ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200338

Extending a Fuzzy Polarity Propagation Method for Multi-Domain Sentiment Analysis with Word Embedding and POS Tagging

Claude Pasquier[1] and **Célia da Costa Pereira**[1] and **Andrea G. B. Tettamanzi**[2]

Abstract. Within multi-domain sentiment analysis, we study how different domain-dependent polarities can be learned for the same concepts. To this aim, we extend an existing approach based on the propagation of fuzzy polarities over a semantic graph capturing background linguistic knowledge to learn concept polarities with respect to various domains and their uncertainty from labeled datasets. In particular, we use POS tagging to refine the association between terms and concepts and word embedding to enhance the construction of the semantic graph. The proposed approach is then evaluated on a standard benchmark, showing that the combined use of POS tagging and word embedding improves its performance. One particularly strong point of the proposed approach is its recall, which is always very close to 100%. In addition, we observe that it exhibits good cross-domain generalization capabilities.

1 INTRODUCTION

The aim of a sentiment analysis task is to determine the polarity (positive, negative or neutral) of a document with respect to a topic [31, 30], given the polarities of the different words in the document. However, as it has been pointed out in [42] for example, the polarity of some words often depends on the domain knowledge considered. By way of example, let us consider (as in [42]) the word "long" which has a positive polarity in the Camera domain, but a negative polarity if we are characterizing the execution time of a computer program. Schouten *et al.* showed in [33] that including concept-based features instead of term-based features always 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.

Several other solutions have been proposed in the literature. For instance, Yoshida *et* al. [42] proposed a solution to improve *transfer learning methods*. To be more precise, while the solutions in the literature learn a model for a given (single) domain and make it applicable to another (single) domain [10], the solution they proposed consists in generalizing that method—the model was constructed from the datasets corresponding to different domains and was also applicable to different domains. However, as it has been well underlined by Abdullah *et al.* [1], the main drawback of transfer learning techniques is that

Even though the methods were able to adapt the relevant sentiment features between different domains, the transfer learning approach imposes the necessity to build a new transfer model, each time a new domain needs to be analysed. This limits its generalization's capability.

Dragoni *et* al. [7, 6, 8, 9] use *fuzzy logic* to model the relationships between the polarity of concepts and the domain. They used 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 has been the winner of the ESWC 2014 Concept-Level Sentiment Analysis Challenge [7], is that it both accounts for the conceptual representation of the terms in the documents by using WordNet and SenticNet, and proposes one possible solution avoiding to build a new model each time a new domain needs to be analysed.

However, by simply using these resources, some problems remain:

1. it is not possible to discard some of the remaining ambiguities due to the fact that a *synset* (roughly, a concept) corresponds to a group of words (nouns, adjectives, verbs and adverbs) that can be interchangeable and that denote a particular meaning or use—depending on the type of the term used (noun, adjective, verb or adverb), the meaning of the word can change; as an example the term "light" that 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").
2. another disadvantage is that in the implementation of their work, they of course consider each domain as independent of each other, but they use the same stopping criterion for the propagation algorithm for the different domains which could be challenged—the iterative process stops as soon as the sum of the variations in polarity for each concept and domain falls below a fixed threshold, without taking into account the fact that the size of the domains may be very different, so that simultaneous stopping of propagation could be premature for some domains and delayed for others;
3. in *MDFSA*, 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 should be the weights associated with them in the graph.

Here, we propose an extension of the *MDFSA* approach whose

[1] Université Côte d'Azur, CNRS, I3S, France, email: claude.pasquier@univ-cotedazur.fr, celia.da-costa-pereira@univ-cotedazur.fr
[2] Université Côte d'Azur, Inria, CNRS, I3S, France, email: andrea.tettamanzi@univ-cotedazur.fr

aim is to propose solutions for the three above-mentioned problems. For the purpose of (1), deciding 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. Concerning (2), the propagation stopping problem, we propose to 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. Finally, concerning (3), the similarities between the related concepts in the graph, in addition to the synonymy relationships defined in WordNet, we use a pre-trained word embedding model to complete the semantic graph with relationships of closeness between terms.

The results obtained are promising and encouraging. Indeed, when applied to the DRANZIERA dataset [9], with the same evaluation protocol as in [7, 6, 8, 9], the average precision obtained over all 20 domains is 0.7617, which constitutes a significant improvement over *MDFSA-NODK* [8] which obtains a precision of 0.7145 and IRSA-NODK [6] with a precision of 0.6784. An important result is that this improvement in precision score does not come at the expense of the recall value, which is considerably higher than the other methods tested on the same dataset.

