




Propagation-Based Domain-Transferable Gradual Sentiment Analysis

Célia da Costa Pereira¹, Claude Pasquier¹ and Andrea G. B. Tettamanzi²

¹ Université Côte d'Azur, I3S, CNRS, Sophia Antipolis, France

² Université Côte d'Azur, I3S, Inria, Sophia Antipolis, France

{celia.da-costa-pereira, claude.pasquier, andrea.tettamanzi}@univ-cotedazur.fr

Keywords: Sentiment Analysis, Fuzzy Polarity Propagation.

Abstract: We propose a novel refinement of a gradual polarity propagation method to learn the polarities of concepts and their uncertainties with respect to various domains from a labeled corpus. Our contribution consists of introducing a positive correction term in the polarity propagation equation to counterbalance negative psychological bias in reviews. The proposed approach is evaluated using a standard benchmark, showing an improved performance relative to the state of the art, good cross-domain transfer and excellent coverage.

1 INTRODUCTION

Sentiment analysis aims at determining the global polarity (positive, negative or neutral) of a document based on the polarities of the words in the document. A supervised method trains a model (classifier) by using the datasets of reviews or labeled texts and use such models to classify the user opinions. However, the polarity of some words in a review might depend on the domain knowledge considered (Rexha et al., 2018; Pirnau, 2018). For example (Yoshida et al., 2011), the word 'long', which has a positive polarity in the Camera domain, has a negative polarity if we are characterizing the execution time of a computer program.

Several solutions have been proposed. Concept-based approaches include the one proposed by Schouten *et al.* (Schouten and Frasincar, 2015), who show that considering concept-based features instead of term-based features helps improving the performance of multi-domain sentiment analysis methods. The good quality of the results obtained with this relatively straightforward setup encourages the use of more advanced ways of handling semantic information.


Yoshida *et al.* (Yoshida et al., 2011) propose a solution to improve *transfer learning methods*. However, as it has been rightly pointed out by Abdullah *et al.* (Abdullah et al., 2019), the transfer learning ap-


proach imposes the necessity to build a new transfer model and this limits its generalization capability.


A serie of works (Dragoni et al., 2014; Dragoni, 2015; Dragoni et al., 2015; Dragoni et al., 2016) use *fuzzy logic* to model the relationships between the polarity of concepts and the domain. They use a two-level graph, where the first level represents the relations between concepts, whereas the second level represents the relations between the concepts and their polarities in the various targeted domains, the idea being to capture the fact that the same concept can be positive in one domain, but negative in another. This is accomplished thanks to a polarity propagation algorithm and without the necessity of starting the learning process for each different domain. The main advantage of that approach, named *MDFSA (Multi-Domain Fuzzy Sentiment Analyzer)*, which won the ESWC 2014 Concept-Level Sentiment Analysis Challenge (Dragoni et al., 2014), is that it both accounts for the conceptual representation of the terms in the documents by using WordNet and SenticNet, and proposes a solution avoiding to build a new model each time a new domain needs to be analysed.

However, *MDFSA* had several issues, including the following:

1. It is not possible to discard some of the remaining word ambiguities due to the fact that a *synset* corresponds to a group of words (nouns, adjectives, verbs and adverbs) that can be interchangeable and depending on the type of the term used, the meaning of the word can change.

^a <https://orcid.org/0000-0001-6278-7740>

^b <https://orcid.org/0000-0001-7498-395X>

^c <https://orcid.org/0000-0002-8877-4654>

2. The *same* stopping criterion for the propagation algorithm is used for the *different independent* domains which could be challenged, i.e. that simultaneous stopping of propagation could be premature for some domains and delayed for others.
3. The propagation of polarities takes place without taking into account the similarity of related concepts in the graph. Indeed, the more similar the concepts are, the higher the weights associated with their relationship in the graph should be.

To solve them, we recently proposed *Sental* (Pasquier et al., 2020), an extension of *MDFSA*, which we take as a starting point. In particular, concerning (1), to decide whether a term occurring in a document is associated to a synset v or not, *Sental* looks at its *part of speech (POS) tag* and considers it an instance of v only if its *POS tag* matches the *POS* of v . Concerning (2), *Sental* tests for convergence for each domain separately. Finally, concerning (3), *Sental* uses a pre-trained word embedding model to complete the semantic graph with a graded relation of semantic similarity, in addition to the crisp relations defined in WordNet.

