construction of specifically annotated corpora. According to this assumption (called "anonymity assumption" [12]), if some content is posted anonymously, it is deemed sensitive, otherwise it is considered as non sensitive. This strategy is adopted, for instance, to apply some machine learning model and analyze anonymous and non anonymous posting behavior in question-and-answer platforms [45], or to compare content posted on anonymous and non-anonymous social media, according to their topics and linguistic features [22, 11]. The largest available corpus supporting this category of studies consists of nearly 90 million posts downloaded from Whisper, an anonymous social media platforms [40]. Another solution is to consider the privacy settings associated to shared items as a proxy for measuring sensitivity: contents posted with more restricting visibility are deemed sensitive, as done by Yu *et al.* to measure the sensitivity of photos and to identify categories of privacy-sensitive objects according to a deep multi-task learning model [67].

The concept of content sensitivity is closely related to the one of self-disclosure, defined as the act of revealing personal information to others. It has been extensively investigated well before the advent of the Internet and social media [32]. In more recent years, self-disclosure has been studied to show that people behave differently in online support groups and discussion forums [5]. Other studies analyze the differences in the degree of positive and negative disclosure according to the visibility (private or public) of discussion channels in online support groups for cancer patients and their relatives. They apply support vector machines on lexical, linguistic, topic-related and word-vector features extracted from a small annotated corpus [65]. In [63], machine learning has been used to detect the degree of self-disclosure of social media posts and to replicate patterns from other empirical and theoretical work using a feature engineering approach. The experiments, conducted on a relatively small and proprietary corpus, identify post length, emotional valence, the presence of certain topics, social distance and social normality, among the most distinctive features for self-disclosure. Instead, in [31], Jaidka *et al.* report the results of a challenge concerning a relatively large corpus consisting of posts collected from Reddit, all annotated according to their degree of informational and emotional self-disclosure. Finally, in [12] the task of *content sensitivity analysis* is defined and different classification models are applied on three different text corpora of short social media posts, including a specifically annotated corpus of Facebook posts. The authors show the results for different lexical-based models as well as several classifiers based on CNNs, RNNs and language models in predicting content sensitivity of short posts.

All the works mentioned so far focus on English text, making it a high-resource language even for content sensitivity analysis. In fact, the few existing

works on self-disclosure on non-English languages are psychological studies based on surveys [4, 35, 27, 23]. To bypass the lack of linguistic resources when dealing with languages other than English, one may consider cross-lingual or multilingual models, which have been applied with success to many problems, including sentiment analysis [25], emotion detection [1] and information retrieval [59]. Multilingual language models have gained popularity thanks to their success in zero-shot transfer learning. The idea behind cross-lingual or multi-lingual models is to learn a shared embedding space for two or more languages to improve their ability for machine translation. One of the early models is XLM [21], which defines a new cross-lingual objective trained on two different languages. Another early approach is CMLM [51], which computes cross-lingual n-gram embeddings and infers an n-gram translation table from them. Both the models are cross-lingual, while, more recently, multilingual approaches have emerged, which are pre-trained once for all languages. Notable examples are mBART [37] and mT5 [64]: the former consists in a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective [34]; the latter is a multilingual variant of a text-to-text transfer transformer trained on a dataset covering 101 languages. Other multilingual representations are based on contrastive learning that samples sentences from the document and constructs positive and negative pairs based on their saliency [62], or computing a contrastive loss on the representations of aligned pairs of sentences (considered as positive examples) and randomly selected non-aligned pairs (considered as negative examples) [43].

3 Identifying privacy-sensitive content

In this section, we describe the task of content sensitivity analysis, introduced in [8, 12]. However, since there is no agreement in the definition of privacy-sensitive content, we first precise what we mean by it in this paper.

3.1 Privacy-sensitive content

User-generated content, in the form of text and/or multimedia items (photos, videos), may carry sensitive information concerning the private life of the author or any other identifiable person and its explicit or implicit disclosure could potentially cause harm or embarrassment to them. In fact, such content could involve financial or medical information, sexual orientation and preferences, religious or political beliefs, or any other kind of personal data that, if posted online, could be exploited by third parties for malicious purposes, such as identity theft, frauds, discrimination, cyberbullying or stalking. A social media post with all these characteristics is defined as *privacy-sensitive* by Battaglia *et al.* [8]. This concept is a generalization of *self-disclosure* since, unlike it, revealing privacy-sensitive information could not only concern the author himself, but also other individuals mentioned (explicitly or implicitly) in the content item. Another important key point is how information is disclosed: sometimes sensitive information is clearly,

directly and voluntarily disclosed. However, often it can be inferred from the context or by using some background knowledge. Some examples of sensitive and non sensitive posts are showed below.

1. *Guys, I'm taking some days off! On my way to Barcelona with my friend Alice. See you in two weeks.*
2. *How would you react if your doctor told you that they diagnosed you with cancer and that you need to start chemotherapy?*
3. *A 5th person is likely cured of HIV, and another is in long-term remission.*

The first text discloses information about the author and their friend *Alice* explicitly, despite neither does it have any sensitive term, nor deals with any sensitive topic. From the text, it is clear that they will be far from their respective homes for two weeks. It also contains hidden spatiotemporal references that are clear from the context.

The second post is a general question that does not disclose any sensitive information apparently, but this assumption might not be true. Is is very likely, indeed, that the author themself was diagnosed with cancer and will have to start chemotherapy. It is an implicit way to reveal very sensitive medical information.

Finally, despite the third text item deals with a sensitive topic, it does not really disclose any private information that could put in danger the privacy of any people, since there is not any direct or indirect reference to a specific identifiable person. This sentence could be a citation from a newspaper or scientific article.

