A Quick Survey on Sentiment Analysis Techniques: a lexical based perspective

Flora Amato[1], Francesco Colace[2], Luca Greco[2] , Vincenzo Moscato[1], Antonio Picariello[1]
[1]DIETI- Department of Electrical Engineering and Information Technology
Università degli Studi di Napoli "Federico II", Napoli - Italy
{flora.amato, vmoscato, picus}@unina.it
[2]DIEM - Department of Information Engineering, Electrical Engineering and Applied Mathematics
Università degli Studi di Salerno, Fisciano (Salerno) - Italy
{fcolace,lgreco}@unisa.it

***Abstract*— With the advent of *World Wide Web* and the widespread of on-line collaborative tools, there is a increasing interest towards automatic tools for *Sentiment Analysis* to provide a quantitative measure of "positivity" or "negativity" about opinions or social comments. In this paper, we provide an overview of the most diffused techniques for sentiment analysis based on the lexical-based approaches as a quick reference guide in the choice of the most suitable methods for solving a specific problem in the sentiment analysis field.**

***Index Terms*— Sentiment analysis, Computational linguistics, Text Classification .**

I. INTRODUCTION

People's opinion has always driven human choices and behaviors, even before the diffusion of Information and Communication Technologies. With the advent of *World Wide Web* and the widespread of on-line collaborative tools such as blogs, focus groups, review web sites, forums, social networks (e.g *Facebook*, *Twitter*, *MySpace*, etc.), users more and more use to make available to everyone their tastes and liking, and in general, their opinions and sentiments about an event, a topic, a public person, a political faction, a TV program, etc.

In such a context, there is an increasing need to have available automatic tools for *Sentiment Analysis* (or *Opinion Mining*) and *Tracking* in order to provide a quantitative measure of "positivity" or "negativity about opinions (*polarity*) or comments related to a certain topic of interest and to track along the time such information.

More in details, sentiment analysis aims at finding the opinions of authors (thought leaders and ordinary people) about specific entities, by analyzing a large number of documents (in any format such as PDF, HTML, XML, etc.).

It can be considered as a sub-discipline of Computational Linguistics, indeed it is a *Natural Language Processing* and *Information Extraction task* [14], or challenged by the use of classical *Machine Learning* based approaches.

DOI reference number: 10.18293/VLSS2015-020

The most studied languages in the opinion mining field are English and Chinese, but there are several researches on other languages like Italian, Thai and Arabic [12].

Opinion mining allows to identify problems by listening, rather than by asking, ensuring an accurate reflection of reality [14].

The analyzed textual information can be of two types: *facts* and *opinions*. The facts are objective expressions that describe entities, conversely the opinions deal with people's emotions, sentiments and feelings and so they are subjective.

Generally, we can see an opinion (or a sentiment) as a quintuple: $<o, f, s, h, t>$, where o is the object evaluated by the opinion holder h, f is a feature of the object o, t is the time when the opinion has been expressed and s is the value of the opinion (for example positive or negative) [1][14].

Sentiment analysis techniques have as main goal the automatic extraction of the polarity measure "attached" to an object and can adopt several methods and techniques derived both from Computational Linguistics and Machine Learning theory. Here, we focus our attention on *lexical-based* techniques belonging to the branch of Computational Linguistics approaches.

The paper is organized as follows. Section II contains a review of the most diffused lexical-based approaches. Finally, Section III reports some conclusions and final considerations about our study.

II. AN OVERVIEW OF LEXICAL BASED SENTIMENT ANALYSIS TECHNIQUES

In lexical-based approach a predefined list of words is used to determine a specific sentiment. A relevant problem regards ambiguity of natural language: sentiment value for a given word depends on the specific context.

There are several approaches to sentiment lexicons' creation. A manual construction is often difficult and very time consuming. In the literature, the most used methods can be classified as Corpus-based and Dictionary-based.

i. *Corpus-based Approach*

In this approach, a set of seed words grows by using a corpus of documents of a specific domain. Therefore a specific domain lexicon is constructed on the basis of a labeled corpus of documents.

One of the first works in this field is [6] where, given some seed adjectives, a corpus is used to identify additional sentiment adjectives. A key point regards the presence of conjunctions: for example the conjunction 'and' between two adjectives can refer to the same sentimental polarity. A graph with same or different orientation links between adjectives is created. These adjectives are then separate with a clustering algorithm into two subsets.

Another example is [8] where a corpus of 10000 blog posts from LiveJournal.com is used; the posts are labeled "happy" or "sad". A happiness factor is assigned to words by calculating their frequency: the ratio between the number of occurrences of a word in the happy blogposts and its frequency in the entire corpus.

