The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis

Marco Vassallo, Giuliano Gabrieli
CREA Research Centre
for Agricultural Policies and Bio-economy
{marco.vassallo,
giuliano.gabrieli}@crea.gov.it

Valerio Basile, Cristina Bosco
Department of Computer Science
University of Turin
{basile,bosco}@di.unito.it

Abstract

English. Sentiment Analysis (SA) based on an affective lexicon is popular because straightforward to implement and robust against data in specific, narrow domains. However, the morpho-syntactic pre-processing needed to match words in the affective lexicon (lemmatization in particular) may be prone to errors. In this paper, we show how such errors have a substantial and statistical significant impact on the performance of a simple dictionary-based SA model on data from Twitter in Italian. We test three pre-trained statistical models for lemmatization of Italian based on Universal Dependencies, and we propose a simple alternative to lemmatizing the tweets that achieves better polarity classification results.[1]

1 Introduction

In the last few years a very large variety of approaches has been proposed for addressing Sentiment Analysis (SA) related tasks. In several approaches, lexical resources play a crucial role: they allow systems to move from strings of characters to the semantic knowledge found, e.g., in an affective lexicon[2]. For achieving this result and calculating the polarity of sentiment, or of some related categories, some shallow morphological analysis has to be applied, which mostly consists in lemmatization.

When we refer to standard text, available resources and robust lemmatizers make lemmatization a practically solved issue, but the presence of misspellings, lingo and irregularities makes the application of lemmatization on user-generated content drawn from social media and micro-blogs not equally easy.

A possible solution consists in applying supervised machine learning techniques in order to create robust lemmatization models. However, the large manually curated datasets necessary for this task are currently very rare, in particular for languages other than English. For what concerns Italian, a good quality gold standard resource in Universal Dependency has been released which includes texts drawn from micro-blogs, namely PoSTWITA-UD (Sanguinetti et al., 2018). Unfortunately it is not nearly large enough to be of practical use in a supervised machine learning setting.

In this paper, we focus on the lemmatization of social media texts, observing and evaluating its impact on SA. The goal of this work is to address the following research questions: *what is the impact of lemmatization in SA tasks? Can we classify lemmatization errors and automatically adjust (a relevant portion of) them?*

We start from the empirical evidence found in a corpus of tweets from the agriculture domain that has initially raised our attention on this problem. After that, we present further experiments on a manually annotated dataset. We further propose some hints about a solution based on an affecting lexicon of inflected forms.

2 Datasets

We collected two datasets of microblogs in Italian language, in order to experiment on realistic data.

AGRITREND is a corpus of Italian posts collected from the Twitter accounts of the main institutional and media actors related to the agricultural sector during the period of January-April 2019. The data related to the first two months of the year have been used for the publication of the

[2]For an informal definition of *affective lexicon* see: http://www.ai-lc.it/lessici-affettivi-per-litaliano/

first issue of the Institutional bulletin of the CREA Research Centre for Agricultural Policies and Bioeconomy (Monda et al., 2019). Institutional motivations drove the initiative of setting up this corpus: exploring the sentiment in agriculture and thus providing insights about current and emerging trends of the agricultural sector. The dataset is composed of 8,883 tweets, including 2,554 retweets (28.75% of the total).

SENTIPOLC is the corpus distributed for the SENTIment POLarity Classification task (Barbieri et al., 2016) within the context of the evaluation campaign EVALITA 2016[3]. The corpus, consisting of 9,392 tweets, was created partly by querying Twitter for specific keywords and hashtags marking political topics, and partly with random tweets on any topic. Experts and crowdsourcing contributors annotated the dataset with subjectivity (binary classification: objective/subjective), polarity (4-fold multiclass classification: positive/negative/neutral/mixed) and irony (binary classification: ironic/not-ironic).

3 Processing the *AGRITREND* corpus

In this section, we describe the processing applied on the *AGRITREND* with the goal of SA, after the pre-processing which consisted in filtering out hashtags, @mentions, URLs and tokenization.

