DRANZIERA: An Evaluation Protocol For Multi-Domain Opinion Mining

Mauro Dragoni[1], **Andrea G. B. Tettamanzi**[2], **Célia da Costa Pereira**[2]
[1] FBK–IRST, Trento, Italy
[2] Université Nice Sophia Antipolis, I3S, UMR 7271, Sophia Antipolis, France
dragoni@fbk.eu, andrea.tettamanzi@unice.fr, celia.pereira@unice.fr

Abstract

Opinion Mining is a topic which attracted a lot of interest in the last years. By observing the literature, it is often hard to replicate system evaluation due to the unavailability of the data used for the evaluation or to the lack of details about the protocol used in the campaign. In this paper, we propose an evaluation protocol, called *DRANZIERA*, composed of a multi-domain dataset and guidelines allowing both to evaluate opinion mining systems in different contexts (Closed, Semi-Open, and Open) and to compare them to each other and to a number of baselines.

Keywords: Opinion Mining, Benchmark, Evaluation Protocol

1. Introduction

Opinion Mining is a natural language processing task whose aim is to classify documents according to the opinion they express on a given subject (Pang et al., 2002). Generally speaking, opinion mining aims at determining the attitude of a speaker or a writer with respect to a topic or the overall tonality of a document. This task has created a considerable interest due to its wide applications. In recent years, the exponential increase of the Web for exchanging public opinions about events, facts, products, etc., led to an extensive usage of opinion mining approaches, especially for marketing purposes.

By formalizing the opinion mining problem, an "opinion" (or "sentiment") has been defined by Liu and Zhang (2012) as a quintuple:

$$\langle o_j, f_{jk}, so_{ijkl}, h_i, t_l \rangle, \quad (1)$$

where o_j is a target object, f_{jk} is a feature of the object o_j, so_{ijkl} is the sentiment value held by person h_i (which will be called an "opinion holder") on feature f_{jk} of object o_j at time t_l. The value of so_{ijkl} can be positive (denoting a state of happiness, bliss, or satisfaction), negative (denoting a state of sorrow, dejection, or disappointment), or neutral (it is not possible to denote any particular sentiment), or a more granular rating. The term h_i encodes the opinion holder, and t_l is the time when the opinion is expressed.

With the growth of the number of solutions available in this field, the comparison between new approaches and the ones available in the state of the art is a mandatory task for measuring the quality of proposed solution. Unfortunately, in many cases it is not trivial to replicate evaluation campaigns mentioned into published papers. Among all possible issues, like the difficulty in replicating the implemented model, we want to focus on the "data" side. More in detail, the issues that can be easily found when we try to replicate experiments are the following.

- The used dataset is not longer available for download.
- The system is not longer maintained or available on the authors' website.
- In case of dataset availability, it is not clear how the dataset has been split for training and validating the model.
- Systems are evaluated without the adoption of a specific protocol; therefore, it is hard to compare one's own system with all the others due to the amount of effort required for replicating the same settings.

Due to the reasons presented above, it is often hard to ensure a fair comparison among all the systems. When we talk about "fair comparison", we mean the possibility of comparing different systems by using the same data and under the same environment (settings, constraints, etc.).

The solution for going in the direction of implementing a "fair comparison" policy, is to provide a methodology describing data, resources, and guidelines, that have to be followed for evaluating a system. To the best of our knowledge, in the field of opinion mining, such a methodology is still missing and, with the *DRANZIERA* protocol presented in this paper, we aim at filling this gap.

In the literature, some datasets have been proposed (Blitzer et al., 2007; Saif et al., 2013) and venues for comparing systems, like SemEval,[1] are organized. However, the two papers mentioned above do not provide details about how to split data for training and validating systems; this way, it is very hard to replicate the experiments performed by systems adopting such datasets. In SemEval, on the other hand, the possibility of building models without constraints on the resources that can be used does not allow to perform a complete comparison under the same environment. Moreover, the adoption of only one training and test sets does not avoid the introduction of overfitting aspects during system evaluation.

The paper is structured as follows. In Section 2., we present the dataset that is part of the *DRANZIERA* protocol; then, in Section 3., guidelines about how to use the dataset on the settings for performing different type of evaluation are provided. Finally, in Section 4., results obtained by two systems evaluated under the *DRANZIERA* protocol are reported.

[1] Last edition can be found at http://alt.qcri.org/semeval2016/

2. The Dataset

In this Section, we present the characteristics of the dataset contained within the *DRANZIERA* protocol.
The dataset is composed of one million reviews crawled from product pages on the Amazon web site. The size of one million has been chosen in order to (i) have an amount of data reflecting a good variety of the user-generated content that is analyzed, (ii) give a quantity of data suitable also for measuring the scalability of the approaches, and (iii) perform the experimental campaign in a reasonable time span. Such reviews belong to twenty different categories (that we will call "domains" for the rest of the paper) listed in Table 1.