Among the most recent studies there is the work in [4]. The key of this approach is searching the connotative polarity between a conative predicate and its semantic argument. It is done by using a graph-based algorithm that use PageRank [9] and HITS [7] that collectively learn connotation lexicon together with connotative predicates.

ii. *Dictionary-Based Approach*

In this approach a small set of seed words is first manually collected and then is expanded with words synonyms and antonyms. This is done by using online resources (dictionaries). The most well-known example is *Wordnet* that is an online lexical database for English language.

A great disadvantage of this approach is that the lexicon acquired is independent from a specific domain.

- *WordNet-Affect*

WordNet-Affect [11] is a linguistic resource, composed by 2,874 synsets and 4,787 words, developed considering WordNet Domains, that is a multilingual extension of Wordnet.

It aims at providing correlations between affective concepts and affective words by using a synset model.

A subset of synsets, which are able to represent affective concepts, is derived from WordNet. Then, these synsets are labeled with one or more affective categories.

The Core of WordNet Affect is created by considering a lexical database, called Affect, composed by 1,903 words that are mostly adjectives and nouns.

Lexical and affective information are associated to each term; they includes parts of speech, definitions, synonyms and antonyms.

In order to assign an affective category to terms, an attribute called Ortony is used. Terms can be classified in emotional terms, non-emotional affective terms, non-affective mental state terms, personality traits, behaviors, attitudes etc.

Ortony information is projected on the subset selected from Wordnet but doesn't cover all Affect items and for this reason some labels are manually assigned. When the subset is completely labeled, WordNet-Affect Core is defined and can be extended exploiting WordNet relations.

- *SentiWordNet*

SentiWordNet is a lexical resource proposed in [2].

SentiWordNet is built with a ternary classification, indeed each synset (set of synonyms) is labeled as positive, negative or objective by using a set of ternary classifiers. If all of them will give to the synset the same label, therefore that label for that synset will have the maximum score; otherwise this score will be proportional.

Each classifier follows a semi-supervised approach that is a learning process where the training set $Tr = L \cup U$ so that: L is a small subset of manually labeled training data, and U is a subset of training data labeled by the process by using L, and other available resource, as input.

In [2] L is divided into: L_P, L_n, that are two small synsets respectively for positive and negative training data, and L_o for the objective ones.

L_p and L_n are expanded with K iterations obtaining the following result for the i-th iteration:

Tr_p^i (resp Tr_n^i) will contain, in adding to Tr_p^{i-1} (resp Tr_n^{i-1}), all the synsets that are related to synsets in Tr_p^{i-1} (resp Tr_n^{i-1}) by WordNet lexical relations and have the same Positive(resp. Negative) polarity, and the synsets that are related to synsets in Tr_n^{i-1} (resp Tr_p^{i-1}) and have the opposite polarity.

Tr_o^K coincides with L_o and it consists of 17,530 synsets that doesn't belong either to Tr_p^K or to Tr_n^K. To each synset is associated a vectorial representation by applying a cosine-normalized tf*idf to its gloss, that is a textual representation of its semantic.

Hence now the training synset, for a class c_i, can be given to a standard supervised learner that generates two binary classifiers. One of these will distinguish *positive* and *not_positive* terms, and takes $Tr_p^K \cup Tr_o^K$ in the training phase, the other one will classify terms as *negative* or *not_negative*, and takes $Tr_n^K \cup Tr_o^K$ in the training phase.

It produces a resulting ternary classifier that will classify the entire WordNet.

SentiWordNet has been developed in several versions, but the most significant is SentiWordNet 3.0 that, in the automatic annotation of WordNet, adds to the semi-supervised learning step a random-walk step for refining the scores. This version is compared with the previous one, and an improvement in accuracy of about 20% is found.

- *Context Dependent Opinion Observer (CDOO)*

CDOO is a system implemented in C++ and it is based on a method that tries to infer the semantic orientation of opinion sentences by associating contextual information to opinion words obtained from WordNet.

This approach goes through four steps.

In the first step, after a preprocessing phase, opinion sentences are extracted from the inputs by using feature keywords directly.

In the second step Context independent opinions that don't

require any contextual information are analyzed to determine the semantic orientation. In this step opinion words from Wordnet are simply considered and in particular are utilized adjective synonym set and antonym set.

In the third step distinct-dependent opinions are analyzed: adjacent sentences are needed to define the semantic orientation by using Linguistic rules, especially conjunction rule.

In the fourth and final step Context indistinct opinions that need contextual information from other reviews are analyzed. In order to collect contextual segments sets for given features, a large number of online reviews are considered. Subsequently, contextual information is extracted from the segment sets by using Emotional-ATFxPDF to compute weight of terms in text segment set. Then the orientation of the opinion is calculated using semantic similarity.