We have also trained our method on each of the domains of the DRANZIERA dataset and use the obtained models to predict the orientation (positive or negative) of movie reviews from distinct datasets [27, 17]. Surprisingly, the best precision was obtained when the training was performed with the DRANZIERA reviews belonging to the 'Music' domain. Other domains that lead to good results are 'Books', 'Movies TV' and 'Video games', which are all somehow broadly related to entertainment or culture.

The rest of the paper is structured as follows. Section 2 presents some background with relevant definitions on Fuzzy Set theory. Section 3 describes all the originalities of our approach with a detailed description of the respurces we used. Section 4 presents the fuzzy polarity propagation algorithm in general, and the extensions we have done in particular. Section 5 presents the experiments and discuss the results. Finally, Section 6 concludes the paper.

2 BACKGROUND ON FUZZY SET THEORY

In this section we provide basic definitions and results about fuzzy sets, which will be used in the rest of the article.

2.1 Fuzzy Sets

Fuzzy sets [43], allow the representation of imprecise information. Information is imprecise when the value of the variable to which it refers cannot be completely determined within a given universe of discourse. Fuzzy sets are then a generalization of classical sets obtained by replacing the characteristic function of a set A, χ_A, which takes up values in $\{0, 1\}$ ($\chi_A(x) = 1$ iff $x \in A$, $\chi_A(x) = 0$ otherwise) with a *membership function* μ_A, which can take up any value in $[0, 1]$. The value $\mu_A(x)$ or, more simply, $A(x)$ is the membership degree of element x in A, i.e., the degree to which x belongs in A.

A fuzzy set is completely defined by its membership function. Therefore, it is useful to define a few terms describing various features of this function, summarized in Figure 1. Given a fuzzy set A, its *core* is the (conventional) set of all elements x such that $A(x) = 1$; its *support*, $\text{supp}(A)$, is the set of all x such that $A(x) > 0$. A fuzzy set is *normal* if its core is nonempty. The set of all elements x of A such that $A(x) \geq \alpha$, for a given $\alpha \in (0, 1]$, is called the α-cut of A, denoted A_α.

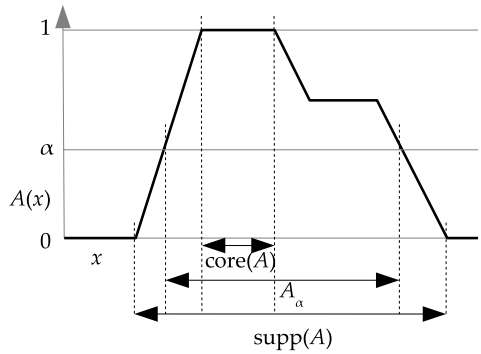

Figure 1: Core, support, and α-cuts of a set A of the real line, having membership function μ_A.

The usual set-theoretic operations of union, intersection, and complement can be defined as a generalization of their counterparts on classical sets by introducing two families of operators, called triangular norms and triangular co-norms. In practice, it is usual to employ the min norm for intersection and the max co-norm for union. Given two fuzzy sets A and B, and an element x,

$$(A \cup B)(x) = \max\{A(x), B(x)\}; \quad (1)$$

$$(A \cap B)(x) = \min\{A(x), B(x)\}; \quad (2)$$

$$\bar{A}(x) = 1 - A(x). \quad (3)$$

2.2 The Extension Principle

The extension principle [44] is the main formal tool for making any mathematical theory fuzzy in a consistent and well-founded way.

Let U be the Cartesian product of n universes $U_1, \ldots, U_n$ and let $A_1, \ldots, A_n$ be an equal number of fuzzy sets defined in $U_1, \ldots, U_n$ respectively.

Suppose $t : U \to V$ is a morphism from U into a new universe V. The question we ask is what the image of a fuzzy subset of U in this new universe V would be under the morphism t. This image would also be a fuzzy set, and its membership function would be calculated from the membership function of the original set and the morphism t.

Let B represent the fuzzy set induced in V by morphism t from the fuzzy sets $A_1, \ldots, A_n$ defined in U. The Extension Principle states that B has membership function, for all $y \in V$,

$$\mu_B(y) = \sup_{(x_1,\ldots,x_n)\in t^{-1}(y)} \min\{\mu_{A_1}(x_1), \ldots, \mu_{A_n}(x_n)\}. \quad (4)$$

B is said to extend fuzzy sets $A_1, \ldots, A_n$ in V.