3.1 Lexicon-based Sentiment Analysis

While most modern SA approaches are supervised[4], our SA approach is unsupervised and based on an affective lexicon. However, given the narrow topic scope of our data of interest and the unavailability of annotated data for agriculture, the application of an unsupervised classifier allowed us to avoid domain adaptation issues. Moreover, the dictionary-based approach is more transparent, allowing us to evaluate its errors at a finer-grained lexical level.

The method is straightforward. Given a preprocessed tweet and an affective lexicon with lemmas paired to their polarity scores, we match the tokens in the tweet to their respective entries in the lexicon, and compute the sum of their values. We use *Sentix* (Basile and Nissim, 2013), an affective lexicon for Italian, created by the alignment of SentiWordNet (Baccianella et al., 2010) and the Italian section of MultiWordNet (Pianta et al., 2002). In particular, we adopt Sentix version 2.0[5].

3.2 Lemmatization

In order to match the tweets' words with a Sentix entry, we need to transform them into their base forms, i.e., lemmatize the tweets. For this purpose *UDPipe* R package with the function *udpipe_annotate* was used, applying all the three available models for Italian language: *ISDT* (Italian-isdt-ud-2.3-181115), *POSTWITA* (Italian-postwita-ud-2.3-181115), and *PARTUT* (Italian-partut-ud-2.3-181115). UDPipe (Straka and Straková, 2017) is an end-to-end NLP pipeline including part-of-speech tagging and syntactic parsing with Universal Dependencies.

We ran the models on *AGRITREND*. In order to automatically estimate the quality of the lemmatization, the produced lemmas were checked against the Hoepli dictionary, a large, general-purpose online Italian dictionary comprising over 500,000 lemmas[6]. The results, in Table 2, show how the UDpipe models generated a substantial amount of improper Italian lemmas. Moreover, for each of the three models, a number between 20% and 30% of incorrect lemmas were generated correctly by at least one of the two other models.
In Table 1 an example is shown of the lemmatization according to the three models: among other errors, the named entity *Adige* was incorrectly lemmatized by all models.

3.3 Polarity detection

We compute the polarity of the lemmatized tweets, including wrong lemmatizations, by matching the produced lemmas in Sentix. Incorrect lemmatization, even for a single word, may cause serious distortions of the polarized scores. For instance, comparing the overall polarity scores calculated for the three models in Table 1, we can see that when PARTUT has been used, a wrong lemma (which is a non-existing verbal form of the noun *acqua* (water)) has been associated to the word *acqua* determining the attribution of negative rather than positive score. This phenomenon often occurs in *AGRITREND* regardless of the lemmatiza-

[3] http://www.evalita.it/2016

[4] Already in 2016, only one team out of 13 participated to the SENTIPOLC shared task on Italian SA with an unsupervised system.

[5] https://github.com/valeriobasile/sentixR

[6] https://dizionari.repubblica.it/italiano.html

Table 1: A tweet from *AGRITREND* with the output of the three UDpipe lemmatization models where the lemmas are alphabetically ordered and the errors marked in bold.

Original	`@ANBI_Nazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgriculturaIT`
ISDT	`acqua` **`adigire`** `allarme crisi` **`d`** `dopo idrico po` - Sentix score: 0.080
POSTWITA	`acqua` **`adigere`** `allarme crisi di dopo idrico po` - Sentix score: 0.080
PARTUT	**`acquare adigere`** `allarme crisi` **`d`** `dopo idrico po` - Sentix score: -0.078

Table 2: Number and rate of lemmas produced by the UDpipe lemmatization models and not found in the Hoepli dictionary.

Model	Incorrect lemmas	%
ISDT	19,707	44.5
POSTWITA	21,444	48.4
PARTUT	22,440	50.7

tion model applied. Table 3 shows the percentages of negative, neutral and positive tweets based on the assigned polarity for each model. Here we consider positive a tweet whose Sentix score is greater than zero, negative when lower than zero, and neutral if it is exactly zero.

