Samskara
Minimal structural features for detecting subjectivity and polarity in Italian tweets

Irene Russo, Monica Monachini
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC CNR)
Lari Lab
firstname.secondname@ilc.cnr.it

Abstract

English. Sentiment analysis classification tasks strongly depend on the properties of the medium that is used to communicate opinionated content. There are some limitations in Twitter that force the user to exploit structural properties of this social network with features that have pragmatic and communicative functions. Samskara is a system that uses minimal structural features to classify Italian tweets as instantiations of a textual genre, obtaining good results for subjectivity classification, while polarity classification needs substantial improvements.

Italiano. *I compiti di classificazione a livello di sentiment analysis dipendono fortemente dalle proprietà del mezzo usato per comunicare contenuti d'opinione. Vi sono limiti oggettivi in Twitter che forzano l'utente a sfruttare le proprietà strutturali del mezzo assegnando ad alcuni elementi funzioni pragmatiche e comunicative. Samskara è un sistema che si propone di classificare i tweets italiani come se appartenessero a un genere testuale, interprentandoli come elementi caratterizzati da strutture minimali e ottenendo buoni risultati nella classificazione della soggettività mentre la classificazione della polarità ha bisogno di sostanziali miglioramenti.*

1 Introduction

After 15 years of NLP works on the topic Sentiment Analysis is still a relevant task, mainly because we assist every day to an exponential growth of opinionated content on the web that require computational systems to be managed. Detected, extracted and classified, opinionated content can also be labeled as positive or negative, but additional categories (ambiguous, neutral etc.) are possible. Resources and methodologies created for the detection and classification of subjectivity and polarity in reviews are not applicable with good results on different data, such as tweets or comments about news from online fora.

There are several reasons behind this: first and foremost, opinions can be expressed more or less explicitly depending on the context; lexical cues from lexical resources such as SentiWordNet (Baccianella et al., 2010) or General Inquirer (Stone, 1966) could be useless when people write their point of views in complex and subtle ways. Secondly, different media and platforms impose different constraints on the structure of the content expressed.

Twitter's limits in terms of characters force the use of abbreviations and the omission of syntactic elements. But users try to exploit creatively these limitations, for example adding pragmatic functions with emoticons.

Features and functionalities anchoring the text to extra-linguistic dimensions (such as mentions and pictures in tweets or *like/agree* from other users in online debates) should be considered in Sentiment Analysis classification tasks because of to their communicative functions.

In this paper we present Samskara, a Lari lab system for the classification of Italian tweets that took part in two tasks at Sentipolc2016 (Task 1,subjectivity and Task 2, polarity classification). The system is described in par. 2, with results presented in 2.2 where we discuss the limitations of the system.

2 System description

Samskara is a classification system based on a minimal set of features that wants to address the issue of subjectivity and polarity classifications of Italian tweets. Tweets are considered as instantiations of a textual genre, namely they have specific structural properties with communicative and pragmatic functions. In our approach, focusing on the structural properties means:

- abstracting the task from lexical values of single words that could be a deceptive cue because of lexical sparseness, ambiguity of words, use of jargon and ironic exploitations of words;
- taking into account features used in authorship attribution to represent abstract patterns characterizing different styles, e.g. PoS tag n-gram frequencies(Stamatos, 2009)[1];
- choosing a tagset for PoS that includes tags peculiar of tweets as a textual genre, i.e. interjection and emoticon.

More generally, we want to capture high-level linguistic and extra-linguistic properties of tweets, also considering basic sequential structures in forms of sequences of bigrams.

2.1 Data analysis, data preprocessing and feature selection

Before starting with the selections of features, data analysis of the training set helped in the investigation of several hypotheses.

Polarised lexical items have been widely used in sentiment analysis classification (Liu and Zhang, 2012) but resources in this field list values at sense level (such as SentiWordNet) or conflate the senses in a single entry (such as General Inquirer and LIWC). Without an efficient word sense disambiguation module, using SentiWordNet is difficult. One strategy is to sum all the values and to select a threshold for words that are tagged as polarised in text. That means to overstimate positive/negative content, without finding a clear boundary between, for example, positive and negative tweets.

Considering the Italian version of LIWC2015 (Pennebaker et al., 2015) we see that frequencies are unable to distinguish between positive and negative tweets in the Sentipolc2016 training data (see Table 1). To avoid this, we defined for internal use a subset of SentiWordNet 3.0 (Baccianella et al., 2010) that we call SWN Core selecting:

class	tokens	LIWC+	LIWC-
pos	92295	234 (0.26%)	225 (0.25%)
neg	114435	78 (0.07%)	683 (0.6%)

Table 1: Absolute and relative frequencies of Italian LIWC2015 lemmas in positive and negative tweets (Sentipolc2016 training set).

- all the words corresponding to senses that are polarised;
- from the set above, all the words corresponding to senses that display single-valued polarity (i.e. they are always positive or always negative);
- from the set above we delete all the words that have also a neutral sense;
- we sum polarity values for every lemma in order to have for example a single value for lemmas listed in SWN with two different positive values or three different negative values.

