Hurtlex: A Multilingual Lexicon of Words to Hurt

Elisa Bassignana and **Valerio Basile** and **Viviana Patti**
Dipartimento di Informatica
University of Turin
{basile,patti}@di.unito.it
elisa.bassignana@edu.unito.it

Abstract

English. We describe the creation of *HurtLex*, a multilingual lexicon of hate words. The starting point is the Italian hate lexicon developed by the linguist Tullio De Mauro, organized in 17 categories. It has been expanded through the link to available synset-based computational lexical resources such as MultiWordNet and BabelNet, and evolved in a multi-lingual perspective by semi-automatic translation and expert annotation. A twofold evaluation of HurtLex as a resource for hate speech detection in social media is provided: a qualitative evaluation against an Italian annotated Twitter corpus of hate against immigrants, and an extrinsic evaluation in the context of the AMI@Ibereval2018 shared task, where the resource was exploited for extracting domain-specific lexicon-based features for the supervised classification of misogyny in English and Spanish tweets.

Italiano. L'articolo descrive lo sviluppo di *Hurtlex*, un lessico multilingue di parole per ferire. Il punto di partenza è il lessico di parole d'odio italiane sviluppato dal linguista Tullio De Mauro, organizzato in 17 categorie. Il lessico è stato espanso sfruttando risorse lessicali sviluppate dalla comunità di Linguistica Computazionale come MultiWordNet e BabelNet e le sue controparti in altre lingue sono state generate semi-automaticamente con traduzione ed annotazione manuale di esperti. Viene presentata sia un'analisi qualitativa della nuova risorsa, mediante l'analisi di corpus di tweet italiani annotati per odio nei confronti dei migranti e una valutazione estrinseca, mediante l'uso della risorsa nell'ambito dello sviluppo di un sistema Automatic Misogyny Identification in tweet in spagnolo ed inglese.

1 Introduction

Communication between people is rapidly changing, in particular due to the exponential growth of the use of social media. As a privileged place for expressing opinions and feelings, social media are also used to convey expressions of hostility and *hate speech*, mirroring social and political tensions. Social media enable a wide and viral dissemination of hate messages. The extreme expressions of verbal violence and their proliferation in the network are progressively being configured as unavoidable emergencies. Therefore, the development of new linguistic resources and computational techniques for the analysis of large amounts of data becomes increasingly important, with particular emphasis on the identification of hate in language (Schmidt and Wiegand, 2017; Waseem and Hovy, 2016; Davidson et al., 2017).

The main objective of this work is the development of a lexicon of hate words that can be used as a resource to analyze and identify *hate speech* in social media texts in a multilingual perspective. The starting point is the lexicon '*Le parole per ferire*' developed by the Italian linguist Tullio De Mauro for the *"Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena* of the Italian Chamber of Deputies. The lexicon consists of more than 1,000 Italian hate words organized along different semantic categories of hate (De Mauro, 2016).

In this work, we present a computational version of the lexicon. The hate categories and lemmas have been represented in a *machine-readable* format and a semi-automatic extension and enrichment with additional information has been provided using lexical databases and ontologies. In particular we augmented the original Italian lexi-

con with translations in multiple languages.

HurtLex, the hate lexicon obtained with the method described in Section 3, has been tested with a corpus-based evaluation, through the analysis of a hate corpus of about 6,000 Italian tweets (Section 4.1), and through an extrinsic evaluation in the context of the shared task on *Automatic Misogyny Identification* at *IberEval 2018*, focusing on the identification of hate against women in Twitter in English and Spanish (Section 4.2).

The resource is available for download at `http://hatespeech.di.unito.it/resources.html`

2 Related Work

Lexical knowledge for the detection of hate speech, and abusive language in general, has received little attention in literature until recently. Even for English, there are few publicly available domain-independent resources — see for instance the novel lexicon of abusive words recently proposed by (Wiegand et al., 2018). Indeed, lexicons of abusive words are often manually compiled specifically for a task, thus they are rarely based on deep linguistic studies and reusable in the context of new classification tasks. Moreover, the lexical knowledge exploited in this context is often limited to inherently derogative words (such as slurs, swear words, taboo words). De Mauro (2016) highlights that this can be a restriction in the compilation of a lexicon of hate words, where the accent is also on derogatory epithets aimed at hurting weak and vulnerable categories of people, targeting individuals and groups of individuals on the basis of race, nationality, religion, gender or sexual orientation (Bianchi, 2014).