3.2 Content sensitivity analysis

Content sensitivity analysis is a data mining task aimed at recognizing whether a given user-generated content item is privacy-sensitive or not, according to the definition given above [8]. This particular task has been extensively investigated for the analysis of text posts written in English [12]. Instead, in this paper, we focus on user-generated content written in different languages, with a special focus on Italian, as it has been shown that most users mainly use their own native language in social media and, consequently, they tend to self-disclose more in their mother tongue(s), than in a different vehicular language, such as English [56]. The original definition of content sensitivity analysis is as follows.

Definition 1 (Content sensitivity analysis). *Given a user-generated content item $c_i \in \mathcal{C}$, where $\mathcal{C}$ is user-generated content domain,* content sensitivity analysis *is a task aimed at defining a function $f_s : \mathcal{C} \rightarrow \{sensitive, non-sensitive\}$, such that:*

$$f(c_i) = \begin{cases} sensitive & \text{if } c_i \text{ is privacy-sensitive} \\ non\text{-}sensitive & \text{otherwise.} \end{cases}$$

In the following, without loss of generality, we will limit the scope of this definition to text posts only. Examples of privacy-sensitive posts are those containing information that violates a person's privacy (not necessarily of the author

of the post), for instance: information about current or future travels; physical or mental well-being; lifestyle habits that may reveal the writer's location or that of others mentioned; romantic relationship status; opinions that may suggest political or religious belief.

According to Definition 1, a simple way to implement an inductive content sensitivity analysis task is by defining it as a binary classification task, where the parameters of the classification function are learned by training the classifier from an annotated corpus of sensitive and non sensitive posts.

On the other hand, the definition does not take into account the possible different nuances of sensitivity, i.e., how much privacy-sensitive a content is. The degree of sensitivity of a post may vary according, for instance, to its topic, its lexical features, or its context. For instance: a post revealing health information is much more sensitive than one about holidays, although both are considered privacy-sensitive. A more precise definition taking into account different degrees of sensitivity has been given in [8], but, for simplicity, in this paper we refer to this task as a binary classification, as it has been done in [12] for English posts.

4 Text corpora for content sensitivity analysis

Training a classifier to make it capable of solving a content sensitivity analysis task, requires to feed it with a huge amount of user-generated text content also including privacy-harming information and, additionally, annotated by experts according to its actual sensitivity. However, finding and collecting such type of information on the Web is very hard and the reason is, mainly, the privacy itself. Indeed, social networks, such as Facebook, Instagram and Twitter, would be very rich sources to accomplish this goal as posts and tweets with sensitive information are very likely published in users' personal profiles, but, although visible to their contacts and friends, they cannot be download using the API made available by those platforms. After the Cambridge Analytica scandal in 2018, Facebook introduced restrictions on data access by developers [54], followed by other social media companies. Currently, albeit with some limitation, they only allow to download public posts or posts published on public pages, which, however, are less relevant to our purposes. Fortunately, there exists some corpora of social media posts collected and released publicly for research purposes before these restrictions were introduced, or downloaded from anonymous blogging platforms. For instance, the *myPersonality* [16] corpus is made up of around 10 000 posts downloaded from Facebook, collected by Cambridge University between 2009 and 2012 for a psychological study [33]. Another example is *CL-Aff #OffMyChest* [31], a corpus containing discussions on family and intimate relationships downloaded from Reddit. All posts of the two corpora, with few exceptions, are in English. In the remainder of this section, we will first introduce a new corpus in Italian for content sensitivity analysis; then, we will briefly describe the existing English corpora already used in similar tasks, that we will employ to train cross-linguals models.

4.1 An Italian corpus for content sensitivity analysis: ITA-SENS

Most of the (annotated) corpora extracted from social networks are mainly in English, due to the huge amount of natural language processing resources available and to the necessity of making the research findings universally accessible. There exists also corpora in other languages (including Italian), but most of them focus on very specific topics and tasks, such as the detection of racial stereotypes [14] or hate speech [53]. To train a content sensitivity classifier, instead, we need posts dealing with various generic topics. Therefore, we construct a new dataset, called *ITA-SENS*, consisting of more than 15 000 social media posts written in Italian. In this work, we rely on the anonymity assumption, i.e., we consider as sensitive all posts that have been shared anonymously, while posts shared publicly are considered non-sensitive. Hence, we focus on Italian and identify two sources of non-anonymous and anonymous posts: Twitter and Insegreto[1]. Regarding the first source, we take into account *Feel-IT* [10] and *SENTIPOLC* [6], two corpora of Italian tweets covering a wide range of generic topics. *FEEL-IT*[2] is a corpus of tweets written in Italian and annotated according to four base emotions: anger, fear, joy and sadness [10]. The curators of this dataset have downloaded tweets at a daily basis, by monitoring trend topic in a three-month period. The dataset consists of 1000 tweets per day covering different topics, including health, sports, societal problems and TV programs. Topics have different time-spans, from hours (e.g., tweets related to TV programs) or days (e.g., major sports events) to the entire observation period (general topics such as COVID-19). *SENTIPOLC*[3], instead, is a dataset consisting of tweets written in Italian and constructed to solve sentiment polarity classification [6]. It has been presented during EVALITA 2016, a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language[4]. Each tweet is annotated according to its topic; however, since most of them are related to a very specific political topic, we only retain 295 tweets addressing more different and general subjects. Additionally, to broaden the variety and quantity of tweets, we downloaded further tweets directly from Twitter, by filtering them according to popular general hashtags or by retrieving them from news accounts, using the official Twitter API.

As anonymous source, we take into consideration posts from Insegreto, an Italian social network that allows people to share their lives, secrets and opinions on different topics, in a totally anonymous way (the Italian locution "in segreto" means "secretly" in English). The posts are organized into several categories ranging from school to health, from politics to religion, from love to sexuality. As such, this is a valuable source of sensitive posts.