Nevertheless, *Sental* still has an important issue: after a few iterations, its propagation process tends to drive polarities towards -1 . This happens because, on average, negative sentiments have more impact and are more resistant to disconfirmation than positive ones.

The aim of this paper is to solve this problem. We propose a new formula with a term to correct the negative bias by exerting a positive pressure which is strongest for neutral polarities and whose effect diminishes for extreme polarities. To the best of our knowledge it is the first time that this aspect, which is very well known in psychology (Baumeister et al., 2001), is taken into account in a sentiment analysis framework.

The rest of the paper is structured as follows. Section 2 presents the original method we are proposing. Section 3 presents the experiments and discuss the results. Finally, Section 4 concludes the paper.

2 METHOD

The algorithm used to learn concept polarities for various domains consists of three phases, namely:

1. Semantic graph construction from background knowledge;
2. Concept polarity initialization, based on a training set of documents, associated with a domain and labeled with a rating;

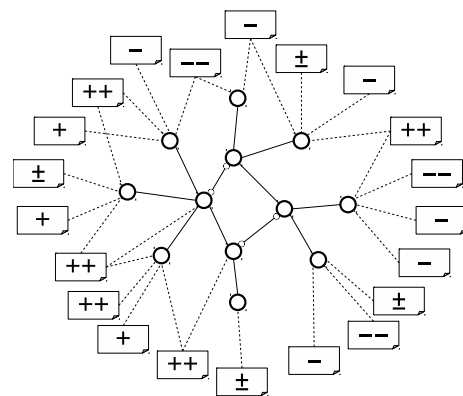


Figure 1: An illustration of the semantic graph constructed by the proposed method. Circles represent the vertices of the graph and solid lines its edges. The documents of the training set are shown around the semantic graph. Dashed lines represent the occurrence, in a document, of a term associated with a vertex of the graph (i.e., a lemma or a WordNet synset). Documents are rated (e.g., on a scale from $--$ to $++$, which may then be mapped to the $[-1, 1]$ interval).

3. Propagation of polarity information over the semantic graph.

Its result is an estimation of polarities, represented as convex fuzzy sets over the $[-1, +1]$ interval, for each concept and for each domain. Figure 1 illustrates the idea of a semantic graph, whose construction and use is detailed below.

2.1 Semantic Graph Construction

The backbone of the semantic graph is based on WordNet (Miller, 1995) and SenticNet (Cambria et al., 2010), combined as in (Dragoni et al., 2015). Both are publicly available lexical databases in which nouns, verbs, adjectives, and adverbs are organized into sets of synonyms (*synsets*), each representing a lexicalized concept. Synsets are linked by semantic relationships, including synonymy, antonymy and hypernymy. We distinguish concepts, which are abstract notions representing meaning, from terms, which are tangible ways of expressing concepts (in written language, they are words or phrases). Now, words, and typically the most frequent words (Casas et al., 2019), can be polysemous; a preprocessing phase is thus needed to link the terms found in the texts as accurately as possible to their corresponding synsets. Unlike WordNet, SenticNet is specifically built for opinion mining, but the polarities it associates to terms are not used, because the very assumption on which our approach is based is that polarity is not an intrinsic property of a term, but an extrinsic, domain-dependent property. Besides covering terms that are not indexed in WordNet, SenticNet allows to resolve

a number of cases of ambiguity.

To further reduce ambiguity, we propose to parse the text and use the POS of a term to associate it to the correct synset. As an example, the same term ‘light’ can be, according to WordNet 3.1, a noun (‘do you have a light?’), a verb (‘light a cigarette’), an adjective (‘a light diet’) or an adverb (‘experienced travelers travel light’), which are represented by distinct synsets.

In addition to the WordNet and SenticNet relations, already used in the literature, we use a word embedding model, pre-trained by applying Word2vec (Mikolov et al., 2013) to roughly 100 billion words from a Google News dataset,¹ to complete the semantic graph with relationships of semantic similarity between terms.

The semantic graph is constructed as a weighted graph (V, E, w) . Each element of V is either a concept (i.e., a synset) or the canonical form (lemma) of a term used in a review; $w : E \rightarrow [-1, 1]$ is a weight function and the edges in E are created:

- between synsets linked by a *hypernym* relationship, a synset and its lemmas, and between lemmas linked by a *synonym* relationship, with weight +1;
- between lemmas linked by an *antonym* relationship, with weight -1;
- between each lemma and the five closest lemmas according to the pre-trained *word2vec* model, with their cosine similarity as weight.