Equation 4 is expressed for morphisms t of general form. If t is a discrete-valued function, the sup operator can be replaced by the max operator.

2.3 Defuzzification Methods

There may be situations in which the output of a fuzzy inference needs to be a crisp number y^* instead of a fuzzy set R. Defuzzification is the conversion of a fuzzy quantity into a precise quantity.

At least seven methods in the literature are popular for defuzzifying fuzzy outputs [15], which are appropriate for different application contexts. The *centroid method* is the most prominent and physically

appealing of all the defuzzification methods. It results in a crisp value

$$y^* = \frac{\int y\mu_R(y)dy}{\int \mu_R(y)dy}, \quad (5)$$

where the integration can be replaced by summation in discrete cases.

3 MATERIAL

Our method is based on fuzzy set theory (cf. Section 2) and exploits background knowledge about concepts. This background knowledge is represented by a semantic graph composed of vertices, which represent either concepts or terms, and edges, which represent semantic relations. In our approach, 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 groups of words).

The backbone of our semantic graph is based on WordNet [22], an online lexical database in which nouns, verbs, adjectives, and adverbs are organized into sets of synonyms (*synsets*), each representing a lexicalized concept. In this database, all synsets are connected to other synsets by means of semantic relationships. In our work, we use the relationships of synonymy, antonymy and hypernymy.

WordNet is built around the notion of concept, but a written text is composed of words, not concepts, and words having several distinct meanings (which is the norm for the most frequent words [5]) are represented as many distinct synsets. It is therefore necessary to undergo a preprocessing phase in order to link the words found in the texts as accurately as possible to their corresponding synsets. The first approach we propose here to reduce ambiguity is to parse the text in order to take into account the part of speech (POS) of the tokens when associating the corresponding synset. Indeed, many terms, taken independently, can belong to different grammatical types (for example the word "course" that, depending on the context, can be a verb, a noun or an adverb). Using the POS of terms enables a more reliable association with synsets.

Another approach, complementary to the use of POS tags, is the integration with other public resources.

In our method, we combine WordNet with SenticNet [4], a publicly available resource for opinion mining. The latest version (SenticNet5) covers 100,000 common-sense concepts, which are assigned values corresponding to various characteristics (polarity, pleasantness, attention, sensitivity, and aptitude) and are semantically linked to other concepts. The use of SenticNet allows, on the one hand, to extend the coverage of WordNet by allowing synsets to be assigned to terms or groups of terms that are not indexed in WordNet and, on the other hand, to resolve a number of cases of ambiguity. The methodology employed to link WordNet and SenticNet entries is identical to that used by Dragoni *et al* [8]. It should be noted here that SenticNet is only used to extend WordNet coverage and to increase the accuracy of associations between terms and synsets. The polarities defined by SenticNet, which represent typical values for each term, 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 property that depends on the relation between a term and a domain.

In addition to the synonymy relationships defined in WordNet, already used in the literature, we use a dataset obtained by word embedding to complete the semantic graph with relationships of closeness (or similarity) between terms. The dataset that we use is a pretrained model computed by applying Word2vec [21] on roughly 100 billion words from a Google News dataset.[3] This dataset allows to obtain the proximity between each pair of terms by calculating the cosine similarity between their vector representations. In the final semantic network, the distance between terms in the vector space is used to link each term to its five closest terms. These edges are weighted with the value of the cosine similarity between the terms, whereas relationships extracted from WordNet are weighted with 1 (for synonymy and hypernymy relationships) or -1 (for antonymy).

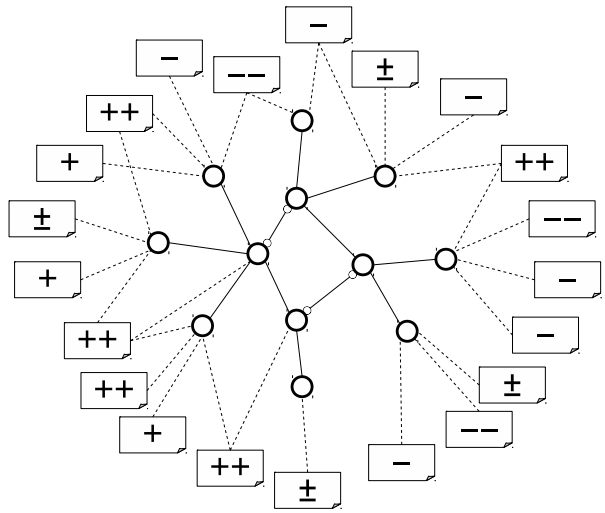

Figure 2: An illustration of the semantic graph constructed by the proposed method. Circles represent the vertices of the graph and solid lines its edges. The documents of the training set are shown around the semantic graph. Dashed lines represent the occurrence, in a document, of a term associated 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).