The statistics about this dataset are shown in Table 1. *ITA-SENS* consists of 15 144 Italian posts, of which 8 419 are labeled as sensitive and 6 725 as non-

[1] https://insegreto.com/it
[2] https://github.com/MilaNLProc/feel-it
[3] https://github.com/evalita2016/data
[4] https://www.evalita.it/

sensitive. We split it randomly by putting 55% of the data into the training set; 25% into the validation set and the remaining 20% into the test set[5].

Annotation Exactly as was done in previous studies [22], for the annotation of our corpus we rely on the anonymity assumption, where the content is considered sensitive if the user has chosen to publish it anonymously, hiding their real identity or if they have made it visible to very few friends. If the content is visible to anyone or the author can be identified from it, then it is considered non-sensitive. After a careful reading and analysis of the collected posts, we have observed that those coming from Insegreto contain sensitive information that could harm the privacy of both the author and other identifiable people. Furthermore, they deal with sensitive topics and are published in a totally anonymous way. For these reasons, we label all Insegreto posts as *sensitive*. On the other hand, we consider all tweets in the corpus as *non-sensitive* because they come from public Twitter pages and profiles, following the anonymity assumption. Although it has been shown that this assumption is simplistic [12], we rely on it for this work, as the main goal is to study whether multilingual text analysis approaches can compete with monolingual ones for the specific task of content sensitivity analysis.

4.2 An auxiliary English corpus: SENS2+OMC

In our work, we also leverage an additional dataset of social media posts written in English to train multilanguage models able to solve the content sensitivity classification task by transferring the learned knowledge from English to Italian. To this purpose, we merge two corpora: *SENS2* [12] and CL-Aff #OffMyChest [31], hereinafter referred to as *OMC*.

SENS2 is a subset of the dataset introduced in [12]. It consists of 8 765 English posts from Facebook covering a wide range of topics. The posts have been manually annotated by a pool of experts according to some guidelines providing privacy-sensitive content definitions and examples. More in detail, *SENS2* contains posts that received the same "sensitive" or "non sensitive" tag by at least two annotators. Therefore, 3 336 posts are annotated as sensitive, the others 5 429 as non-sensitive.

OMC[6], instead, is a corpus of English conversations about family and intimate relationships, extracted from two subreddits in Reddit: *r/CasualConversations*, a subcommunity where users share their opinions about different topics; *r/OffmyChest*, a mutually supportive community where deep sentiments and emotions are shared. Each post is annotated depending on how much informational and emotional disclosure it contains. We exploit such annotations to assign a new label to each post: "sensitive" if post discloses informational or emotional data, "non sensitive" otherwise. Consequently, the dataset contains 17 860 posts, of which 10 793 are annotated as "sensitive" and 7 067 as "non sensitive".

[5] Our dataset is available online at https://github.com/federicopeiretti/ITA-SENS
[6] https://github.com/kj2013/claff-offmychest

Table 1. Details on the datasets used in our study.

Dataset	Language	#posts	#sens	#nosens
ITA-SENS	Italian	15 144	8 419	6 725
SENS2	English	8 765	3 336	5 429
OMC	English	17 860	10 793	7 067

4.3 Preprocessing

Before feeding posts to a language model, they need to be preprocessed. To this end, we use a Python library optimized for texts from social networks, called Ekphrasis [9] that allow us to perform tokenization, word normalization, word segmentation and spell correction. We also remove URLs, emoticons and emojis that are not that important for analyzing the sensitivity of the text and, additionally, may introduce biases in the machine learning processes. However, we keep hashtags (removing the # symbol) because they are often used as terms in a sentence. In addition, we sanitize the text by replacing e-mail addresses, dates, hours, currencies and phone numbers, with a generic placeholder using the format $\langle entity\ type \rangle$.

5 Monolingual and cross-lingual content classification

In this section, we introduce the classification strategies used to solve the content sensitivity analysis task for Italian posts. Our methodology is based on the findings reported by the authors of [8] and [12], that show how to solve the same task for English. The authors train and compare different classifiers based on several types of models, from the most traditional ones (e.g. k-NN, SVM, Random Forest) to more sophisticated deep neural network models (e.g. CNN, LSTM, Google BERT). They observe that the former are not suitable because they fail to capture the manifold of privacy-sensitivity with sufficient accuracy. Instead, the latter perform better due to the ability of deep learning models to take into account the context of words and sentences in their training processes. In conclusion they find that BERT, the only model based on Transformers considered in their study, outperforms all other models.

Hence, following these results, we focus on the most recent and accurate Language Models (LMs) based on the Transformer architecture in order to classify Italian posts according to their sensitivity and compare them. We also examine two recent alternative approaches: LASER [3], performing multilingual sentence embedding for over 93 languages in a shared space, and MultiFiT [26], a fine-tuning technique that is much cheaper to pre-train but more efficient than Transformers in terms of space-time complexity.

In the following, we firstly illustrate the two experimental settings. Then, we describe the language models we use and how we fine-tune them.

5.1 Experimental settings

In our study, two different experimental settings are considered. The first is the most traditional one and consists in fine-tuning the language models using the training and validation sets of *ITA-SENS*. Then we test the learnt models on the test set and report the results. In the second experiment, we exploit the zero-shot transfer capabilities of multilingual models in order to perform zero-shot cross-lingual transfer learning [52, 19]. It consists in the transfer of the knowledge learned using the data available for a reference language (English in our case) to solve the task in another target language (Italian in our application). This is useful when the former language is a high-resource one and latter is a lower-resource one. To this end, we first fine-tune the model on the English corpus (*SENS2+OMC*), then, we transfer the learned model on Italian, performing the inference on the test set of *ITA-SENS*.

All the experiments have been implemented in Python with the support of some libraries, especially PyTorch, Scikit-learn and Keras, and have been executed on a server with 32 Intel Xeon Skylake cores running at 2.1 GHz, 256 GB RAM, and one NVIDIA Tesla T4 GPU. In the remainder of the section, we provide the details about the different language models used in our study.