Each vertex $v \in V$ is labeled by a vector $\vec{p}(v)$ of polarities, so that $p_i(v)$ is the polarity of v in the i th domain.

2.2 Concept Polarity Initialization

The initial polarities $\vec{p}_i^{(0)}(v) \in [-1, 1]$ of all the vertices v of the semantic graph are computed, for each domain i , as the average polarity of the documents of domain i in the training set, in which at least a term of v occurs. If no term associated to v occurs in a document of domain i , $p_i^{(0)}(v) = 0$.

As explained above, to decide whether a term occurring in a document is associated to a synset v or not, we look at its POS tag and we consider it an instance of v only if its POS tag matches the POS of v .

2.3 Polarity Propagation

In this phase, information about the polarity of vertices is propagated through the edges of the graph, so

¹https://frama.link/google_word2vec

that concepts for which no polarity information could be directly extracted from the training set (i.e., those v such that $p_i^{(0)}(v) = 0$ for some i) can “assimilate”, as it were, the polarity of their close relatives. In addition, this propagation process may contribute to correct or fine-tune the polarity of incorrectly initialized concepts and, thus, reduce noise.

Polarity propagation through the graph is carried out iteratively. At each iteration $t = 1, 2, \dots$, the polarity $p_i^{(t)}(v)$ of each vertex v for domain i is updated taking into account both the values of its neighbors $N(v) = \{v' \mid (v, v') \in E\}$ and its distinctiveness from the other terms in the domain.

$$\begin{aligned} \vec{p}^{(t+1)}(v) &= (1 - \lambda)\vec{p}^{(t)}(v) \\ &+ \lambda \frac{1}{\|N(v)\|} \sum_{v' \in N(v)} \vec{p}^{(t)}(v') w(v, v') \quad (1) \\ &+ \varphi(1 - |\vec{p}^{(t)}(v)|), \end{aligned}$$

where $w(v, v')$ denotes the weight of the edge between vertices v and v' , $0 < \lambda < 1$, the *propagation rate* and $0 < \varphi < 1$, the *positive correction strength* are parameters of the algorithm.

Notice that the propagation of polarity for one domain does not interact with the same process for the other domains and we can thus consider that polarity propagation is carried out in parallel and independently for each domain.

Inspired by simulated annealing (Kirkpatrick et al., 1983), the propagation rate is decreased at each iteration, according to a parameter A called *annealing rate*. Thus, the value of λ at iteration t is calculated according to the value of λ at iteration $t - 1$ as follow:

$$\lambda_t = A\lambda_{t-1}. \quad (2)$$

In the method proposed by Dragoni et al. (Dragoni et al., 2015), the iterative process stops as soon as the sum of the variations of the polarity for each concept and domain falls below a fixed threshold. The drawback of using a fixed convergence limit is that it depends on the dataset used. Indeed, a dataset composed of many domains using lots of different terms will logically generate a greater variation than a smaller dataset. In our proposed method, we specify a threshold that applies to each domain separately and that is relative to the number of different nodes composing the semantic graph of the domain. Thus, for the i th domain, the polarity propagation stops when the average polarity variation,

$$\Delta_i^{(t)} = \frac{1}{\|V\|} \sum_v |p_i^{(t)}(v) - p_i^{(t-1)}(v)|, \quad (3)$$

falls below a threshold L , which is the *convergence limit*. We denote by t_i^{stop} the total number of iterations carried out for the i th domain until $\Delta_i^{(t)} < L$.

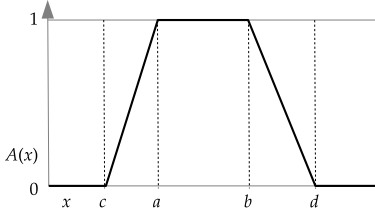


Figure 2: A fuzzy set with a trapezoidal membership function, defined by the four parameters (a, b, c, d) .