Table 3: Polarity classification on *AGRITREND* lemmatized with different UDpipe models.

Model	Negative	Neutral	Positive
ISDT	32.6%	9.5%	57.9%
POSTWITA	32.3%	10.2%	57.5%
PARTUT	33.8%	11.1%	55.1%

At the fist glance, from percentages only, we might argue that the lemmatization models, each one with its own bias, classified the tweets in a similar manner. However, at this step of analysis, we cannot say anything about statistical differences in the size and in the signs of the polarity scores between each model.

3.4 Statistical significance

If the differences between the scores were not statistically significant, the incorrect lemmatization should not impact on the polarity scores. Conversely, if significant differences exist, the lemmatization models will generate different polarity scores, severely affected by the incorrect lemmatization. In order to verify this hypothesis, we applied the non-parametric statistical signed rank test of Wilcoxon (1945) for paired samples to the polarity scores for each pair of models. This test is commonly used to verify if the difference between two scores from the same respondents (i.e., samples) is significantly different without the need for the data to follow a known probability distribution or high precision in the measures to be tested for. In our case the samples are coupled, since they are composed of the same tweets with potential different lemmas and the scores are the polarity of the tweets after lemmatization. As a consequence, the test is able to simply evaluate if the difference between the polarity of the tweets is due to the sign and the magnitude of the score simultaneously. The results of the Wilcoxon test, computed with the statistical package SPSS, are presented in Table 4.

The results of the Wilcoxon test are not statistically significant between ISDT and POSTWITA. The polarity obtained with the PARTUT lemmatization is significantly different from the other two, in line with the observation of a higher number of incorrect lemmas (51%, see Table 2). The result of this test indicates that an incorrect lemmatization produces *statistically significant* differences between the subsequent polarity scores and confirms our hypothesis.

4 Experiments on *SENTIPOLC*

In the previous section, we analyzed the lemmatization errors produced by three UDpipe models on *AGRITREND* and we observed how statistically significant is the failure in lemmatization on the result of dictionary-based SA. Nevertheless, being the *AGRITREND* corpus not annotated for sentiment polarity, we could not say anything about the *accuracy* of the prediction. To bridge this gap, we repeated the experiment on *SENTIPOLC*, where ground truth labels (also called *gold standard labels*) were manually annotated, starting by running the same processing pipeline as for *AGRITREND*. Table 5 shows an example tweet with the corresponding polarity scores. In this dataset, the percentages of incorrect lemmas, according to the Hoepli dictionary, is generally smaller than in the *AGRITREND* data, but still substantial: 35% for ISDT, 41% for POSTWITA, 44% for PARTUT (see Table 2 for a comparison with the other dataset).

Comparing the predictions obtained with Sentix with the labels annotated in *SENTIPOLC*, we eval-

Table 4: Wilcoxon signed rank test results between pairs of UDPipe models.

	ISDT vs. POSTWITA	ISDT vs. PARTUT	POSTWITA vs. PARTUT
Standardized test statistic	-1.317	-6.996	6.208
Asymtotic Sign. (2-sided test)	0.188 ($p > 0.05$)	0.000 ($p < 0.05$)	0.000($p < 0.05$)
Positive differences	2,190	2,250	2,913
Negative differences	2,281	2,824	2,404
Number of Ties	4,412	3,809	3,566

Table 5: Example tweet from *SENTIPOLC* with the output of three UDpipe lemmatization models. The lemmas are ordered alphabetically, since they are further processed as a bag of words.