5.2 Language models

In this section, we present the language models (LMs) used in our comparative analysis and, for each of them, we provide the details on parameter fine-tuning.

Transformer-based LMs As first category of methods, we consider several language models with a Transformer architecture, based on attention mechanisms [58]. These models are pre-trained on large text corpora (e.g. Wikipedia, CommonCrawl, Europarl, Books) on one or more typical NLP tasks (e.g. Next Sentence Prediction, Masked Language Modeling). Pre-training is useful to learn general language patterns and features, and avoids training the models from scratch. Consequently, it reduces the computation costs. More in detail, we take into account the most popular multilingual Transformers: mBERT [24] and XLM-RoBERTa [20]. We also consider monolingual versions specifically trained for Italian: AlBERTo [49] based on BERT, GilBERTo [50] and UmBERTo[7] [44] based on RoBERTa. We employ their respective versions for sequence classification made available in *HuggingFace*[8] by means of *Transformers APIs*, as they have already a linear layer for sequence classification on top of the pooled output. We use AdamW optimizer [38] with linear scheduler, with $\epsilon = 10^{-8}$ as default value, and the Binary Cross Entropy as loss function.

[7] We consider two versions of UmBERTo: *wikipedia-uncased*, uncased version trained on Wikipedia; *commoncrawl-cased*, cased version trained on CommonCrawl.

[8] https://huggingface.co/

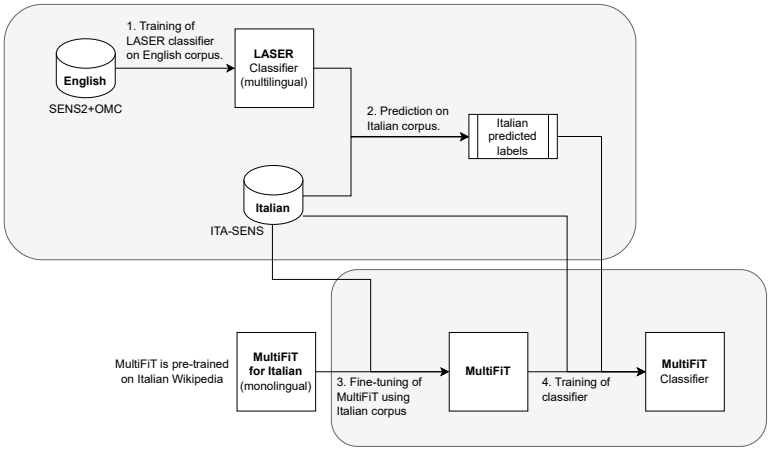

Fig. 1. Bootstrapping method adopted to perform zero-shot transfer with MultiFiT using LASER classifier as cross-lingual teacher.

LASER As first alternative multilingual model, we consider LASER, a new architecture to learn joint multilingual sentence embeddings for over 93 languages, proposed by Meta [3]. More specifically, LASER uses a single Bidirectional LSTM (long short-term memory) encoder with a shared BPE (Byte Paired Encoding) vocabulary for all languages, which, in its turn, is coupled with an auxiliary decoder and pre-trained on parallel corpora. We refer the reader to the original paper for further details. In our experiments, we use the pre-trained model downloaded from the official GitHub repository[9] as follows: first, we encode all input texts into LASER embeddings. Then, we create a sequential neural network that works as a decoder for classification and consisting in: (i) an input layer taking a LASER embedding with fixed size of 1024 input neurons; (ii) 4 hidden dense layers with 512, 128, 32, 8 neurons, respectively, LeakyReLU as activation function, dropout rate of 0.25 and batch normalization; (iii) a dense output layer with one neuron, which produces the predicted class, using the sigmoid as activation function. For the learning process, we use the AdamW optimizer with default value of $\epsilon = 10^{-8}$ and Binary Cross Entropy as loss function.

MultiFiT As second alternative multilingual model, we use MultiFiT, an alternative approach for fine-tuning monolingual models proposed by Eisenschlos *et al.* [26]. MultiFiT is an extended version of ULMFiT [29], designed to enhance its efficiency and applicability to NLP tasks in languages other than English. Its architecture includes subword tokenization and a Quasi-Recurrent Neural Network (QRNN) [15]. Instead of fine-tuning the classifier directly, MultiFiT first fine-tune the pre-trained model on the input corpus and then use that as the base for the classifier. To this purpose, several monolingual models, including

[9] https://github.com/facebookresearch/LASER

for Italian, are pre-trained on Wikipedia on the Next Word Prediction task. Additionally, the authors recommend to apply one-cycle policy with cosine annealing [55] and label smoothing techniques: the former is to make the training and convergence of complex models faster, the latter to avoid overfitting and overconfidence. We use the pre-trained model for Italian downloaded from the official GitHub repository[10] and rely on the *fast.ai* Python library in order to fine-tune and train both the language model and the classifier.

It is worth briefly describing the bootstrapping method we adopt to perform zero-shot cross lingual transfer learning, proposed by the authors of the paper and illustrated in Fig. 1. First, we exploit a LASER classifier previously trained on the English corpus (*SENS3+OMC*) as cross-lingual teacher and we make inference on the Italian corpus (*ITA-SENS*) obtaining the predicted labels. Then, we perform zero-shot transfer with MultiFiT for Italian pre-trained on Wikipedia: we fine-tune MultiFiT on *ITA-SENS* and train the classifier on top using the pseudo-labels predicted by LASER.