Experience shows that, after a few iterations, the propagation process described above tends to drive polarities towards -1 . This happens because, on average, negative sentiments in reviews are more strongly expressed than positive sentiments. Such difference of intensity between negative and positive sentiments has already been observed and is well-known in psychology (Baumeister et al., 2001). That is why we have added a term to correct this negative bias (third line of Equation 1), by exerting a positive pressure which is strongest for neutral polarities and whose effect tapers off for extreme polarities. The overall strength of this positive correction term is controlled by parameter ϕ .

At each iteration $t = 0, 1, 2, \dots$, the vectors $\vec{p}^{(t)}(v)$ are saved in order to exploit them for the calculation of the shapes of the fuzzy membership functions describing the polarity of concept v for each domain. Indeed, the final polarities are represented as trapezoidal fuzzy membership functions, whose core is the interval between the initial polarity computed from the training set, $p_i^{(0)}(v)$, and the polarity resulting from the propagation phase, $p_i^{(t_i^{\text{stop}})}(v)$ and whose support extends beyond the core on either side by half the variance $\sigma_{v,i}^2$ of the distribution of $p_i^{(t)}(v)$, $t = 0, \dots, t_i^{\text{stop}}$. To sum up, for each domain i , $\mu_{v,i}$ is a trapezoid with parameters (a, b, c, d) , like the one depicted in Figure 2, where

$$\begin{aligned} a &= \min\{p_i^{(0)}(v), p_i^{(t_i^{\text{stop}})}(v)\}, \\ b &= \max\{p_i^{(0)}(v), p_i^{(t_i^{\text{stop}})}(v)\}, \\ c &= \max\{-1, a - \sigma_{v,i}^2/2\}, \\ d &= \min\{1, b + \sigma_{v,i}^2/2\}. \end{aligned}$$

The idea here is that the most likely values for the polarity of v for a domain are those comprised between the initial and final value of the polarity propagation phase and the more quickly the polarity values converged during that phase, the least uncertainty there is about the resulting polarity estimate. Conversely, a polarity value that converged slowly or with many fluctuations is going to yield a less reliable, and thus more uncertain, estimate.

2.4 Document Polarity Calculation

Once the model is trained according to the algorithm described in the previous sections, the (fuzzy) polarity of a novel document D of the i th domain is computed as the average of the fuzzy polarities (represented by their trapezoidal membership functions) of all the terms v occurring in the document:

$$\mu_i = \frac{1}{\|V_i\|} \sum_{v \in V_i} \mu_{v,i}, \quad (4)$$

where $V_i = \{v \in V \mid v \text{ occurs in } D\}$. This average of fuzzy sets is computed by applying the extension principle, thus yielding, for all $x \in [-1, 1]$,

$$\mu_i(x) = \sup_{x = \frac{1}{\|V_i\|} \sum_{v \in V_i} x_v} \min_{v \in V_i} \mu_{v,i}(x_v). \quad (5)$$

However, given that all $\mu_{v,i}$ are trapezoidal with parameters

$$(a_v, b_v, c_v, d_v),$$

as pointed out in (Dragoni et al., 2015), μ_i will always be trapezoidal as well, with parameters

$$\frac{1}{\|V_i\|} \left(\sum_{v \in V_i} a_v, \sum_{v \in V_i} b_v, \sum_{v \in V_i} c_v, \sum_{v \in V_i} d_v \right). \quad (6)$$

This fuzzy polarity reflects the uncertainty of the estimate obtained by the model. A single polarity figure can be obtained by applying a defuzzification method. For the empirical validation of our method, we used the centroid method for this purpose.

3 EXPERIMENTS AND RESULTS

The proposed system (that we call Sental+) has been evaluated using the DRANZIERA evaluation protocol (Dragoni et al., 2016), a multi-domain sentiment analysis benchmark, which consists of a dataset containing product reviews from 20 different domains, crawled from the Amazon web site, as well as guidelines allowing the fair evaluation and comparison of opinion mining systems. In the dataset of the DRANZIERA benchmark, each domain is composed of 5,000 positive and 5,000 negative reviews that are split in five folds containing 1,000 positive and 1,000 negative reviews each.

3.1 Experimental Protocol

As suggested by the guideline of the DRANZIERA evaluation protocol (Dragoni et al., 2016), the performance of the method has been assessed by performing

a 5-fold cross validation. For each specific domain, the method was trained on four of the five folds provided with the benchmark and tested on the remaining fold. The process is repeated five times so that each fold is in turn used for testing. Experiments were performed on a computer running Ubuntu 18.04 and based on a Intel®Core™ i7-7700 @ 3.60GHz with 32 Gb main memory.