Table 1: Categories/Domains

Amazon Instant Video	Automotive
Baby	Beauty
Books	Clothing Accessories
Electronics	Health
Home Kitchen	Movies TV
Music	Office Products
Patio	Pet Supplies
Shoes	Software
Sports Outdoors	Tools Home Improvement
Toys Games	Video Games

For each domain, we extracted five thousands positive and five thousands negative reviews that have been split in five folds containing one thousand positive and one thousand negative reviews each. This way, the dataset is balanced with respect to both the polarities of the reviews and the domain to which they belong. The choice between positive and negative documents has been inspired by the strategy used in (Blitzer et al., 2007) where reviews with 4 or 5 stars have been marked as positive, while the ones having 1 or 2 stars have been marked as negative. Furthermore, in each domain, the dataset is balanced with respect to both positive and negative polarity. The rationale behind the choice of having a balanced dataset is that our idea is not to provide a dataset reflecting the proportion detected in the sources from which documents have been retrieved. The proportion measured in such sources reflects only a part of the reality; therefore, an unbalanced dataset would be misleading during systems training. Instead, by providing a balanced dataset, systems are able to analyze the same number of positive and negative contexts for each domain. This way, the built models are supposed to be fairly used in any kind of opinion inference testing environment.
In the proposed dataset, we did not take into account neutral documents. The rationale for ignoring neutral documents is the difficulty of validating their real neutrality. For single sentences it is easy to annotate them as "neutral"; for instance, the sentence "The weather is hot" does not contain any polarized information. However, the annotation of documents as "neutral" is more controversial due to the complex structure of documents themselves. Let us consider the text below taken from the Amazon.com website, which, according to the Blitzer strategy, would be annotated as a "neutral" document.

Love this camera. That said, I bought it the first day you could get it. When it arrived all was fine. I was able to pair it with my iPad and an android, however, two days later the wifi connection failed. I tried everything I could from my experience with a Gopro Hero4 Black with no good news. It took about an hour on the phone for Gopro to admit we have a problem. They plan on replacing the camera, but can not tell me when. Waite until they get the bugs out. Hope that it will be the camera I think it can be.

As it is possible to notice, the document contains a mix of both "positive" and "negative" opinions. Such a mix is not always balanced, therefore, it is not possible to infer the real neutrality of a document by simply aggregating all the expressed polarities. Another reason is that when user-generated contents are taken into account for building opinion models, we are not working with annotations provided by experts, but with annotations that are the results of user impressions, moods, etc., and the boundary between neutrality and "small" polarization is too fuzzy. For these reasons, we decided to exclude "neutral" documents from the dataset in order to avoid the adoption of borderline documents that actually introduce noise in building the opinion models.
The split of each domain in five folds allows to easily have a clear distinction between the samples used for training the system and the ones used for testing it. Most of the work in the literature cites the dataset they used without specifying which part has been used for the training phase and which part has been used for the testing phase. Moreover, multiple folds can be exploited for performing a k-fold cross validation; hence, all samples may be used either for training and testing the system. Different settings for the value of k has been adopted in the literature, but in general k remains an unfixed parameter. Our assignment of $k = 5$ allows to have a good trade-off between the verification of the model generalization and the computational time needed for performing the experimental campaign.

3. Evaluation Guidelines

Comparison between effectiveness measured by different systems is possible only if common guidelines are followed in the usage of both data and external resources. The *DRANZIERA* protocol provides three evaluation settings that differ on the type of external resources that can be used for building the model used by a system. Such settings are presented below.

3.1. "Closed" setting

Within the "closed" setting, only samples contained in the folds chosen as training set can be used for training systems. This kind of setting is thought for approaches mainly based on pure Natural Language Processing analysis or statistical techniques. Evaluation has to be performed by adopting the cross-fold strategy using, iteratively, all folds for measuring the effectiveness of the system. The activities that have to be carried out are:

- Perform five training-test rounds by using four folds for each round for training and one fold for testing. Each of the folds has to be used as test set in one round.
- In each round, compute precision, recall, and f-measure;
- Compute the average of precision, recall, and f-measure over all rounds and the related root mean square errors (RMSE).