4 METHOD

The algorithm used to learn concept polarities for various domains is an extension of the one described in [8]. It consists of three phases, which are:

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;
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 2 illustrates the idea of a semantic graph, whose construction and use is detailed below.

4.1 Semantic Graph Construction

The semantic graph is constructed as a graph (V, E). Each element of V is either a concept (in our case a synset defined in WordNet) or the canonical form (lemma) of a term used in the description of the reviews.

The edges in E are created:

- between pairs of synsets linked by an `is-a` relationship in WordNet, with weight $+1$;

[3] the dataset is downloadable from https://frama.link/google_word2vec

- between lemmas associated to WordNet synsets and linked by an `synonym` relationship, with weight 1;
- between lemmas associated to WordNet synsets and linked by an `antonym` relationship, with weight -1;
- between each lemma and the five closest lemmas according to the pretrained *word2vec* model, with their cosine similarity as weight.

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

4.2 Concept Polarity Initialization

The initial polarities $\vec{p}^{(0)}(v)$ of all the vertices of the semantic graph are computed, for each domain i, as

$$p_i^{(0)}(v) = \bar{p}_i(v) \in [-1, 1], \tag{6}$$

where $\bar{p}_i(v)$ is 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 of v occurs in a document of domain i, $p_i^{(0)}(v) = 0$.

Here, unlike in the literature, for the purpose of deciding 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.

4.3 Polarity Propagation

In this phase, information about the polarity of vertices is propagated through the edges of the graph, so that concepts for which no polarity information could be directtly 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 $t = 1, 2, \ldots$, the polarity $p_i^{(t)}(v)$ of each vertex v for domain i is updated based on the values of its neighbors $N(v) = \{v_j \mid (v, v_j) \in E\}$ as follows:

$$\vec{p}^{(t+1)}(v) = (1 - \lambda)\vec{p}^{(t)}(v) + \lambda \frac{1}{\|N(v)\|} \sum_{v' \in N(v)} \vec{p}^{(t)}(v'), \tag{7}$$

where $0 < \lambda < 1$ is the *propagation rate*, a parameter 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 the principle of simulated annealing [18], 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:

$$\lambda_t = A\lambda_{t-1}. \tag{8}$$

In the method proposed by Dragoni *et* al. [8], 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

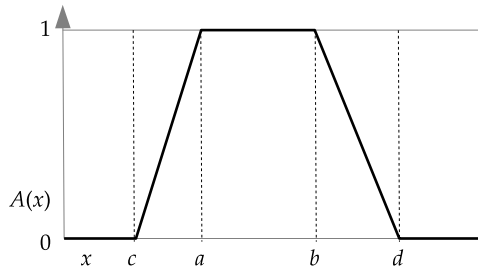

Figure 3: A fuzzy set with a trapezoidal membership function, defined by the four parameters (a, b, c, d).

to each domain separately and that is relative to the number of different nodes composing the semantic graph of the domain. Thus, for a domain i, the polarity propagation stops when the average variation in term polarity,

$$\Delta_i^{(t)} = \frac{1}{\|V\|} \sum_v \left| p_i^{(t)}(v) - p_i^{(t-1)}(v) \right|, \tag{9}$$

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

At each iteration $t = 0, 1, 2, \ldots$, 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 $p_i^{(t)}(v)$, $t = 0, \ldots, t_i^{\text{stop}}$. 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 3, 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.