The algorithm depends on four different parameters: the *propagation rate* λ , which determines the diffusion rate of the polarity values between concepts, the *positive correction strength* ϕ that maintains the diversity of polarities, the *convergence limit* L , which represents the criterion for stopping the polarity propagation phase for each domain, and the *annealing rate* A , used to decrease, at each iteration, the *propagation rate* (cf. Equation 2). In order not to bias the method settings by selecting parameters that could be specifically suited to the DRANZIERA dataset, we carried out the method tuning using a separate dataset. For this purpose we used the Blitzer dataset (Blitzer et al., 2007) which is composed of product reviews belonging to 25 domains. A quick scan of the dataset shows both that there is a large discrepancy in the number of reviews available for each domain and that positive reviews are much more frequent. For example, the *books* category contains 975194 reviews (123899 negatives and 851295 positives) while the *tools & hardware* domain contains only 14 negative reviews and 98 positives. In order not to bias the system, we have taken care to balance the reviews between positive and negative. We limited the maximum number of reviews to 1600 (composed by the 800 first positive and the 800 first negative reviews found in the XML files). Using a small portion of the dataset, we have therefore experimented different configurations of parameters (λ, ϕ, L, A) by varying the *propagation rate* between 0.1 and 0.9 in 0.1 steps, by testing all values for *annealing rate* between 0.0 and 1.0 in 0.1 steps and using 10^{-1} , 10^{-2} and 10^{-3} as values for the *convergence limit*. Our experiments show that using a *propagation rate* of 0.3, an *annealing rate* of 0.5, a *positive correction strength* of 0.3 and a *convergence limit* of 0.05 lead to the best results. The setting that leads to the best results is $\lambda = 0.3$, $\phi = 0.3$, $L = 0.05$, and $A = 0.5$.

3.2 Results of DRANZIERA Evaluation

When applied to the DRANZIERA dataset with the settings previously identified,

the average precision obtained over all 20 domains is 0.8191, which constitutes a significant improvement over MDFSA (Dragoni et al., 2015), MDFSA-

Table 1: Average precision, recall, F1 score and standard deviation of the F1 score obtained on the 20 domains of the DRANZIERA dataset.

Method	Precision	Recall	F1 score	SD
MDFSA	0.6832	0.9245	0.7857	0.0000
MDFSA-NODK	0.7145	0.9245	0.8060	0.0001
IRSA	0.6598	0.8742	0.7520	0.0002
IRSA-NODK	0.6784	0.8742	0.7640	0.0003
Sental w/ embedding	0.7527	0.9942	0.8551	0.0446
Sental w/ POS	0.7617	0.9947	0.8612	0.0435
Sental+	0.8191	0.9941	0.8975	0.0275

NODK (Dragoni et al., 2015), IRSA (Dragoni, 2015) and IRSA-NODK (Dragoni et al., 2016) which obtain a precision of 0.6832, 0.7145, 0.6598 and 0.6784 respectively.

It should be noted that this improvement in precision is not detrimental to the recall value, which is higher than the other methods. As a result, the proposed method obtains an even greater improvement with respect to the other methods if performance is measured in terms of the F1 score, in particular a 9.15% improvement with respect to the best of them, MDFSA-NODK. Table 1 provides a summary of the comparison of the results obtained by our method with four other methods evaluated on the DRANZIERA dataset, whose results are provided in (Dragoni et al., 2016). Using word embedding (Sental w/ embedding) provides an improvement over existing methods; incorporating POS tags (Sental w/ POS) further improves precision and recall, **but the best results are achieved by adding a positive correction term to the polarity propagation equation (Sental+)**.

The breakdown of the results obtained by our method implementing the four proposed solutions on the 20 domains of the DRANZIERA dataset is presented in Table 2.

3.3 Cross-Domain Transfer Experiment

To test the generalization capabilities of our approach, we have strived to use our model, trained with DRANZIERA data, on other datasets. Sentiment analysis has become extremely popular but datasets available for use by multi-domain sentiment analysis are still scarce. Two datasets, proposed by Hutto and Gilbert (Hutto and Gilbert, 2014), correspond to domains that could benefit from being processed with our method. The first one, ‘Product’, contains 3708 customer reviews of five electronics products. The second one, ‘Movie’, contains 10,605 sentence-level snippets from the site <http://rotten.tomatoes.com> reanalyzed by 20 independent human raters. ‘Product’ reviews should be efficiently

Table 2: Detail of the results obtained on the 20 domains of the DRANZIERA dataset.