Precision is defined as:

$$\frac{true\_positive}{true\_positive + false\_positive}, \quad (2)$$

while recall is defined as

$$\frac{true\_positive}{true\_positive + false\_negative}. \quad (3)$$

For the sake of completeness, here is the meaning of the terminology used for computing precision and recall:

- "true positive" will be the instances of the test set predicted correctly;
- "false positive" will be the instances of the test set predicted wrongly;
- "false negative" will be the instances of the test set for which no prediction is made;
- "true negative" are not present in the dataset.

3.2. "Semi-Open" Setting

Here, systems can build their models by exploiting a predefined set of external resources. Such a setting allows the comparison between systems using semantic representation of information and exploiting external resources for building their model. We identified three linguistic resources, described below, that can be used in the "semi-open" evaluation. Beside them, other resources are available for working in the opinion mining domain. These three resources have been selected because the polarities expressed for each term (and in case of SenticNet also some more complex expressions) can be combined with further features included in such resources. To the best of our knowledge, these are the only resources giving the possibility of working with a multidimensional polarity of terms.

WordNet WordNet[2] (Fellbaum, 1998) is a large lexical database of English nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms called *synsets*, where each synset expresses a distinct concept. In particular, each synset represents a list of synonyms, intended as words that denote the same concept and that are interchangeable in many contexts. WordNet contains around 117,000 synsets linked to each other by a small set of "conceptual relations", i.e., synonymy, hypernymy, hyponymy, etc.. Additionally, a synset contains a brief definition ("gloss") and, in most cases, one or more short sentences illustrating the use of the synset members. Words having several distinct meanings are represented in as many distinct synsets. Even if WordNet superficially resembles a thesaurus, there are some important distinctions with respect to it. Firstly, WordNet does not define links between words, but between specific senses of words; this way, words that are found in close proximity to one another in the network are semantically disambiguated. Secondly, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than the similarity of their meanings.

SenticNet SenticNet[3] (Cambria et al., 2014) is a publicly available resource for opinion mining, which exploits both artificial intelligence and semantic Web techniques to infer the polarities associated with common-sense concepts and to represent them in a semantic-aware format. In particular, SenticNet uses dimensionality reduction to calculate the affective valence of a set of Open Mind[4] concepts and it represents them in a machine accessible and processable format.

The development of SenticNet was inspired by SentiWordNet (Baccianella et al., 2010), a lexical resource in which each WordNet synset is associated to three numerical scores describing how objective, positive, and negative the terms contained in each synset are. The differences between SenticNet and SentiWordNet are basically three: (i) in SentiWordNet, each synset is associated to a three-valued representation (the objectivity of the synset, its positiveness, and its negativeness), while in SenticNet there is only one value belonging to the $[-1, 1]$ interval for representing the polarity of the concept; (ii) SenticNet provides the sentiment model of more complex common-sense concepts, while SentiWordNet is focused on assigning polarities to WordNet synsets: for instance, in SenticNet, complex concepts like "make good impression", "look attractive", "show appreciation" , "being fired", "leave behind", or "lose control" are used for defining positive or negative situations; and (iii) completely neutral concepts are not reported.

In order to represent SenticNet in a machine-accessible and processable way, information about each concept is encoded as a set of RDF triples using an XML syntax. In particular, concepts are identified using the ConceptNet Web API and statements, which have the form "concept-hasPolarity-polarityValue", are encoded in the RDF/XML format on the basis of the human emotion ontology (HEO) (Grassi, 2009), a high-level ontology of human emotions, which supplies the most significant concepts and properties which constitute the centerpiece for the description of every human emotion.

As an example, the representation of the concept "a lot of fun" contained in SenticNet is shown in Figure 1.

SenticNet contains more than 5,700 polarity concepts (nearly 40% of the Open Mind corpus) and it may be connected with any kind of opinion mining application. For example, after the de-construction of the text into concepts through a semantic parser, SenticNet can be used to associate polarity values to these and, hence, infer the overall polarity of a clause, sentence, paragraph, or document by averaging such values.