4.4 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 domain i is computed as the average of the fuzzy polarities (represented by their trapezoidal membership functions) of all the synsets v whose terms occur in the document:

$$\mu_i = \frac{1}{\|V_i\|} \sum_{v \in V_i} \mu_{v,i}, \tag{10}$$

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

$$\mu_i(x) = \sup_{x=\frac{1}{\|V_i\|}\sum_{v\in V_i} x_v} \min_{v\in} \mu_{v,i}(x_v). \quad (11)$$

However, given that all $\mu_{v,i}$ are trapezoidal with parameters (a_v, b_v, c_v, d_v), as pointed out in [8], μ_i will always be trapezoidal as well, with parameters

$$\frac{1}{\|V_i\|}\left(\sum_{v\in V_i} a_v, \sum_{v\in V_i} b_v, \sum_{v\in V_i} c_v, \sum_{v\in V_i} d_v\right). \quad (12)$$

This fuzzy polarity reflects the uncertainty of the estimate obtained by the model. If a single polarity figure is needed for the application at hand, that can be obtained by applying a defuzzification method, like the centroid method of Equation 5. This is, at least, what we did for the empirical validation of our method, presented in Section 5.

5 EXPERIMENTS AND RESULTS

The proposed system has been evaluated using the DRANZIERA evaluation protocol [9], 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 five thousand positive and five thousand negative reviews that are split in five folds containing one thousand positive and one thousand negative reviews each.

5.1 Experimental Protocol

As suggested by the guideline of the DRANZIERA evaluation protocol [9], 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.

The algorithm depends on three different parameters: the *propagation rate* λ, which determines the diffusion rate of the polarity values between concepts, 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 8). Using a small portion of the dataset, we have therefore experimented 297 different configurations of parameters (λ, 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 and a *convergence limit* of 0.01 lead to the best results.

5.2 Results of DRANZIERA Evaluation

When applied to the DRANZIERA dataset with the settings previously identified, namely $\lambda = 0.3$, $L = 0.01$, and $A = 0.5$, the average precision obtained over all 20 domains is 0.7617, which constitutes a significant improvement over MDFSA-NODK [8], which obtains a precision of 0.7145, and IRSA-NODK [6] with a precision of 0.6784. 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 6.8% improvement with respect to 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 [9].

Table 1: Average precision and recall obtained on the 20 domains of the DRANZIERA dataset.

Method	Precision	Recall	F1 score
MDFSA [8]	0.6832	0.9245	0.7857
MDFSA-NODK [9]	0.7145	0.9245	0.8060
IRSA [6]	0.6598	0.8742	0.7520
IRSA-NODK [9]	0.6784	0.8742	0.7640
Our Method	**0.7617**	**0.9947**	**0.8627**

The breakdown of the results obtained on the 20 domains of the DRANZIERA dataset are listed in Table 2.

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

Domain	precision	recall	F1 score
Music	0.8123	0.9918	0.8931
Books	0.6789	0.994	0.8068
Video Games	0.7993	0.9907	0.8848
Movies TV	0.684	0.9954	0.8108
Software	0.7344	0.9948	0.8450
Amazon Instant Video	0.6131	0.9974	0.7594
Electronics	0.8166	0.9949	0.8970
Beauty	0.8662	0.9947	0.9260
Toys Games	0.7746	0.9947	0.8710
Sports Outdoors	0.8022	0.9951	0.8883
Health	0.7874	0.993	0.8783
Office Products	0.7548	0.9965	0.8590
Patio	0.7611	0.9942	0.8622
Tools Home Improvement	0.7704	0.9932	0.8677
Pet Supplies	0.7361	0.9945	0.8460
Clothing Accessories	0.8302	0.998	0.9064
Shoes	0.8797	0.9961	0.9343
Automotive	0.7364	0.9951	0.8464
Home Kitchen	0.7274	0.9947	0.8403
Baby	0.6687	0.9955	0.8000
Average	0.7617	0.9947	0.8627

5.3 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.

Ribeiro *et* al. [32] benchmarked 24 sentiment analysis methods on 18 datasets. Most of these datasets contain messages posted on various channels (discussion forums, Twitter). However, two datasets do represent movie reviews that could benefit from being processed with our method (although the DRANZIERA dataset does not contain the

'movies' 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).

We chose to use the dataset proposed by Hutto *et* al. [17], a reanalysis of Pang and Lee's dataset [27] by 20 independent human raters which can be considered as gold-standard quality.

Our method, trained on each of the domains of the DRANZIERA dataset was used to predict the orientation of each movie review between positive and negative. The best precision was obtained when the training was performed with the DRANZIERA reviews belonging to the 'Music' domain. The other domains for which the transfer of the learned model is effective are 'Books', 'Movies TV' and 'Video games'.

Although these are not exactly the categories for which one would have expected to obtain the best cross-domain transfer quality, the result is still consistent. 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', 'Automotive', 'Home Kitchen', 'Baby'). Table 3 lists the precision obtained on both datasets according to the category of the DRANZIERA dataset used for training. Recall values are not displayed because the variation is small; they range from 0.9901 to 0.9965.