Domain	Precision	Recall	F1 score	SD
Amazon Video	0.7253	0.9946	0.8389	0.0310
Automotive	0.8402	0.9943	0.9108	0.0219
Baby	0.7357	0.9945	0.8457	0.0702
Beauty	0.8523	0.9947	0.9180	0.0085
Books	0.7910	0.9911	0.8798	0.0410
Clothing	0.8807	0.9973	0.9354	0.0401
Electronics	0.7998	0.9950	0.8868	0.0293
Health	0.8388	0.9947	0.9101	0.0208
Home Kitchen	0.8142	0.9951	0.8956	0.0399
Movies TV	0.8064	0.9934	0.8902	0.0273
Music	0.7764	0.9933	0.8716	0.0333
Office Products	0.8256	0.9949	0.9024	0.0168
Patio	0.8378	0.9945	0.9094	0.0201
Pet Supplies	0.7970	0.9912	0.8836	0.0265
Shoes	0.9183	0.9965	0.9558	0.0123
Software	0.8090	0.9943	0.8921	0.0216
Sports Outdoors	0.8320	0.9915	0.9048	0.0179
Tools Home Impr.	0.8430	0.9947	0.9126	0.0248
Toys Games	0.8581	0.9950	0.9215	0.0265
Video Games	0.7998	0.9916	0.8854	0.0201
Average	0.8191	0.9941	0.8975	0.0275

analyzed with our method trained on the ‘Electronics’ domain of DRANZIERA dataset. Although the DRANZIERA dataset does not contain the ‘Movie’ domain as such, it includes some related domains, for example ‘Movies TV’ or ‘Amazon Instant Video’, whose reviews may be used to infer the rating of movies.

Our method, trained on each of the domains of the DRANZIERA dataset was used to predict the orientation of each ‘Movie’ and ‘Product’ review between positive and negative. Table 3 lists the precision values obtained on ‘Movie’ and ‘Product’ reviews according to the category of the DRANZIERA dataset used for training. Precisions displayed in bold highlight the best score obtained for each domain. Recall values are not displayed because the variation is small; they range from 0.9916 to 0.9944 for ‘Movie’ and from 0.9889 to 0.9964 for ‘Product’. Overall, there is a decrease in performance when the method, trained on DRANZIERA domains, is applied to reviews originating from different datasets; which seems perfectly logical. We can notice, in Table 3, that the domain used for training has a great influence on the results.

For a transfer to ‘Movie’, the best precision was obtained when the training was performed with the DRANZIERA reviews belonging to the ‘Movie TV’ domain. Although ‘Music’ is not exactly the category for which one would have expected to obtain the best cross-domain transfer quality, the result is still consistent. The other domains for which the transfer of the learned model is effective are ‘Books’, ‘Amazon

Table 3: Precision and recall of cross-domain transfer from the 20 categories of the DRANZIERA dataset to independent ‘Movie’ and ‘Product’ reviews.

DRANZIERA domain	precision on ‘Movie’	precision on ‘Product’
Amazon Instant_Video	0.6962	0.5704
Automotive	0.6008	0.6490
Baby	0.5830	0.5735
Beauty	0.5992	0.6812
Books	0.7021	0.6196
Clothing Accessories	0.5885	0.6889
Electronics	0.5728	0.7309
Health	0.5899	0.6821
Home Kitchen	0.5875	0.6656
Movies TV	0.7134	0.6000
Music	0.6770	0.6507
Office Products	0.6135	0.6813
Patio	0.6099	0.6885
Pet Supplies	0.5960	0.6720
Shoes	0.5795	0.6843
Software	0.6154	0.7098
Sports Outdoors	0.5797	0.6891
Tools Home Improvement	0.5910	0.7086
Toys Games	0.6200	0.6103
Video Games	0.6595	0.6942

Instant Video’ and ‘Music’. Overall, the reviews that most closely mirror those of the ‘Movie’ reviews are more oriented towards cultural goods, while the most distant ones concern more tangible goods (‘Shoes’, ‘Sport Outdoors’ and ‘Electronics’). The result is exactly the opposite for the transfer to ‘Product’ reviews. The best precision is obtained when the training was performed on ‘Electronics’ and the transfer works better overall when the method has been trained on reviews about concrete objects (‘Software’ and ‘Tools Home Improvement’ are both domains that allow to exceed a precision of 70%).