Regarding Italian, apart from the lexicon of hate words developed by Tullio De Mauro described in Section 3, the literature is sparse, but it is worth mentioning at least the study by Pelosi et al. (2017) on mining offensive language on social media and the project reported in D'Errico et al. (2018) on distinguishing between pro-social and anti-social attitudes. Both the works rely on the use of corpora of Facebook posts. In particular, in Pelosi et al. (2017) the focus is on automatically annotating hate speech in a corpus of posts from the Facebook page "Sesso Droga e Pastorizia", by exploiting a lexicon-based method using a dataset of Italian taboo expressions.

To conclude, let us mention that a new shared task on hate speech detection has been proposed in the context of the EVALITA 2018 evaluation campaign[1], which provides a stimulating setting for discussion on the role of lexical knowledge in the detection of hate in language.

3 Method

Our lexicon was created starting from preexisting lexical resources. In this section we give an overview of such resources and of the process we followed to create HurtLex.

3.1 "Parole per Ferire"

We started from the lexicon of "words to hurt" *Le parole per ferire* by the Italian linguist Tullio De Mauro (De Mauro, 2016). This lexicon includes more than 1,000 Italian words from 3 macro-categories: *derogatory* words (all those words that have a clearly offensive and negative value, e.g. slurs), words bearing *stereotypes* (typically hurting individuals or groups belonging to vulnerable categories) and words that are neutral, but which can be used to be derogatory in certain contexts through semantic shift (such as metaphor). The lexicon is divided into 17 finer-grained, more specific sub-categories that aim at capturing the context of each word (see also Table 1):

Negative stereotypes ethnic slurs (PS); locations and demonyms (RCI); professions and occupations (PA); physical disabilities and diversity (DDF); cognitive disabilities and diversity (DDP); moral and behavioral defects (DMC); words related to social and economic disadvantage (IS).

Hate words and slurs beyond stereotypes plants (OR); animals (AN); male genitalia (ASM); female genitalia (ASF); words related to prostitution (PR); words related to homosexuality (OM).

Other words and insults descriptive words with potential negative connotations (QAS); derogatory words (CDS); felonies and words related to crime and immoral behavior (RE); words related to the seven deadly sins of the Christian tradition (SVP).

3.2 Lexical Resources

WordNet (Fellbaum, 1998) is a lexical reference system for the English language based on psycholinguistic theories of human lexical memory.

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

Category	Percentage	Category	Percentage
PS	3,85%	ASM	7,07%
RCI	0,81%	ASF	2,78%
PA	7,52%	PR	5,01%
DDF	2,06%	OM	2,78%
DDP	6,00%	QAS	7,34%
DMC	6,98%	CDS	26,68%
IS	1,52%	RE	3,31%
OR	1,52%	SVP	4.83%
AN	9,94%		

Table 1: Distribution of sub-categories in *Le parole per ferire*.

Category	Percentage	Category	Percentage
PS	2,76%	ASM	6,21%
RCI	0,41%	ASF	1,66%
PA	5,38%	PR	1,66%
DDF	1,52%	OM	2,76%
DDP	8,55%	QAS	11,03%
DMC	7,45%	CDS	26,07%
IS	1,38%	RE	4,69%
OR	2,34%	SVP	6.07%
AN	10,07%		

Table 2: Distribution of the words not present in BabelNet along the 17 sub-categories of De Mauro.

WordNet is structured around *synsets* (sets of synonyms) and their 4 coarse-grained parts of speech: noun, verb, adjective and adverb.

MultiWordNet (Pianta et al., 2002), is an extension of WordNet that contains mappings between the English lexical items in Wordnet and lexical items of other languages, including Italian.

BabelNet (Navigli and Ponzetto, 2012) is a combination of a multilingual encyclopedic dictionary and a semantic network that links concepts and named entities in a very wide network of semantic relationships.