Table 3: Precision of cross-domain transfer from the 20 categories of the DRANZIERA dataset to gold-standard quality movie reviews.

Domain	expert-rated movies [17]
Music	0.7003
Books	0.6658
Video Games	0.6622
Movies TV	0.6634
Software	0.6456
Amazon Instant Video	0.6384
Electronics	0.6116
Beauty	0.6075
Toys Games	0.6035
Sports Outdoors	0.6012
Health	0.6014
Office Products	0.5939
Patio	0.5932
Tools Home Improvement	0.5873
Pet Supplies	0.5885
Clothing Accessories	0.5875
Shoes	0.5853
Automotive	0.5788
Home Kitchen	0.5623
Baby	0.5524

We can notice, in Table 3, that the domain used for training has a great influence on the results. The precision therefore varies by nearly 0.15 between a transfer done from the 'Music' domain and a transfer from the 'Baby' domain. However, the difference can also be partly explained by the intrinsic score obtained on each DRANZIERA domain. We can indeed notice in Table 2, that the difference in precision between the 'Music' domain and the 'Baby' domain is slightly more than 0.14. Overall, there is a decrease in accuracy when the method, trained on a DRANZIERA domain, is applied to an independent dataset; which seems perfectly logical. 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.

Considering precision, our method is only outperformed by 4 of the 24 methods tested by Ribeiro *et* al.: SenticNet [3], Stanford recursive deep model [35], Sentiment140 [11] and SO-CAL [36] that obtain a precision of 0.9630, 0.8270, 0.7308 and 0.7165 respectively. The best method, consisting in using only the polarity reported in SenticNet, obtains a high precision but only on a relatively small part of the reviews since its coverage is 0.6941. The second and fourth best methods, Stanford DM and SO-CAL, have a coverage of the same order, 0.9192 and 0.8910 respectively while Sentiment140 is able to attribute a rating to only 18.67% of the reviews. The very good recall of 0.9922 obtained by our method is only exceeded by Emoticons DS. However, the high recall of 0.9979 scored by Emoticon DS is associated with a modest accuracy of 0.5027.

When recall is considered together with precision (cf. Table 4), our method obtains an F1 score of 0.8223, which is higher than the F1 score of the most precise method, 0.8067, but short of Stanford DM, which has the highest F1 score with 0.8707, while SO-CAL is lower, at 0.7943, and Sentiment140 is far behind with an F1 score of only 0.2974.

Table 4: A comparison of our method (trained on the Music domain of DRANZIERA) with all the methods benchmarked by Ribeiro *et* al. [32], when applied to the user-rated movies dataset. The highest score of each column is highlited in boldface

Method	Precision	Recall	F1 score
AFINN [26]	0.6593	0.7259	0.6910
ANEW SUB [40]	0.5680	0.9630	0.7145
Emolex [25]	0.6477	0.7439	0.6925
Emoticons [12]	0.6000	0.0005	0.0010
Emoticons DS [14]	0.5027	**0.9979**	0.6686
NRC Hashtag [23]	0.6234	0.9347	0.7480
LIWC07 [37]	0.6300	0.6608	0.6450
LIWC15 [29]	0.6335	0.6608	0.6469
Opinion Finder [41]	0.2655	0.4912	0.3447
Opinion Lexicon [16]	0.6977	0.7728	0.7333
PANAS-t [13]	0.6630	0.0340	0.0647
Pattern.en [34]	0.6784	0.6559	0.6670
SANN [28]	0.6181	0.6176	0.6178
SASA [39]	0.5741	0.5824	0.5782
Semantria [20]	0.6964	0.6880	0.6922
SenticNet [3]	**0.9630**	0.6941	0.8067
Sentiment140 [11]	0.7308	0.1867	0.2974
Sentiment140 L [24]	0.6318	0.9351	0.7541
SentiStrength [38]	0.6754	0.2698	0.3856
SentiWordNet [2]	0.6145	0.6253	0.6199
SO-CAL [36]	0.7165	0.8910	0.7943
Stanford DM [35]	0.8270	0.9192	**0.8707**
Umigon [19]	0.6344	0.5395	0.5831
VADER [17]	0.6519	0.8270	0.7291
Our Method	0.7021	0.9922	0.8223

Overall, our method is characterized by its excellent coverage. Indeed, this one is always above 0.99, which means that less than 1% of the reviews are not classified.