The precision therefore varies by nearly 0.15 between a transfer done from the ‘Music’ domain and a transfer from the ‘Electronics’ domain. However, the difference can also be partly explained by the intrinsic score obtained on each DRANZIERA domain. We can indeed notice in Table 3, that the difference in precision between the ‘Music’ domain and the ‘Baby’ domain is slightly more than 0.14. However, we can observe significant differences according to the domains. Some cross-domain transfers go very well, such as the transfer from domains like ‘Books’, ‘Movies TV’ or ‘Amazon Instant Video’ to movie reviews since the accuracy decline is less than 0.03. Other cross-domain transfers are more problematic, such as those from ‘Electronics’, ‘Beauty’, ‘Sports Outdoors’, ‘Clothing Accessories’ or ‘Shoes’ to movies reviews since the accuracy falls by more than 0.2. These observations also seem to make perfect sense.

Table 4: A comparison of our method (trained on the ‘Movies TV’ domain of DRANZIERA) with the best methods benchmarked by Ribeiro *et al.*, when applied to the ‘Movie’ dataset. The highest score of each column is highlighted in boldface.

Method	Precision	Recall	F1 score
AFINN	0.6593	0.7259	0.6910
LIWC15	0.6335	0.6608	0.6469
Opinion Lexicon	0.6977	0.7728	0.7333
Pattern.en	0.6784	0.6559	0.6670
Semantria	0.6964	0.6880	0.6922
SenticNet	0.9630	0.6941	0.8067
SO-CAL	0.7165	0.8910	0.7943
Stanford DM	0.8270	0.9192	0.8707
VADER	0.6519	0.8270	0.7291
Sental+	0.7134	0.9944	0.8308

To assess the performance of our method we compared it with the benchmark of 24 sentiment analysis methods performed by Ribeiro *et al.* (Ribeiro *et al.*, 2016).

Since the setting we have chosen leads to a high recall value at the expense of precision, we list in tables 4 and 5 the prediction performance of all methods that achieve, in either domain, an F1 score greater than 0.5 and a higher precision than our method. Methods meeting these criteria are AFINN (Nielsen, 2011), LIWC15 (Tausczik and Pennebaker, 2010), Opinion Lexicon (Hu and Liu, 2004),

Pattern.en (Smedt and Daelemans, 2012), Semantria (Lexalytics, 2015), SenticNet (Cambria *et al.*, 2018),

SO-CAL (Taboada *et al.*, 2011), Stanford DM (Socher *et al.*, 2013), and VADER (Hutto and Gilbert, 2014).

Regarding the ‘Movie’ domain, the best method, consisting in using only the polarities reported in SenticNet, obtains a high precision but only on a relatively small part of the reviews since its coverage is 69.41%. The second and third best methods, Stanford DM and SO-CAL, have a coverage of the same order, 91.92% and 89.10% respectively, which are both below Sental+’s coverage value, 99.44%.

All other methods are outperformed by Sental+, by both criteria.

When recall is considered together with precision (cf. Table 4), our method obtains an F1 score of 83.08%, which is higher than the F1 score of the most precise method, 80.67%, but short of Stanford DM, which has the highest F1 score with 87.07%, while SO-CAL is lower, at 79.43%.

Regarding the ‘Product’ domain,

a group of methods (AFINN, LIWC15, Opinion Lexicon, Pattern.en and Semantria) achieves a precision close to 80% with a recall ranging from 57% to

Table 5: A comparison of our method (trained on the ‘Electronic’ domain of DRANZIERA) with the best methods benchmarked by Ribeiro *et al.*, when applied to the ‘Product’ dataset. The highest score of each column is highlighted in boldface.