3.3 A Computational Lexicon of Hate Words

The first step for the creation of our lexicon consisted in extracting every item from the lexicon *Le parole per ferire*. We obtain 1,138 items, but 1,082 unique items because several items were duplicated in multiple categories. We also removed 10 lemmas that belong to idiomatic multi-word-expressions, e.g., "coccodrillo" (crocodile) in the expression "lacrime di coccodrillo" (crocodile tears), leaving us to 1,072 unique lemmas.

As a second step, we use MultiWordNet to augment the words with their part-of-speech tags. We use the Italian index of MultiWordNet, comprising, for each lemma, four fields containing the identifiers of the synsets in which the lemma is intended like a noun, an adjective, a verb and a pronoun. By joining this index with our lexicon, we obtain all the possible part-of-speech for 59,2 % of the lemmas, bringing the total number of lemmas from 1,072 to 1,156 to include duplicates with different part of speech. The remaining lemmas were annotated manually.

The third step consists of linking the lemmas of the lexicon with a definition. We use the BabelNet API to retrieve the definitions, aiming for high coverage. In total, we were able to retrieve a definition for 71,1% of the lemmas. Table 2 shows the distribution of the words not present in BabelNet across the HurtLex categories. All the information about the entries of HurtLex (lemma, part of speech, definition) and the hierarchy of categories is collected in one XML structured file for distribution in machine-readable format.

3.4 Semi-automatic Multilingual Extension of the Lexicon

We leverage BabelNet to translate the lexicon into multiple languages, by querying the API[2] to retrieve all the senses of all the words in the lexicon.

Next, we queried the BabelNet API again to retrieve all the lemmas in all the supported languages, thus creating a basis for a multilingual lexicon starting from an Italian resource.

Not surprisingly, some of the senses retrieved in the first step were unrelated to the offensive context, therefore their translation to other languages would generate unlikely candidates for a lexicon of hate words. For instance, BabelNet senses of named entities which are homograph to words in the input lexicon are extracted along with the other senses, but they are typically to exclude from a resource such as HurtLex.

Therefore, we performed a manual filtering of the senses prior to the automatic translation, with the aim of translating the original words only according to their offensive meaning. We manually annotated each pair lemma-sense according to one of three classes: **Not offensive** (used for senses that are totally unrelated to any offensive context), **Neutral** (senses that are not inherently offensive, but are linked to some offensive use of the word, for example by means of a semantic shift), and **Offensive** (senses that embody a crystallized offensive use of a word). To check the consistency

[2] https://babelnet.org/guide#java

Definition	Annotation
Finocchio is a station of Line C of the Rome Metro.	Not offensive
Aromatic bulbous stem base eaten cooked or raw in salads.	Neutral[3]
Offensive term for an openly homosexual man.	Offensive

Table 3: Annotation of three senses of the Italian word "Finocchio".

Category	Occurrence	Category	Occurrence
RE	45,10%	DDP	1,90%
QAS	23,32%	IS	1,60%
CDS	8,30%	SVP	0,50%
PS	7,10%	RCI	0,30%
ASM	2,70%	PR	0,30%
OM	2,20%	DDF	0,30%
AN	2,10%	OR	0,20%
PA	2,00%	ASF	0,00%
DMC	1,90%		

Table 4: Percentage of messages in the hate speech corpus containing words from the 17 HurtLex categories.

of the annotation, a subset of 200 senses were annotated by two experts, reporting an agreement on 87.6% of the items. Table 3 shows examples of the different annotation of senses of the same word.

After discussing the results of the pilot annotation, we decided to split the *Neutral* class into two additional classes. One of the new classes covers the cases where a sense is **not literally pejorative**, but it is used to insult by means of a semantic shift, e.g. metaphorically. The other additional class is for the senses which have a clear **negative connotation**, but not necessarily a direct derogatory use in a derogatory way, e.g., the main senses of "criminal". Subsequently, the lexicon was annotated by two other experts reporting an agreement on 61% of the items. Most disagreement was concentrated in the distinctions *Not offensive*/*Not literally pejorative* (43% of the disagreement cases) and *Negative connotation*/*Offensive* (25% of the disagreement cases).

After the annotation, we discarded all the senses marked "not offensive", and created two different versions of the multilingual lexicon in 53 languages: one containing only the translations of "offensive" senses (more conservative), and the other containing translations of "offensive", "not literally pejorative" and "negative connotation" senses (more inclusive).