Original text	`Capitale Europea della Cultura che combacia con la fine consultazioni de #labuonascuola: gran bel segnale :)`
Bag of words	`bel Capitale combacia consultazioni Cultura della Europea fine gran segnale`
ISDT	`bello capitale combaciare consultazione cultura di europeo fine grande segnale` - Sentix score: 0,8449
POSTWITA	`bello capitale combaciare consultazione cultura` **`da`** `europeo fine grande segnale` - Sentix score: 1,0739
PARTUT	**`bel`** `capitale` **`combacia`** `consultazione cultura` **`dere`** `europeo fine grande segnale` - Sentix score: -0,2715

Model	F1 (pos.)	F1 (neg.)	F1 (avg.)
ISDT	0.404	0.535	0.470
POSTWITA	0.414	0.540	0.477
PARTUT	0.409	0.540	0.474

Table 6: Performance of the dictionary-based SA, with different lemmatization models.

uate the performance of the dictionary-based approach in terms of precision, recall, F1-measure, and thus simultaneously measuring the impact of the different lemmatization models on the prediction accuracy. The results are shown in Table 6, in terms of F1-score for the positive polarity, negative polarity, and their average, following the official evaluation metrics of the SENTIPOLC task.
The Wilcoxon test applied on *SENTIPOLC* gave very similar results to those achieved on *AGRITREND*, confirming the similarity of the classification obtained with ISDT and POSTWITA, while PARTUT tends to stand apart. Moreover, errors in lemmatization have a statistically significant impact on the SA on the *SENTIPOLC* dataset to the same extent as *AGRITREND*.

5 Morphologically-inflected Affective Lexicon

The analyses presented in the previous sections highlight how low coverage and errors in lemmatization have a negative impact on the performance of downstream tasks such as SA. In an attempt to mitigate this issue, we propose an alternative approach to link the lexical items found in tweets with the entries of an affective lexicon such as Sentix without an explicit lemmatization step.

We expand the lexicon by considering all the acceptable forms of its lemmas. Each form takes the same polarity score of the original lemma. When different lemmas can assume the same form, we assign it the arithmetic mean of the lemmas' polarity scores. We use the *morph-it* morphological resource for Italian (Zanchetta and Baroni, 2005) to extract all possible forms from the lemmas of Sentix 2.0, and create a Morphologically-inflected Affective Lexicon (MAL) of Italian. The MAL comprises 148,867 forms, more than three times the size of Sentix 2.0 (41,800 lemmas).

The classification performance obtained using the MAL instead of a lemmatization model is in line with the results of the experiment in Table 6: 0.408 F1 (positive), 0.542 F1 (negative), and 0.475 F1 (average). However, so far we have employed a heuristic to map the Sentix score to polarity classes which is highly polarizing, that is, only tweets with an exact score of zero are classified as neutral. We therefore investigated a more conservative approach, where a parametric threshold T is introduced. After computing the polarity score of a message by summing up the polarity of its constituent words (or lemmas), we assign it a positive polarity label if the score is greater than T and negative if the score is lower than -T. The results of this experiment are shown in Figure 1. Several observations can be drawn from these results. First, using a threshold to assign polarity classes is indeed beneficial, with the right threshold empirically estimated around 5. Second, using the MAL instead of a lemmatization step improves the SA performance overall, in particular due to a better prediction of the negative polarity. Finally, the variation in threshold has opposite impact on the prediction of negative and positive tweets. We speculate that this may be due to asymmetries in