Method	Precision	Recall	F1 score
AFINN	0.7869	0.6280	0.6985
LIWC15	0.7674	0.5645	0.6505
Opinion Lexicon	0.8082	0.6715	0.7335
Pattern.en	0.7571	0.5953	0.6665
Semantria	0.8159	0.5945	0.6878
SenticNet	0.6991	0.9748	0.8142
SO-CAL	0.7823	0.7152	0.7472
Stanford DM	0.6853	0.8028	0.7394
VADER	0.7705	0.7133	0.7408
Sental+	0.7309	0.9956	0.8430

67%. SO-CAL and VADER score well in both precision and recall. In contrast to the ‘Movie’ domain, SenticNet and Stanford DM achieve modest precisions of less than 70%, compensated by high recall values.

The precision of 73.09% scored by Sental+, (i.e. better than 70%, as for the ‘Movie’ domain) and its recall of 99.56% allows the method to obtain the best F1 score. Overall, Sental+ appears to have a more consistent performance. It also features an excellent coverage, which is always above 99%, meaning that less than 1% of the reviews are not classified.

4 CONCLUSION

We have proposed Sental+, an extension of the Sental method inspired by the *MDFSA* approach, which incorporates an original solution to overcome a major issue of Sental: a positive correction term in the polarity propagation phase effectively counteracts the negative bias of the reviews.

The resulting method has been validated using a standard evaluation protocol, showing a significant improvement with respect to the state of the art. We have also tested the cross-domain generalization capabilities of our approach with very promising results.

Injecting more linguistic background knowledge, as Sental does (here, POS tagging and word embedding) into the semantic graph appears to improve both the precision and the coverage of the method. This suggests focusing future work in this direction, for instance by exploiting large language models.

REFERENCES

- Abdullah, N. A., Feizollah, A., Sulaiman, A., and Anuar, N. B. (2019). Challenges and recommended solutions in multi-source and multi-domain sentiment analysis. *IEEE Access*, 7:144957–144971.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Review of general psychology*, 5(4):323–370.
- 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., Poria, S., Hazarika, D., and Kwok, K. (2018). Sentinet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*, pages 1795–1802. AAAI Press.
- Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Sentinet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium: Commonsense Knowledge*, volume FS-10-02 of *AAAI Technical Report*, pages 14–18. AAAI Press.
- Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i Cancho, R., and Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, 58:19–50.
- 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.
- Dragoni, M., Tettamanzi, A. G. B., and da Costa Pereira, C. (2014). A fuzzy system for concept-level sentiment analysis. In *SemWebEval@ESWC*, volume 475 of *Communications in Computer and Information Science*, pages 21–27. Springer.
- Dragoni, M., Tettamanzi, A. G. B., and da Costa Pereira, C. (2015). Propagating and aggregating fuzzy polarities for concept-level sentiment analysis. *Cognitive Computation*, 7(2):186–197.
- Dragoni, M., Tettamanzi, A. G. B., and da Costa Pereira, C. (2016). DRANZIERA: an evaluation protocol for multi-domain opinion mining. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France*, pages 267–272, Paris, France. European Language Resources Association (ELRA).
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.
- Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Kirkpatrick, S., Jr., D. G., and Vecchi, M. P. (1983). Optimization by simulated annealing. *SCIENCE*, 220(4598):671–680.
- Lexalytics (2015). Sentiment extraction - measuring the emotional tone of content. *Technical Report*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.
- Pasquier, C., da Costa Pereira, C., and Tettamanzi, A. G. B. (2020). Extending a fuzzy polarity propagation method for multi-domain sentiment analysis with word embedding and POS tagging. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2140–2147. IOS Press.
- Pirna, M. (2018). Sentiment analysis for the tweets that contain the word “earthquake”. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6.
- Rexha, A., Kröll, M., Dragoni, M., and Kern, R. (2018). The CLAUSY system at ESWC-2018 challenge on semantic sentiment analysis. In *SemWebEval@ESWC*, volume 927 of *Communications in Computer and Information Science*, pages 186–196. Springer.
- Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23.
- Schouten, K. and Frasincar, F. (2015). The benefit of concept-based features for sentiment analysis. In *SemWebEval@ESWC*, volume 548 of *Communications in Computer and Information Science*, pages 223–233. Springer.
- Smedt, T. D. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K. D., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Yoshida, Y., Hirao, T., Iwata, T., Nagata, M., and Matsumoto, Y. (2011). Transfer learning for multiple-domain sentiment analysis—identifying domain dependent/independent word polarity. In *AAAI*, pages 1286–1291.