4 Evaluation

We evaluated the quality of the lexicon of hate words created with the method described in the previous section in two settings: by studying the occurrence of its words and their categories in a corpus of hate speech (Section 4.1), and by extracting features from HurtLex for supervised classification of misogyny in social media text (Section 4.2).

4.1 Qualitative Evaluation

In order to gain insights on the composition of the HurtLex lexicon, we evaluated it against an annotated corpus of Hate Speech on social media, recently published by Sanguinetti et al. (2018b). The corpus consists of 6,008 tweets selected according to keywords related to immigration and ethnic minorities. Each tweet in the corpus is annotated following a rich schema, including hate speech (yes/no), aggressiveness (strong/weak/none), offensiveness (strong/weak/none), irony (yes/no) and stereotype (yes/no).

We searched the lemmas of HurtLex in the version of the hate speech corpus enriched with Universal Dependencies annotations[4], by matching the pairs (lemma, POS-tag) in HurtLex with the morphosyntactic annotation of the corpus, and computed several statistics on the actual usage of such words in a specific abusive context of hate against immigrants. Table 4 shows the rate of messages in the corpus featuring words from each HurtLex category in the corpus.

For a more in-depth analysis, we also examined the relative frequency of single words in HurtLex with respect to the finer-grained annotation of the messages where they occur. Figures 1, 2, 3, 4 and 5 show examples of such analysis.
It can be noted how the relative frequency of words like "terrorismo" (*terrorism*), "ladro" (*thief*) and "rubare" (*stealing*) decrease drastically as the tweets become more aggressive, *offensive* or with a higher level of hate speech (perhaps because, albeit negative, they are not swear words)), while

[3]The derogatory use of the word "finocchio" (fennel) in Italian is thought to originate from the middle ages, linking the fennel plant to the execution of gay men at the burning stake.

[4]The corpus of hate speech by Sanguinetti et al. (2018b) has been annotated with a method similar to that described in Sanguinetti et al. (2018a).

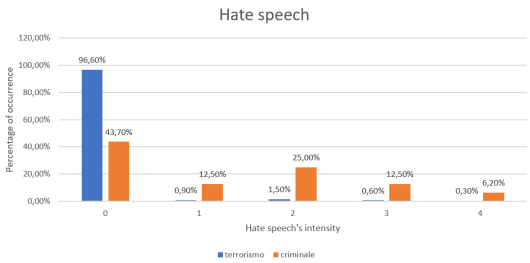

Figure 1: Relative frequency of the words "terrorismo" (*terrorism*) and "criminale" (*criminal*) with respect to the hate speech annotation.

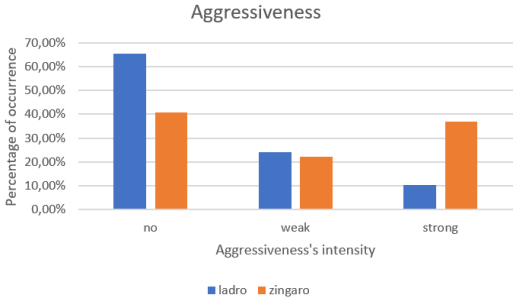

Figure 2: Relative frequency of the words "ladro" (thief) and "zingaro" (gypsy) with respect to the aggressiveness annotation.

words like "bastardo" (*bastard*) occur more as the tweets become more offensive (possibly also because they belong to the swearing sphere). Another class of words, like "zingaro" (*gypsy*), show a parabolic distribution. We hypothesize that this behavior is typical of words with an apparently neutral connotation that are sometimes used in abusive context with an offensive connotation. We plan to leverage this method of analysis for further studies on this line.