6 CONCLUSION

We have proposed solutions to solve three remaining problems of the *MDFSA* approach. The first solution allows to eliminate some ambiguities arising when associating different types of words (nouns,

adjectives, verbs, adverbs) with their synset, thus allowing to propagate the polarity of the document where they occur to the correct vertex of the semantic graph. The second solution that we have proposed consists of making the "propagation stopping time" dependent on the domain size—the propagation phase depends on the number of nodes composing the semantic graph of the domain—, unlike in the *MDFSA* approach, where the propagation phase stops at the same time for all the domains. This solution allows to save computing cycles for those domains whose polarities converge fast, while guaranteeing that the polarities of all the domains converge. Finally, we improved the construction of the semantic graph by adding weighted edges connecting each vertex to the vertices of its five most similar terms according to a pre-trained word embedding model, thus favoring the propagation of polarities between semantically similar concepts and terms.

The resulting approach has been validated using a standard evaluation protocol. The results of the evaluation show 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.

Our results seem to confirm that making the construction of the semantic graph more accurate by disambiguating the terms occurring in documents and connecting semantically similar vertices helps to improve both the precision and the coverage of the approach.

The extensions we have proposed solve, if only partially, some known issues of the approach, but we know there is still much room for improvement by injecting more linguistic knowledge, which is, therefore, the main direction for future work. That might include using a dependency parser to analyze the texts and applying graph embedding techniques to the dependency graph of the sentences where the terms occur. Another possibility would be to use a more sophisticated disambiguation method than just looking at the POS tags of the terms.

ACKNOWLEDGEMENTS

Célia da Costa Pereira acknowledges support of the PEPS AIRINFO project funded by the CNRS.

Andrea Tettamanzi has been supported by the French government, through the 3IA Côte d'Azur "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

REFERENCES

[1] Nor Aniza Abdullah, Ali Feizollah, Ainin Sulaiman, and Nor Badrul Anuar, 'Challenges and recommended solutions in multi-source and multi-domain sentiment analysis', *IEEE Access*, **7**, 144957–144971, (2019).

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, 'Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining', in *LREC*, pp. 2200–2204, (2010).

[3] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok, 'Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings', in *AAAI*, pp. 1795–1802. AAAI Press, (2018).

[4] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain, 'Senticnet: A publicly available semantic resource for opinion mining', in *AAAI Fall Symposium: Commonsense Knowledge*, (2010).

[5] Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i Cancho, and Jaume Baixeries, 'Polysemy and brevity versus frequency in language', *Computer Speech & Language*, **58**, 19–50, (2019).

[6] Mauro Dragoni, 'Shellfbk: an information retrieval-based system for multi-domain sentiment analysis', in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 502–509, (2015).

[7] Mauro Dragoni, Andrea G. B. Tettamanzi, and Célia da Costa Pereira, 'A fuzzy system for concept-level sentiment analysis', in *SemWebEval@ESWC*, volume 475 of *Communications in Computer and Information Science*, pp. 21–27. Springer, (2014).

[8] Mauro Dragoni, Andrea G. B. Tettamanzi, and Célia da Costa Pereira, 'Propagating and aggregating fuzzy polarities for concept-level sentiment analysis', *Cognitive Computation*, **7**(2), 186–197, (2015).

[9] Mauro Dragoni, Andrea G. B. Tettamanzi, and Célia da Costa Pereira, '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*. European Language Resources Association (ELRA), (2016).

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, 'Domain adaptation for large-scale sentiment classification: A deep learning approach', in *ICML*, pp. 513–520. Omnipress, (2011).

[11] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Standford University, 2009.

[12] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha, 'Comparing and combining sentiment analysis methods', in *Proceedings of the first ACM conference on Online social networks*, pp. 27–38, (2013).

[13] Pollyanna Gonçalves, Fabrício Benevenuto, and Meeyoung Cha, 'Panas-t: A pychometric scale for measuring sentiments on twitter', *arXiv preprint arXiv:1308.1857*, (2013).

[14] Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald, 'Tweetin'in the rain: Exploring societal-scale effects of weather on mood', in *Sixth International AAAI Conference on Weblogs and Social Media*, (2012).

[15] Hans Hellendoorn and Christoph Thomas, 'Defuzzification in fuzzy controllers', *Intelligent and Fuzzy Systems*, **1**, 109–123, (1993).