5.3 Hyperparameter selection

For each language model, with the exception of MultiFiT, we tune the hyperparameters by applying a grid search over a set of pre-defined values for the learning rate and for the batch size. To avoid overfitting, we use the early stopping criterion on the validation loss, initializing the epoch number to 100. The selected hyperparameter values of each language model in both the categories considered here are listed in Table 2. It is worth noting that all multilinguage models (mBERT, XML-RoBERTa, LASER and MultiFiT) have been configured for both experiments types: in the first one they are trained with *ITA-SENS*, as for the other monolanguage models, in the second one they are trained on *SENS2+OMC* and tested on *ITA-SENS*.

6 Results

In this section, we show and discuss the classification results of two experiments. In the first one, all language models are trained directly on *ITA-SENS*. In the second experiment, the models capable of performing cross-language transfer are trained on the English corpus and transferred to the Italian one. In both experiments, the results are reported for the test set of *ITA-SENS*.

We compare the different language models by computing the following evaluation metrics: accuracy, precision, recall, macro F1-score and Matthews correlation coefficient (MCC) [17]. Although accuracy and F1-score are the most popular metrics for evaluating binary classifiers, they can show overoptimistic inflated results, especially on unbalanced datasets. MCC is more reliable and yields a high score only if the outcome of the prediction is such that the values

[10] https://github.com/n-waves/multifit

Table 2. Best hyperparameter values for each language model considered in this study.

Experimental setting	Language Model	Batch size	Learning rate	# Epochs
Traditional ITA → ITA	mBERT	32	$5 \cdot 10^{-7}$	4
	XLM-RoBERTa	32	$1 \cdot 10^{-6}$	9
	AlBERTo	32	$5 \cdot 10^{-7}$	5
	GilBERTo	16	$5 \cdot 10^{-7}$	4
	UmBERTo-wiki	32	$5 \cdot 10^{-6}$	3
	UmBERTo-commoncrawl	32	$2 \cdot 10^{-6}$	3
	LASER	32	$2 \cdot 10^{-5}$	28
	MultiFiT	20	$1 \cdot 10^{-3}$	8
Zero-shot ENG → ITA	mBERT	32	$1 \cdot 10^{-6}$	3
	XLM-RoBERTa	32	$5 \cdot 10^{-6}$	2
	LASER	32	$2 \cdot 10^{-5}$	28
	MultiFiT	20	$1 \cdot 10^{-3}$	7

of the four confusion matrix categories (true positives, true negatives, false positive and false negatives) are all high, proportionally to the number of positive and negative samples in the dataset. In fact it is defined as

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP, TN, FP, FN represent the entries of the confusion matrix. The greater their correlations, the more accurate the model.

6.1 Experiment 1: language models trained on ITA-SENS

Table 3 reports the results obtained by the language models trained on *ITA-SENS* directly. We indicate the metrics on the columns and the methods on the rows. The language models are grouped by type: Transformer-based multilingual ones, Transformer-based monolingual ones, LASER and MultiFiT.

The language models with the highest accuracy are MultiFiT (0.9366), AlBERTo (0.9260) and UmBERTo-wikipedia-uncased (0.9260). It is worth noting that, since accuracy is sensitive to class unbalance, the models cannot be compared based on this metric only. For our task, it is important to analyze the recall on the positive class as well, as we want the model to capture all posts which are really sensitive. The model with the highest recall is UmBERTo-wikipedia-uncased (0.9458), followed by XLM-RoBERTa (0.9387) that, however, has a lower accuracy (0.9181). Other models with almost the same recall as XML-RoBERTa are AlBERTo and MultiFiT, both with a value equal to 0.9352. As regards the precision, MultiFiT is the model that predicts the highest percentage (95%) of posts truly belonging to the positive class as sensitive. It is followed by UmBERTo commoncrawl-cased (0.9347). Unlike MultiFiT, its precision is much higher than its recall. The only Transformer-based model exhibiting the highest precision and the highest recall at the same time is AlBERTo. On the other hand,

Table 3. Results of the model trained on *ITA-SENS*, reported for the test set.

Model	Accuracy	Precision	Recall	F1-score	MCC
mBERT	0.8752	0.8691	0.9151	0.8722	0.746
XLM-RoBERTa	0.9181	0.9171	0.9387	0.9166	**0.927**
AlBERTo (it)	0.9260	0.9330	0.9352	0.9249	0.855
GilBERTo (it)	0.9098	0.9228	0.9157	0.9086	0.817
UmBERTo-wiki-U (it)	0.9260	0.9240	**0.9458**	0.9246	0.850
UmBERTo-CC-C (it)	0.9187	0.9347	0.9193	0.9177	0.836
LASER	0.9125	0.9168	0.9281	0.9110	0.822
MultiFiT (it)	**0.9366**	**0.9508**	0.9352	**0.9358**	0.872

precision does not gives us any information on the number of posts of the positive class that are not labeled correctly. Therefore, to capture the balance between precision and recall, we compare all language models according to the macro F1-score. Among the Transformer models, AlBERTo (0.9249) and UmBERTo wikipedia (0.9246) have the highest value. Despite that, MultiFiT continues to outperform all models also in terms of macro F1-score: its value (0.9358), indeed, is the highest one. Finally, as precision, recall and macro F1-score ignore the true negatives, we also analyze the Matthews correlation coefficient (MCC). In this case, the best model (XLM-RoBERTa with a MCC of 0.927) has also a lower accuracy, precision and F1-score than the other models discussed above. Interestingly, XLM-RoBERTa is followed by MultiFiT with MCC equal to 0.872, thus confirming the predominance of this latter model in this experiment.

Summing up, monolingual models (AlBERTo, UmBERTo-wikipedia and MultiFiT) achieve the highest performances in our first experiment. More specifically, MultiFiT, the only model based on Quasi-Recurrent Neural Networks, always ranks first or second for all metrics considered in our study, outperforming all Transformer-bases models in the content sensitivity analysis task.

6.2 Experiment 2: zero-shot cross-lingual transfer learning

Table 4 shows the classification results obtained when multilanguage models, trained on the English corpus, transfer the acquired knowledge on Italian.