[2] https://wordnet.princeton.edu/

[3] http://sentic.net/

[4] https://en.wikipedia.org/wiki/Open_Mind_Common_Sense

```
<rdf:Description
    rdf:about="http://sentic.net/api/en/concept/a_lot_of_fun">
  <rdf:type rdf:resource="http://sentic.net/api/concept"/>
  <text xmlns="http://sentic.net/api">a lot of fun</text>
  <semantics xmlns="http://sentic.net/api"
    rdf:resource="
       http://sentic.net/api/en/concept/radical"/>
  <semantics xmlns="http://sentic.net/api"
    rdf:resource="
       http://sentic.net/api/en/concept/enjoyable"/>
  <semantics xmlns="http://sentic.net/api"
    rdf:resource="
       http://sentic.net/api/en/concept/just_fun"/>
  <semantics xmlns="http://sentic.net/api"
    rdf:resource="
       http://sentic.net/api/en/concept/
       good_mental_health"/>
  <semantics xmlns="http://sentic.net/api"
    rdf:resource="
       http://sentic.net/api/en/concept/fun_play"/>
  <pleasantness xmlns="http://sentic.net/api"
    rdf:datatype="
       http://www.w3.org/2001/XMLSchema#float">
       0.814</pleasantness>
  <attention xmlns="http://sentic.net/api"
    rdf:datatype="
       http://www.w3.org/2001/XMLSchema#float">
       0</attention>
  <sensitivity xmlns="http://sentic.net/api"
    rdf:datatype="
       http://www.w3.org/2001/XMLSchema#float">
       0</sensitivity>
  <aptitude xmlns="http://sentic.net/api"
    rdf:datatype="
       http://www.w3.org/2001/XMLSchema#float">
       0.856</aptitude>
  <polarity xmlns="http://sentic.net/api"
    rdf:datatype="
       http://www.w3.org/2001/XMLSchema#float">
       0.557</polarity>
</rdf:Description>
```

Figure 1: The RDF/XML representation of concept “a lot of fun” in SenticNet.

General Inquirer dictionary The General Inquirer dictionary[5] (Stone et al., 1966) is an English-language dictionary containing almost 12,000 elements associated with their polarity in different contexts. Such dictionary is the result of the integration between both the “Harvard” and “Lasswell” general-purpose dictionaries as well as a dictionary of categories defined by the dictionary creator. If necessary, for ambiguous words, specific polarity for each sense is specified.
For every word, a set of tags is provided in the dictionary. Among them, only a subset is relevant to the opinion mining topic. The following tags are usually exploited:

- Valence categories: the two well-known “positive” and “negative” classification.
- Semantic dimensions: these tags reflect semantic differential findings regarding basic language universals. These dimensions are: “hostile”, “strong”, “power”, “weak”, “submit”, “active”, and “passive”. A word may be tagged with more than one dimension, if appropriate.
- Words of pleasure: these tags are usually also classified positive or negative, with virtue indicating strength and vice indicating weakness. They provide more focus than the categories in the previous two bullets. Such categories are “pleasure”, “pain”, “feel”, “arousal”, “emotion”, “virtue”, “vice”.
- Words reflecting presence or lack of emotional expressiveness: these tags indicate the presence of overstatement and understatement; trivially, such tags are “overstated” and “understated”.

Other categories indicating ascriptive social categories rather than references to places have been considered out of the scope of the opinion mining topic and, in general, they are not taken into account.

No guidelines are provided concerning the integration of these resources. Therefore, any kind of linking between them may be done without any constraint.
Evaluation procedure follows the same steps described for the “close” setting. Hence, a cross-fold validation has to be performed and the indicators described in the “closed” setting have to be provided.

3.3. “Open” setting

In this last setting, systems can build their model by using any kind of external resource. However, it is important to highlight that for adhering to the protocol, all used resources have to be published and reachable in order to allow the replicability of the evaluation campaign.
Rules described for the “closed” setting about cross-fold validation and indicators to report are applied in this setting as well.

Concerning the resources presented in the “Semi-Open” setting, we want to point out that the authors are aware of SentiWordNet (Baccianella et al., 2010), a lexical resource in which each WordNet synset is associated to three numerical scores describing how objective, positive, and negative the terms contained in each synset are.
However, SentiWordNet has not been included in the *DRANZIERA* protocol due to following three reasons:

1. in SentiWordNet, each synset is associated to a three-valued representation (the objectivity of the synset, its positiveness, and its negativeness), while only one value belonging to the $[-1, 1]$ interval was desired for representing the polarity of a concept;
2. SentiWordNet is focused only on the assignment of polarities to WordNet synsets. Within the protocol we already have SenticNet providing such polarities and, beside them, SenticNet provides also the polarities of more complex common-sense concepts like “make good impression”, “look attractive”, “show appreciation” , “being fired”, “leave behind”, or “lose control” used for defining positive or negative situations;
3. in SentiWordNet completely neutral concepts are not reported.