4.2 Misogyny Identification on Social Media

HurtLex was one of the resources used by the Unito's team to participate to the shared task *Automatic Misogyny Identification* (AMI) at IberEval 2018 (Pamungkas et al., 2018). The task consists of identifying misogynous content in Twitter messages (first sub-task) and classifying their misogynist behavior (second sub-task). The Unito's team employed different subsets of the 17 categories of HurtLex by extracting lexicon-based features for a supervised classifier. They identified the *Prostitution*, *Female and Male Sexual Apparatus* and *Physical and Mental Diversity and Disability* categories as the most informative for this task. The

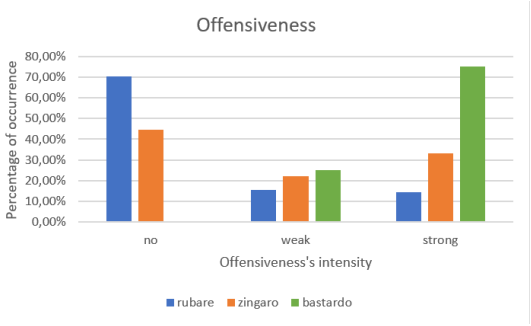

Figure 3: Relative frequency of the words "rubare" (stealing), "zingaro" (*gypsy*) and "bastardo" (bastard) with respect to the offensiveness annotation.

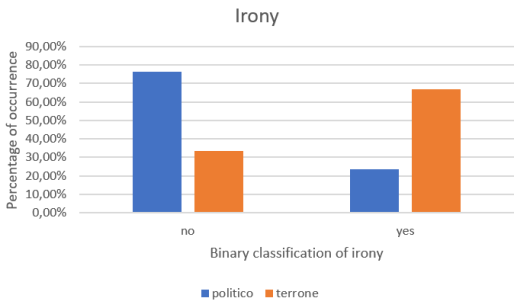

Figure 4: Relative frequency of the words "politico" (politician) and "terrone" (slur referring to southern Italians) with respect to the irony annotation.

Unito classifier obtained the best result in the first sub-task for both languages and the best result in the second sub-task for Spanish.

5 Conclusion and Future Work

Our main contribution is a machine-readable version of the hate words lexicon by De Mauro, enriched with lexical features from available computational resources. We make *HurtLex* available for download as a tool for hate speech detection. A first evaluation of the lexicon against corpora featuring different targets of hate (immigrants and women) has been presented. The multilingual evaluation of HurtLex showed also promising results. Although we are aware that hate speech-related phenomena tend to follow regional and cultural patterns, our semi-automatically produced resource was able to partially fill the gap towards hate speech detection in less represented languages. To this end, we aim at investigating the potential and pitfalls of semi-automating mappings further. In particular, two possible ex-

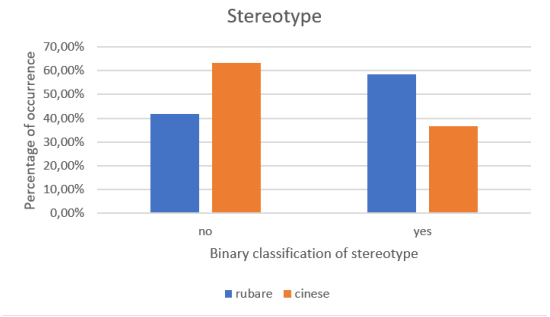

Figure 5: Relative frequency of the words "rubare" (stealing) and "cinese" (chinese) with respect to the stereotype annotation.

tensions of our method involve using distributional semantic models to automatically expand the lexicon with synonyms and lemmas semantically related to the original ones, and exploiting De Mauro's derivational rules.

Acknowledgments

Valerio Basile and Viviana Patti were partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media-IhatePrejudice*, S1618_L2_BOSC_01).

References

Claudia Bianchi. 2014. The speech acts account of derogatory epithets: some critical notes. In J. Dutant, D. Fassio, and Meylan A., editors, *Liber Amicorum Pascal Engel, University of Geneva*, pages pp. 465–480.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. 27 settembre 2016. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena, of the Italian Chamber of Deputies, which issued a Final Report in 2017.

Francesca D'Errico, Marinella Paciello, and Matteo Amadei. 2018. Prosocial words in social media discussions on hosting immigrants. insights for psychological and computational field. In *Symposium on Emotion Modelling and Detection in Social Media and Online Interaction, In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)*.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. 14-ExLab@UniTo for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. In *Proc. of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with SEPLN 2018)*, volume 2150 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Serena Pelosi, Alessandro Maisto, Pierluigi Vitale, and Simonetta Vietri. 2017. Mining offensive language on social media. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.*

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

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)*, Paris, France, may. European Language Resources Association (ELRA).

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. ACL.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056. Association for Computational Linguistics.