[16] Minqing Hu and Bing Liu, 'Mining and summarizing customer reviews', in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, (2004).

[17] Clayton J Hutto and Eric Gilbert, 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', in *Eighth international AAAI conference on weblogs and social media*, (2014).

[18] Scott Kirkpatrick, D. Gelatt Jr., and Mario P. Vecchi, 'Optimization by simulated annealing', *SCIENCE*, **220**(4598), 671–680, (1983).

[19] Clement Levallois, 'Umigon: sentiment analysis for tweets based on lexicons and heuristics', *Proceedings of the International workshop on Semantic Evaluation*, (2013).

[20] Lexalytics, 'Sentiment extraction - measuring the emotional tone of content', *Technical Report*, (2015).

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, (2013).

[22] George A Miller, 'Wordnet: a lexical database for english', *Communications of the ACM*, **38**(11), 39–41, (1995).

[23] Saif M. Mohammad, '# emotional tweets', in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 246–255. Association for Computational Linguistics, (2012).

[24] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu, 'Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets', *arXiv preprint arXiv:1308.6242*, (2013).

[25] Saif M. Mohammad and Peter D. Turney, 'Crowdsourcing a word–emotion association lexicon', *Computational Intelligence*, **29**(3), 436–465, (2013).

[26] Finn Årup Nielsen, 'A new anew: Evaluation of a word list for sentiment analysis in microblogs', *arXiv preprint arXiv:1103.2903*, (2011).

[27] Bo Pang and Lillian Lee, 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp. 271–278. Association for Computational Linguistics, (2004).

[28] Nikolaos Pappas and Andrei Popescu-Belis, 'Sentiment analysis of user comments for one-class collaborative filtering over ted talks', in *Proceedings of the 36th international ACM SIGIR conference on Research*

and development in information retrieval, pp. 773–776, (2013).
[29] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn, 'The development and psychometric properties of liwc2015', Technical report, (2015).
[30] Mironela Pirnau, 'Sentiment analysis for the tweets that contain the word "earthquake"', in *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–6, (June 2018).
[31] Andi Rexha, Mark Kröll, Mauro Dragoni, and Roman Kern, 'The CLAUSY system at ESWC-2018 challenge on semantic sentiment analysis', in *SemWebEval@ESWC*, volume 927 of *Communications in Computer and Information Science*, pp. 186–196. Springer, (2018).
[32] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto, 'Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods', *EPJ Data Science*, **5**(1), 23, (2016).
[33] Kim Schouten and Flavius Frasincar, 'The benefit of concept-based features for sentiment analysis', in *SemWebEval@ESWC*, volume 548 of *Communications in Computer and Information Science*, pp. 223–233. Springer, (2015).
[34] Tom De Smedt and Walter Daelemans, 'Pattern for python', *Journal of Machine Learning Research*, **13**(Jun), 2063–2067, (2012).
[35] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts, 'Recursive deep models for semantic compositionality over a sentiment treebank', in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, (2013).
[36] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede, 'Lexicon-based methods for sentiment analysis', *Computational Linguistics*, **37**(2), 267–307, (2011).
[37] Yla R. Tausczik and James W. Pennebaker, 'The psychological meaning of words: Liwc and computerized text analysis methods', *Journal of language and social psychology*, **29**(1), 24–54, (2010).
[38] Mike Thelwall, 'Heart and soul: Sentiment strength detection in the social web with sentistrength (summary book chapter)', *Cyberemotions: Collective emotions in cyberspace. Berlin, Germany: Springer*, 119–134, (2013).
[39] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan, 'A system for real-time twitter sentiment analysis of 2012 us presidential election cycle', in *Proceedings of the ACL 2012 system demonstrations*, pp. 115–120. Association for Computational Linguistics, (2012).
[40] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert, 'Norms of valence, arousal, and dominance for 13,915 english lemmas', *Behavior research methods*, **45**(4), 1191–1207, (2013).
[41] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 'Recognizing contextual polarity in phrase-level sentiment analysis', in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 347–354, (2005).
[42] Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto, 'Transfer learning for multiple-domain sentiment analysis—identifying domain dependent/independent word polarity', in *AAAI*, pp. 1286–1291, (2011).
[43] Lotfi A. Zadeh, 'Fuzzy sets', *Information and Control*, **8**, 338–353, (1965).
[44] Lotfi A. Zadeh, 'The concept of a linguistic variable and its application to approximate reasoning - i', *Inf. Sci.*, **8**(3), 199–249, (1975).