- *SenticNet*

SenticNet is inspired by SentiWordNet but it assigns to each concept *c* only one value p_c belonging to [-1,1].

The polarity of a concept *c* is defined in the following way:

$$p_c = \frac{Plsn(c) + |Attn(c)| - |Snst(c)| + Aptt(c)}{9}$$

where *Plsn* is *Pleasantness*, *Attn* is *Attention*, *Snst* is *Sensitivity*, *Aptt* is *Aptitude*.

They start from Hourglass model and for example, in order to find positive concepts correlated with Pleasantness, they begin to search concepts semantically correlated to words like "joy", "serenity" and uncorrelated to words like "sadness".

Two different techniques are used: Blending and Spectral Assumption. When polarity is assigned, SenticNet is encoded in RDF triples using a XML syntax.

The current version of SenticNet contains almost 15,000 concepts.

In recent studies SenticNet is often associated to WordNet-Affect. For example in [10] researchers assign to SenticNet concepts, which are not present in WordNet-Affect, emotion labels. It is actually an expansion of WordNet-Affect based on SenticNet. By analyzing several features and utilizing a SVM framework for classification, they obtain an accuracy of 85.12% in their best result.

- *Panas-t*

The original PANAS is created by Watson and Clark and they analyzed 10 moods on a 5-point scale [13].

They also expanded it in PANAS-x where eleven specific affects are considered: Fear, Sadness, Guilt, Hostility, Shyness, Fatigue, Surprise, Joviality, Self-Assurance, Attentiveness, and Serenity. To each affect a list of adjectives is associated.

In [5] it is expanded in Panas-t which is an adaptation that analyzes short text from Online Social Media and in particular from Twitter.

They consider a dataset composed by tweets from all the public accounts registered before August 2009. First tweets that explicitly contain feelings (and hence tweets that contain words like "I am", "feelings", "myself") are identified.

Then a preprocessing phase is performed where individual terms are isolated, using white-space boundaries, and punctuation and other non-alphanumeric characters are removed.

It is assumed that a tweet can be mapped to the first sentiment *s* that appears in the tweet. This can be done by verifying the position of the adjectives.

III. CONCLUSIONS

The paper provided an overview of the most diffused techniques for sentiment analysis based on the lexical-based approaches and the related systems.

The paper wants to be a quick reference guide in the choice of the most suitable lexical-based approaches for a specific problem of sentiment analysis.

REFERENCES

[1] F. Colace, L. Casaburi, M. De Santo, L. Greco, "Sentiment detection in social networks and in collaborative learning environments", Computers in Human Behavior, Available online 27 December 2014, ISSN 0747-5632, http://dx.doi.org/10.1016/j.chb.2014.11.090.

[2] A.Esuli, and F.Sebastiani – "Sentiwordnet: A publicly available lexical resource for opinion mining", Proceedings of LREC. Vol. 6. 2006.

[3] R.Feldman – "Techniques and Applications for Sentiment Analysis", Magazine - Communication of the ACM (April,2013).

[4] S.Feng, B.Ritwik, and C.Yejin – "Learning general connotation of words using graph-based algorithms", Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[5] P.Gonçalves, F.Benevenuto, and M.Cha - "Panas-t: A pychometric scale for measuring sentiments on twitter", arXiv preprint arXiv:1308.1857 (2013). 14

[6] V.Hatzivassiloglou and K.R.McKeown – "Predicting the semantic orientation of adjectives", Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics, 1997.

[7] J.M.Kleinberg – "Authoritative sources in a hyperlinked environment", Journal of the ACM (JACM) 46.5 (1999): 604-632. 1999.

[8] R.Mihalcea, and H.Liu – "A Corpus-based Approach to Finding Happiness", AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[9] L.Page, S.Brin, R.Motwani and T.Winograd – "The PageRank citation ranking: Bringing order to the web" 1999.

[10] S.Poria, A.Gelbukh, E.Cambria, P.Yang, A.Hussain and T.Durrani - "Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis", Signal Processing (ICSP), 2012 IEEE 11th International Conference on , vol.2, no., pp.1251,1255, 21-25 Oct. 2012.

[11] C.Strapparava, and A.Valitutti - "WordNet Affect: an Affective Extension of WordNet", LREC. Vol. 4. 2004.

[12] G.Vinodhini, R.M.Chandrasekaran – "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2,Issue 6 (June 2012).

[13] D.Watson, L.A.Clark, and A.Tellegen - "Development and validation of brief measures of positive and negative affect: the PANAS scales", Journal of personality and social psychology 54.6 (1988).

[14] M.Shelke, S.Deshpande, V.Thakre – "Survey of Techniques for Opinion Mining", International Journal of Computer Applications (0975-8887) Volume 57-No.13 (November 2012).