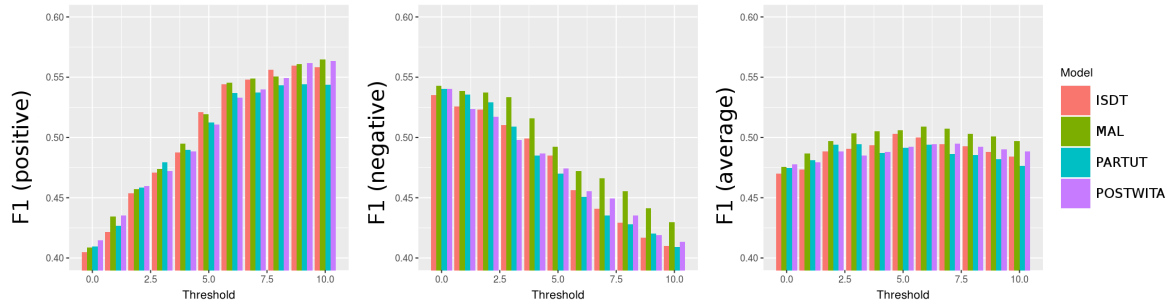

Figure 1: F1-score for the positive polarity (right), negative polarity (center) and average F1 (left) of the prediction of the dictionary-based SA approach on the *SENTIPOLC* test set.

the data, in the lexicon, or both, and intend to carry out future studies to understand this result.

6 Discussion

Our empirical study highlights important issues arising from language analysis errors (in lemmatization, in particular) propagating down the pipeline of a simple dictionary-based SA model. Without double-checking the outcome of the lemmatization step against a dictionary, a significant amount of noise is introduced in the system, leading to unstable results. The problem is even more substantial when dealing with data in a specific domain, such as the *AGRITREND* dataset of tweets about the agricultural domain, which indeed raised our attention on this problem.

We confronted the POS distribution of the parsed Agritrend and SENTIPOLC corpora with the set of UD-parsed corpora in Italian. In the Twitter data, content words are slightly more prominent, while function words are less present, although the general POS distributions have similar shapes. We report however an inverse correlation between the correctness of the lemmatization and the frequency of the POS, that is, words with infrequent POS are more likely to be wrongly lemmatized.

We tested the performance in a setting with no lemmatization at all, and measured a relatively good performance on the *SENTIPOLC* benchmark with some of the parameter configurations. This is unsurprising, following our observations on the significant impact of incorrect lemmatization on the SA performance. However, such a setting is linguistically questionable (matching only an arbitrary subset of words in a lemma-based resources) and its results are highly variable.

It is also important to notice that an incorrect lemmatization is likely hurtful not only to SA. The high reported number of non-existent lemmas created by the UDpipe models may severely alter the results of large-scale statistical studies on social media data, such as the ones planned by the creators of the *AGRITREND* data. Moreover, evaluating the correctness of a word by checking an external dictionary (in our case, Hoepli), is sensible to potential drawbacks of that resource, e.g., leading to overestimating lemmatization errors.

In sum, when choosing a pre-processing strategy for dictionary-based SA, the need arises to strike a balance between two extremes: 1) potentially incorrect lemmatization provided by a statistical model, that possibly *underestimates* the polarity; 2) an inclusive approach like MAL, that possibly *overestimates* the polarity.

7 Conclusion and Future Work

In this paper, we presented an empirical and statistical study on the impact of lemmatization on a NLP pipeline for SA based on an affective lexicon. We found that lemmatization tools need to be used carefully, in order to not introduce too much noise, deteriorating the performance downstream. Then we propose an alternative approach that skips the lemmatization step in favor of a morphologically rich affectve resource, in order to alleviate some of the observed issues.[7] We plan on integrating the proposed solutions, including the MAL and an automatic check of the lemma produced by UDpipe, in a pre-processing pipeline based on UDpipe.

[7]The MAL is available for download at `https://github.com/valeriobasile/sentixR/blob/master/sentix/inst/extdata/MAL.tsv`

Acknowledgments

The work of Marco Vassallo and Giuliano Gabrieli is funded by the Statistical Office of CREA. The work of Valerio Basile and Cristina Bosco is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Languages Resources Association (ELRA).

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, December.

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.

Mafalda Monda, Giuliano Gabrieli, and Marco Vassallo. 2019. Sentiment in agricoltura: Il termometro dell'agricoltura - i principali temi discussi su Twitter e gli umori degli addetti. In *I numeri dell'Agricoltura Italiana.* CREA, Centro Politiche e Bio-economia, June.

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. 2018. 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. European Languages Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).