The models achieving the highest accuracy are MultiFiT (0.7487) and LASER (0.7411). Both models also correctly classify the highest percentage of true positives, as their recall is, respectively, 0.7550 and 0.7544. They also have the highest macro F1-scores (0.7463 and 0.7383, respectively). When the precision on the positive class is considered, XLM-RoBERTa predicts the highest percentage (about 88%) of posts belonging to the positive class as sensitive, but this model also has a relatively low recall. Consequently, the macro F1-score of XLM-RoBERTa (0.7297) is similar to that of LASER and MultiFiT which, however, have a more balanced precision and recall values. The same applies to mBERT, which has an even lower macro F1-score (0.6990). Finally, it can be observed that XLM-RoBERTa has the highest MCC (0.508), which means that this model exhibits a high correlation between real and predicted labels despite

Table 4. Results of zero-shot cross-lingual transfer from English to Italian.

Model	Accuracy	Precision	Recall	F1-score	MCC
mBERT	0.7002	0.8458	0.5689	0.6990	0.447
XLM-RoBERTa	0.7306	**0.8814**	0.6001	0.7297	**0.508**
LASER	0.7411	0.7773	0.7544	0.7383	0.477
MultiFiT	**0.7487**	0.7879	**0.7550**	**0.7463**	0.494

having a much lower accuracy and recall than LASER and MultiFiT, whose MCC is not far from the one of XLM-RoBERTa (0.494).

In conclusion, despite being trained on noisy pseudo-labels predicted by the LASER classifier, even in the zero-shot learning scenario, MultiFiT turns out to be the most accurate language model and seems to be the best choice for solving a content sensitivity analysis task.

6.3 Discussion and limitations

As seen in the previous sections, many language models achieve either high accuracy and F1-score or high values of the MCC, despite the task being generally known as difficult. One may argue that this is due to the fact that the language models are not learning to discriminate between sensitive and non sensitive posts; rather, they are learning to distinguish the sources of the posts (Insegreto or Twitter). In fact this could be a possible bias in our study and we think that, in part, this could explain the very good results obtained in the previous experiments. To try to dispel any doubt, we set up an additional experiment using the best classification models trained on *ITA-SENS* as predictors and, as test set, a further annotated single-source Italian corpus. As our goal is to study how *ITA-SENS* is adapted to the specific goal of content sensitivity analysis, we do not use the transfer learning models trained on *SENS2+OMC* (the English corpus) here. However, since we do not have an additional unbiased and annotated corpus of Italian posts (which will be part of our future work), we use a collection of posts taken from the *OMC* dataset and translated into Italian using DeepL[11], a famous and accurate automatic translator based on transformers. To limit the number of mistranslated posts, we include in this set only those posts with at least 20 words, as it is known that neural machine translation struggle with short texts [61]. The final dataset consists of 4 380 posts (3 182 sensitive and 1 198 non sensitive posts) annotated as described in Section 4.2. The results are reported in Table 5. As expected, the results are lower than those obtained when we apply the models to the test set of *ITA-SENS*. However, they are in general way better than a baseline consisting of a classifier assigning all posts to the majority class (the sensitive one in this case). In this experiments, the most accurate models according to the different performance indicators are mBERT (for the accuracy), MultiFiT (for the precision), XLM-RoBERTa (for the recall, if we exclude the baseline that, as expected, achieve 100% recall on the sensitive

[11] https://www.deepl.com/it/translator

Table 5. Results on the translated version of OMC using (some of) the language models trained on ITA-SENS compared with the majority class classifier as baseline.

Language Model	Accuracy	Precision	Recall	F1-score	MCC
Baseline	0.7264	0.7264	1.0	0.4207	0.0
mBERT	**0.7312**	0.7519	0.9402	0.5497	0.180
XLM-RoBERTa	0.7184	0.7392	**0.9462**	0.5053	0.104
AlBERTo	0.6212	0.7999	0.6382	0.5820	0.193
GilBERTo	0.6863	0.7880	0.7771	**0.6097**	**0.220**
UmBERTo-wiki-U	0.6214	0.7741	0.6763	0.5646	0.141
UmBERTo-CC-C	0.7166	0.7380	0.9456	0.5011	0.095
LASER	0.6687	0.7857	0.7479	0.5985	0.199
MultiFiT	0.6262	**0.8103**	0.6338	0.5906	0.216

class) and GiLBERTo (for the macro F1-score and the MCC). More interestingly, the results are in line with (and even better than) those reported in [12] for the same dataset, where, however, the models where trained and tested on English only. According to these results, even if we can not totally exclude the source bias, we can safely confirm the conclusion drawn at the end of our experiments.

7 Conclusion

In this paper, we have proposed a new corpus specifically annotated for the content sensitivity analysis task. We use it to feed several state-of-the-art language models based on transformers and attention mechanisms to detect privacy-sensitive content in social media posts. We also show the performances of different multilingual models, including two alternative architectures based on bidirectional long short-term memory and quasi-recurrent neural networks. We have set up two different experiments, also including zero-shot cross-lingual transfer methods where the model is trained with an English corpus and tested on Italian posts. Despite some promising results, the models trained directly with the Italian corpus are still the best performing ones. Some limitations will be addressed in future works. First, our corpus is annotated following the anonymity assumption, which consists in labeling as sensitive every content item posted anonymously, and as non sensitive contents posted publicly with identifiable profiles. However, as shown by some recent work, this assumption does not hold in many cases, hence we plan to launch a manual annotation campaign involving several domain experts. Second, in this work we only rely on text posts, but it is well-known that the most successful social media platforms are now fostering the sharing of audio-visual content, such as images and short videos. As future work we will investigate on sensitive content in such modalities and, more specifically, on multimodal content sensitivity analsyis, to exploit the manifold of the information provided by different representations of the same posted message.