The English SWN Core is composed by 6640 exclusively positive lemmas and 7603 exclusively negative lemmas. Since in these lists items have a polarity value ranging from 0.125 to 3.25, with the idea of selecting lemmas that are strongly polarised we set 0.5 as threshold; as a consequence of this decision we have 1844 very positive and 3272 very negative lemmas. After deletion of multiword expressions these strongly opinionated words have been translated to Italian using Google Translate, manually checked and annotated with PoS and polarity.

We clean the lists, deleting lemmas that appear two times, lemmas that have been translated as multiword expressions and lemmas that do not have polarity in Italian. At the end we have 890 positive and 1224 negative Italian lemmas. Considering their frequencies in the training set (see Table 2) we find out that only negative items are distinctive. Because of the presence of ironic tweets positive lemmas tend to occur in tweets that

[1] For the moment we think that sequences of syntactic relations are not useful because of the poor performance of Italian syntactic parsers on tweets.

have been tagged as negative. The exploitation of positive words in ironic communication is a well-known phenomenon (Dews and Winner, 1995) - the positive literal meaning is subverted by the negative intended meaning - and neglecting this aspect of the Sentipolc2016 training set could imply lower classification performances. If we allow positive items from SWN Core in the system the classification of negative tweets is made difficult. As we mention above, structural properties

	SWN Core+	**SWN Core-**
obj	536 (0.76%)	264 (0.37%)
subj	2307 (1.4%)	1608 (1%)
pos	1055 (4.8%)	200 (0.9%)
neg	839 (2%)	1096 (2.6%)

Table 2: Absolute and relative frequencies of SWN Core lemmas in Sentipolc2016 training set.

of tweets can be treated as sequences of PoS. To reduce data sparseness and to include dedicated tags for Twitter we choose the tagset proposed by PoSTWITA, an Evalita2016 task (Bosco et al., 2016). It looks promising because it contains categories that:

- could be easily tagged as preprocessing step with regular expressions (for example MENTION and LINK);
- are suitable for noisy data, tagging uniformly items that can be written in several, non-predictable ways (*ahahahha, haha* as INTJ);
- contains tags that have communicative and pragmatic functions, such as emoticon and interjection (see Table 4).

We preprocessed all the tweets in the training set substituting elements that are easy to find, such as mention, hashtags, email, link, emoticon (all tags included in PoSTWITA).
After that, Sentipolc2016 training set has been tagged with TreeTagger (Schmid, 1997); TreeTagger tags have been converted to PostTWITA tagset (see Table 3) and additional tags from PosTWITA have been added, building dedicated lists for them that include items from PoSTWITA training set plus additional items selected by the authors (see Table 4).
Thanks to TreeTagger we have all the words lemmatized and so all the lemmas included in the negative counterpart of SWN Core can be substituted

TreeTagger	**PoSTWITA**
AUX	[A-Z a-z]+ AUX
DET	[A-Z a-z]+ DET
PRO	[A-Z a-z]+ PRON
NPR	[A-Z a-z]+ PROPN
PUN	PUNCT
SENT	PUNCT
VER[A-Z a-z]+cli	VERB_CLIT
VER	[A-Z a-z]+ VERB

Table 3: Comparison between TreeTagger and PoSTWITA tagsets.

by the tag VERYNEG. At this point, with the intention to have a minimal sequence of significant tags, we created 4 version of the training set according to 4 minimal structures, deleting all lemmas and leaving only PoS tags:

- minimal structure 1 (MSTRU1): EMO, MENTION, HASHTAG, URL, EMAIL;
- minimal structure 2 (MSTRU2): EMO, MENTION, HASHTAG, URL, EMAIL, PROPN, INTJ;
- minimal structure 3 (MSTRU3): EMO, MENTION, HASHTAG, URL, EMAIL, PROPN, INTJ, ADJ, ADV;
- minimal structure 4 (MSTRU4): EMOTICON, MENTION, HASHTAG, URL, EMAIL, PROPN, INTJ, VERYNEG.

We performed classification experiments with these features and we get better results with MSTRU4 (see par. 2.2).
For Samskara each tweet is represented as a sequence including its EMO, MENTION, HASHTAG, URL, EMAIL, PROPN (Proper Noun), INTJ and VERYNEG lemmas from SWN Core (see tweet in example 1 represented in example 2). This minimal, very compact way to represent a tweet is very convenient because partially avoids any noise introduced by PoS tagger (containing only VERYNEG and PROPN as elements that should be properly tagged with this tool).

(1) @FGoria Mario Monti Premier! #Italiaresiste.

(2) MENTION PROPN HASHTAG.

Additional features for the classification of subjective and positive or negative tweets are listed in

new tag	type	examples
PART	particle	*'s*
EMO	emoticon	*:DD, :-)))), u___u*
INTJ	interjection	*ah, boh, oddioo*
SYM	symbol	*%, &, <*
CONJ	coordinating conjunction	*ebbene, ma, oppure*
SCONJ	subordinating conjunction	*nonostante, mentre, come*

Table 4: Examples of lemmas tagged according to Twitter-specific PoSTWITA tags.