[5] http://www.wjh.harvard.edu/ inquirer/

4. Preliminary Baselines

Here, we report some results obtained by systems evaluated by adopting the *DRANZIERA* protocol. Tables 2, 3, and 4 contain results obtained by applying the different settings described in Section 3..
Three baselines have been used:

- Most Frequent Class (MFC): accuracy obtained if a system always predicts the same polarity for all samples contained in the test set.
- Multi-Domain Fuzzy Sentiment Analyzer described in (Dragoni et al., 2015) (MDFSA): this approach implements a multi-domain algorithm exploiting aggregations of fuzzy polarities for computing the opinion expressed in a document.
- Information Retrieval approach for Multi-Domain Sentiment Analysis described in (Dragoni, 2015) (IRSA): this work describes an information retrieval approach building domain-specific indexes containing related polarities. Documents are used as queries for retrieving an estimation of their global polarity.

The list above contains the results obtained by the systems that have already been evaluated by following the *DRANZIERA* protocol and dataset. An updated list of systems and results can be found in the *DRANZIERA* protocol website.[6]
Two considerations have to be highlighted: (i) the first baseline reports, obviously, the same results for all the three evaluation settings described in Section 3., while (ii) the second and third systems have been evaluated both with and without domain knowledge (indicated by the NODK suffix in the table) in order to provide different baselines for future comparisons.

Table 2: Results obtained by applying the "Close" evaluation setting. All values are the average computed through the cross-fold validation.

Baseline	Close			
	Precision	Recall	F-Measure	Deviation
MFC	0.5000	1.0000	0.6667	0.0000
MDFSA	0.6356	0.9245	0.7533	$0.6 \cdot 10^{-4}$
MDFSA-NODK	0.6694	0.9245	0.7765	$1.1 \cdot 10^{-4}$
IRSA	0.6232	0.8742	0.7277	$0.4 \cdot 10^{-4}$
IRSA-NODK	0.6527	0.8742	0.7474	$0.9 \cdot 10^{-4}$

5. Conclusion And Related Work

In this paper, we have presented the *DRANZIERA* protocol, the first methodology allowing a fair comparison between opinion mining systems. Such a methodology includes:

- A set of one million product reviews belonging to twenty domains. For each domain, reviews are split in five folds allowing the execution of a cross-fold validation.

[6] http://goo.gl/7jK4Rp

Table 3: Results obtained by applying the "Semi-Open" evaluation setting. All values are the average computed through the cross-fold validation.

Baseline	Semi-Open			
	Precision	Recall	F-Measure	Deviation
MFC	0.5000	1.0000	0.6667	0.0000
MDFSA	0.6832	0.9245	0.7857	$0.2 \cdot 10^{-4}$
MDFSA-NODK	0.7145	0.9245	0.8060	$0.7 \cdot 10^{-4}$
IRSA	0.6598	0.8742	0.7520	$2.3 \cdot 10^{-4}$
IRSA-NODK	0.6784	0.8742	0.7640	$2.8 \cdot 10^{-4}$

Table 4: Results obtained by applying the "Open" evaluation setting. All values are the average computed through the cross-fold validation.

Baseline	Open			
	Precision	Recall	F-Measure	Deviation
MFC	0.5000	1.0000	0.6667	0.0000
MDFSA	0.6832	0.9245	0.7857	$0.2 \cdot 10^{-4}$
MDFSA-NODK	0.7145	0.9245	0.8060	$0.7 \cdot 10^{-4}$
IRSA	0.6598	0.8742	0.7520	$2.3 \cdot 10^{-4}$
IRSA-NODK	0.6784	0.8742	0.7640	$2.8 \cdot 10^{-4}$

- Three different evaluation settings. Due to the possibility of building sentiment models in many ways, three evaluation settings ("Close", "Semi-Open", "Open") have been foreseen for easing the comparison between systems.
- Preliminary baselines. Some systems already available in the literature have been evaluated by using the *DRANZIERA* protocol. Such systems have been evaluated both with and without using domain knowledge.

6. Bibliographical References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, pages 2200–2204.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 187–205.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI*, pages 1515–1521.

Dragoni, M., Tettamanzi, A. G., and da Costa Pereira, C. (2015). Propagating and aggregating fuzzy polarities for concept-level sentiment analysis. *Cognitive Computation*, 7(2):186–197.

Dragoni, M. (2015). Shellfbk: An information retrieval-based system for multi-domain sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, pages 502–509, Denver, Colorado, June. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Grassi, M. (2009). Developing heo human emotions ontology. In *COST 2101/2102 Conference*, pages 244–251.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal et al., editors, *Mining Text Data*, pages 415–463. Springer.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, Philadelphia, July. Association for Computational Linguistics.

Saif, H., Fernández, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. In Cristina Battaglino, et al., editors, *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 3, 2013.*, volume 1096 of *CEUR Workshop Proceedings*, pages 9–21. CEUR-WS.org.

Stone, P., Dunphy, D., and Marshall, S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Oxford, England: M.I.T. Press.