Acknowledgements The work presented in this paper is supported by Fondazione CRT (Grant No. 2022-0720).

References

1. Ahmad, Z., Jindal, R., Ekbal, A., Bhattachharyya, P.: Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. Expert Systems with Applications **139**, 112851 (2020)
2. Alemany, J., del Val Noguera, E., Alberola, J.M., García-Fornes, A.: Metrics for Privacy Assessment When Sharing Information in Online Social Networks. IEEE Access **7**, 143631–143645 (2019)
3. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguistics **7**, 597–610 (2019)
4. Baiocco, R., Laghi, F., Di Pomponio, I., Nigito, C.S.: Self-disclosure to the best friend: Friendship quality and internalized sexual stigma in italian lesbian and gay adolescents. Journal of Adolescence **35**(2), 381–387 (2012)
5. Barak, A., Gluck-Ofri, O.: Degree and reciprocity of self-disclosure in online forums. Cyberpsychology Behav. Soc. Netw. **10**(3), 407–417 (2007)
6. Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: Overview of the evalita 2016 sentiment polarity classification task. In: Proceedings of CLiC-it 2016 & EVALITA 2016. CEUR-WS.org (2016)
7. Barth, S., de Jong, M.D.T.: The privacy paradox - investigating discrepancies between expressed privacy concerns and actual online behavior - A systematic literature review. Telematics Informatics **34**(7), 1038–1058 (2017)
8. Battaglia, E., Bioglio, L., Pensa, R.G.: Towards content sensitivity analysis. In: Berthold, M.R., Feelders, A., Krempl, G. (eds.) Proceedings of IDA 2020, Konstanz, Germany, April 27-29, 2020. pp. 67–79. Springer (2020)
9. Baziotis, C., Pelekis, N., Doulkeridis, C.: DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In: Proceedings of SemEval-2017. pp. 747–754. ACL (2017)
10. Bianchi, F., Nozza, D., Hovy, D.: FEEL-IT: emotion and sentiment classification for the italian language. In: Proceedings of WASSA@EACL 2021. pp. 76–83. ACL (2021)
11. Biega, J.A., Gummadi, K.P., Mele, I., Milchevski, D., Tryfonopoulos, C., Weikum, G.: R-Susceptibility: An IR-Centric Approach to Assessing Privacy Risks for Users in Online Communities. In: Proceedings of ACM SIGIR 2016. pp. 365–374 (2016)
12. Bioglio, L., Pensa, R.G.: Analysis and classification of privacy-sensitive content in social media posts. EPJ Data Sci. **11**(1), 12 (2022)
13. Blanco-Herrero, D., Rodríguez-Contreras, L.: The risks of new technologies in black mirror: A content analysis of the depiction of our current socio-technological reality in a TV series. In: González, M.Á.C., Rodríguez-Sedano, F.J., Llamas, C.F., García-Peñalvo, F.J. (eds.) Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality, TEEM 2019, León Spain, October, 2019. pp. 899–905. ACM (2019)
14. Bosco, C., Patti, V., Frenda, S., Cignarella, A.T., Paciello, M., D'Errico, F.: Detecting racial stereotypes: An italian social media corpus where psychology meets NLP. Inf. Process. Manag. **60**(1), 103118 (2023)
15. Bradbury, J., Merity, S., Xiong, C., Socher, R.: Quasi-recurrent neural networks. In: Proceedings of ICLR 2017. OpenReview.net (2017)
16. Celli, F., Pianesi, F., Stillwell, D., Kosinski, M.: Workshop on Computational Personality Recognition: Shared Task. In: Proceedings of ICWSM 2013 (2013)

17. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics **21**, 1–13 (2020)
18. Choi, H., Park, J., Jung, Y.: The role of privacy fatigue in online privacy behavior. Computers in Human Behavior **81**, 42–51 (2018)
19. Choi, H., Kim, J., Joe, S., Min, S., Gwon, Y.: Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9608–9613. IEEE (2021)
20. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
21. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 7057–7067 (2019)
22. Correa, D., Silva, L.A., Mondal, M., Benevenuto, F., Gummadi, K.P.: The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In: Proceedings of ICWSM 2015. pp. 71–80 (2015)
23. Danet, M., Miljkovitch, R., Deborde, A.S.: Online self-disclosure: Validation study of the french version of the real me on the net questionnaire. Current Psychology **39**, 2366–2370 (9 2018)
24. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
25. Dong, X., de Melo, G.: Cross-lingual propagation for deep sentiment analysis. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 5771–5778. AAAI Press (2018)
26. Eisenschlos, J., Ruder, S., Czapla, P., Kardas, M., Gugger, S., Howard, J.: MultiFiT: Efficient Multi-lingual Language Model Fine-tuning. In: Proceedings of EMNLP-IJCNLP 2019. pp. 5701–5706. ACL (2019)
27. El Ouirdi, M., Segers, J., El Ouirdi, A., Pais, I.: Predictors of job seekers' self-disclosure on social media. Computers in Human Behavior **53**, 1–12 (2015)
28. Gill, A.J., Vasalou, A., Papoutsi, C., Joinson, A.N.: Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In: Proceedings of ACM CHI 2011. pp. 3227–3236 (2011)
29. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of ACL 2018. pp. 328–339. ACL (2018)
30. Jaidka, K., Guntuku, S., Ungar, L.: Facebook versus twitter: Differences in self-disclosure and trait prediction. In: Proceedings of ICWSM 2018. pp. 141–150. AAAI Press (2018)
31. Jaidka, K., Singh, I., Liu, J., Chhaya, N., Ungar, L.: A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure. In: Proceedings of AffCon@AAAI 2020. pp. 118–129. CEUR-WS.org (2020)
32. Jourard, S.M.: Self-disclosure: An experimental analysis of the transparent self. John Wiley (1971)
33. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. PNAS **110**(15), 5802–5805 (2013)

34. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of ACL 2020. pp. 7871–7880. ACL (2020)
35. Liu, D., Brown, B.B.: Self-disclosure on social networking sites, positive feedback, and social capital among chinese college students. Computers in Human Behavior **38**, 213–219 (2014)
36. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. TKDD **5**(1), 6:1–6:30 (2010)
37. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Linguistics **8**, 726–742 (2020)
38. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: Proceedings of ICLR 2019. OpenReview.net (2019)
39. Ma, X., Hancock, J.T., Naaman, M.: Anonymity, intimacy and self-disclosure in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016. pp. 3857–3869. ACM (2016)
40. Mondal, M., Correa, D., Benevenuto, F.: Anonymity effects: A large-scale dataset from an anonymous social media platform. In: Gadiraju, U. (ed.) Proceedings of ACM HT 2020, Virtual Event, USA, July 13-15, 2020. pp. 69–74. ACM (2020)
41. Oukemeni, S., Rifà-Pous, H., i Puig, J.M.M.: IPAM: Information Privacy Assessment Metric in Microblogging Online Social Networks. IEEE Access **7**, 114817–114836 (2019)
42. Oukemeni, S., Rifà-Pous, H., i Puig, J.M.M.: Privacy Analysis on Microblogging Online Social Networks: A Survey. ACM Comput. Surv. **52**(3), 60:1–60:36 (2019)
43. Pan, X., Wang, M., Wu, L., Li, L.: Contrastive learning for many-to-many multilingual neural machine translation. In: Proceedings of ACL/IJCNLP 2021. pp. 244–258. ACL (2021)
44. Parisi, L., Francia, S., Magnani, P.: Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto (2020)
45. Peddinti, S.T., Korolova, A., Bursztein, E., Sampemane, G.: Cloak and Swagger: Understanding Data Sensitivity through the Lens of User Anonymity. In: Proceedings of IEEE SP 2014. pp. 493–508 (2014)
46. Peddinti, S.T., Ross, K.W., Cappos, J.: User Anonymity on Twitter. IEEE Security & Privacy **15**(3), 84–87 (2017)
47. Pensa, R.G., Di Blasi, G.: A privacy self-assessment framework for online social networks. Expert Syst. Appl. **86**, 18–31 (2017)
48. Pensa, R.G., Di Blasi, G., Bioglio, L.: Network-aware privacy risk estimation in online social networks. Social Netw. Analys. Mining **9**(1), 15:1–15:15 (2019)
49. Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V., et al.: Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In: CEUR Workshop Proceedings. vol. 2481, pp. 1–6. CEUR (2019)
50. Ravasio, G., Di Perna, L.: Gilberto: An italian pretrained language model based on roberta. https://github.com/idb-ita/GilBERTo (2020)
51. Ren, S., Wu, Y., Liu, S., Zhou, M., Ma, S.: Explicit cross-lingual pre-training for unsupervised machine translation. In: Proceedings of EMNLP-IJCNLP 2019. pp. 770–779. ACL (2019)
52. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)

53. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M.: An italian twitter corpus of hate speech against immigrants. In: Proceedings of LREC 2018. ELRA (2018)
54. Schroepfer, M.: An update on our plans to restrict data access on Facebook. https://about.fb.com/news/2018/04/restricting-data-access/ (2018)
55. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. CoRR **abs/1803.09820** (2018), http://arxiv.org/abs/1803.09820
56. Tang, D., Chou, T., Drucker, N., Robertson, A., Smith, W.C., Hancock, J.T.: A tale of two languages: strategic self-disclosure via language selection on facebook. In: Proceedings of ACM CSCW 2011. pp. 387–390. ACM (2011)
57. Vasalou, A., Gill, A.J., Mazanderani, F., Papoutsi, C., Joinson, A.N.: Privacy dictionary: A new resource for the automated content analysis of privacy. JASIST **62**(11), 2095–2105 (2011)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
59. Vulic, I., Moens, M.: Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In: Proceedings of ACM SIGIR 2015. pp. 363–372. ACM (2015)
60. Wagner, I., Eckhoff, D.: Technical privacy metrics: A systematic survey. ACM Comput. Surv. **51**(3), 57:1–57:38 (2018)
61. Wan, Y., Yang, B., Wong, D.F., Chao, L.S., Yao, L., Zhang, H., Chen, B.: Challenges of neural machine translation for short texts. Comput. Linguistics **48**(2), 321–342 (2022)
62. Wang, D., Chen, J., Zhou, H., Qiu, X., Li, L.: Contrastive Aligned Joint Learning for Multilingual Summarization. In: Proceedings of ACL/IJCNLP 2021. pp. 2739–2750. ACL (2021)
63. Wang, Y., Burke, M., Kraut, R.E.: Modeling self-disclosure in social networking sites. In: Proceedings of ACM CSCW 2016. pp. 74–85. ACM (2016)
64. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proceedings of NAACL-HLT 2021. pp. 483–498. ACL (2021)
65. Yang, D., Yao, Z., Kraut, R.E.: Self-disclosure and channel difference in online health support groups. In: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017. pp. 704–707. AAAI Press (2017)
66. Yu, J., Kuang, Z., Zhang, B., Zhang, W., Lin, D., Fan, J.: Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-Grained Privacy Settings for Social Image Sharing. IEEE Trans. Inf. Forensics Secur. **13**(5), 1317–1332 (2018)
67. Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J.: iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. IEEE Trans. Inf. Forensics Secur. **12**(5), 1005–1016 (2017)
68. Zlatolas, L.N., Welzer, T., Hericko, M., Hölbl, M.: Privacy antecedents for SNS self-disclosure: The case of facebook. Comput. Hum. Behav. **45**, 158–167 (2015)