Table 5, with BOOL meaning boolean feature and NUM numeric feature (they correspond to absolute frequencies). The features have been selected thinking about their communicative function: *a1* for example is useful because there is a tendency to communicate opinionated content in discussions with other users while we choose *a2* because neutral tweets often advertise newspapers' articles in a non opinionated way including the link at the end of the tweet, but the URL is significant in other positions *a6, a6_1*. Together with emoticons, interjections are items that signal the presence of opinionated content. For the kind of asynchronous communication that characterize them, tweets can contain questions that don't expect an answer, that are rethorical *a8_1*, thus making the tweet opinionanted.

2.2 Results and Discussion

The system adopts the Weka[2] library that allows experiments with different classifiers. Due to better performance of Naive Bayes (default settings, 10- fold cross validation) with respect to Support Vector Machine we choose the first; best performances were obtained with MSTRU4 considering frequencies of unigrams and bigrams of PoS as features. We took part to Sentipolc2016 only with a constrained run, choosing slightly different set of features for subjectivity and polarity evaluation.

Adding the additional features in Table 5 we selected for Task 1 a subset of them after an ablation test. More specifically, the feature set 1 (FS1 in Table 7) is composed by features *a1, a2, a4, a4_1, a6, a6_1, a7, a7_1, a8_1, a9*. The system performance is reported in terms of F-score, according to the measure adopted by the task organizers (Barbieri et al., 2016). Results on the training data look promising for Task 1, less promising for Task 2 (see Table 8). We didn't succeed in optimising features for the polarity detection sub-task. The performance on the training set was not satisfying but nevertheless we decided to submit results for Task 2 on test set using all the features. In Table 9 the official results submitted for the competition are reported. Samskara was first among the constrained systems for subjectivity classification, while not surprisingly the performance in Task 2 was bad. Results in Task 2 can be explained by the absence in the system of structural features that are meaningful for the positive-negative distinctions or by the unsuitability of such a minimal approach for the task. It is possible that richer semantic features are necessary for the detection and the classification of polarity and polarised lexical items should be revised, for example, representing each lemma as a sentiment specific word embedding (SSWE) encoding sentiment information (Tang et al., 2014).

With Samskara we prove that classification of tweets should take into account structural properties of content on social media, especially properties that have communicative and pragmatic functions. The minimal features we selected for Samskara were successful for the classification of subjective Italian tweets. The system is based on a minimal set of features that are easy to retrieve and tag; the classification system is efficient and fast for Task 1 and as such it is promising for real-time processing of big data stream.

[2]http://www.cs.waikato.ac.nz/ml/weka/

References

Stefano Baccianella and Andrea Esuli and Fabrizio Sebastian. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Barbieri, Francesco and Basile, Valerio and Croce, Danilo and Nissim, Malvina and Novielli, Nicole and Patti, Viviana. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Pierpaolo Basile, Anna Corazza, Franco

features	description	type
a1	the tweet starts with MENTION	BOOL
a2	the tweet ends with a LINK	BOOL
a3	the tweet has PoS of type PUNCT	BOOL
a3_1	number of PoS of type PUNCT in each tweet	NUM
a4	the tweet has PoS of type VERYNEG	BOOL
a4_1	number of PoS of type VERYNEG in each tweet	NUM
a5	the tweet has PoS of type INTJ	BOOL
a5_1	number of PoS of type INTJ in each tweet	NUM
a6	the tweet has PoS of type URL	BOOL
a6_1	number of PoS of type URL in each tweet	NUM
a7	the tweet has PoS of type EMOTICON	BOOL
a7_1	number of PoS of type EMOTICON in each tweet	NUM
a8_1	the tweet contains a question	BOOL
a8_2	the tweet contains a question at the end	BOOL
a9	the tweet contains two consecutive exclamation marks ('!!')	BOOL
a10	the tweets contains connectives such as *anzitutto, comunque, dapprima, del resto*	BOOL

Table 5: Additional features for subjectivy and polarity classification of tweets.

	MSTRU4 + FS1
obj F-score	0.532
subj F-score	0.811
Avg F-score	0.724

Table 6: Classification results for Task 1 obtained on Sentipolc2016 training set.

	MSTRU4 + AllF
pos F-score	0.424
neg F-score	0.539
both F-score	0.047
neu F-score	0.526
Avg F-score	0.48

Table 7: Classification results for Task 2 obtained on Sentipolc2016 training set.

	F-score	**Rank**
Task 1	0.7184	1
Task 2	0.5683	13

Table 8: Classification results for Task 1 and Task 2 on Sentipolc2016 test set.

Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *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)* Associazione Italiana di Linguistica Computazionale (AILC).

Bosco, Cristina and Tamburini, Fabio and Bolioli, Andrea and Mazzei, Alessandro. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *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)* Associazione Italiana di Linguistica Computazionale (AILC).

Shelly Dews and Ellen Winner. 1995. Muting the meaning: A social function of irony. *Metaphor and Symbolic Activity*, 10(1):319.

Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.) *Mining Text Data*, pp. 415–463, US: Springer.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*.

Helmut Schmid. 1997. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *New Methods in Language Processing*, UCL Press, pp. 154-164.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*.

Stone, Philip J. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.