Extending General Sentiment Lexicon to Specific Domains in (Semi-)Automatic Manner

Pavel Brazdil[1,2], Purificação Silvano[3], Fátima Silva[3], Shamsuddeen Muhammad[2,4], Fátima Oliveira[3], João Cordeiro[2,5] and António Leal[3]

[1]FEP, University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
[2]INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[3]FLUP/CLUP, University of Porto, Via Panorâmica, s/n, 4150-564 Porto, Portugal
[4]FCUP, University of Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal
[5]FE/HULTIG, University of Beira Interior, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal

Abstract

This paper describes an approach to the construction of a sentiment analysis system that uses both automatic and manual processes. The system includes a domain-specific sentiment lexicon, modifier patterns and rules that are used to derive the sentiment values of sentences in new texts. The lexicon that includes single words (unigrams) is obtained in an automatic manner from the distribution of ratings for all words in the labelled training data. The sentiment values of phrases is derived from a list of modifier patterns, built/developed manually. These include a modifier and a focal element. The modifiers can be of different types, depending on whether the operation is intensification, downtoning or reversal. This approach was applied to texts on economics and finance in European Portuguese. In our view, this line of work deserves more attention in the community, as the system not only has reasonable performance, but also can provide understandable explanations to the user.

Keywords

sentiment analysis, automatic lexicon generation, domain-specific lexicon, corpus annotation

1. Introduction

The aim of our work is to describe a methodology that has been used to develop a domain-sensitive sentiment lexicon for European Portuguese (EP), which, contrary to English, lacks a rich variety of Natural Language Processing (NLP) tools at its disposal, namely for sentiment analysis. In this study, we focus on the domain of economics and finance. Although deep neural networks have achieved good performance in many tasks, including sentiment analysis, we follow the line that involves sentiment lexicons, as it is relatively easy to explain to the user why

SALLD-1: Workshop on Sentiment Analysis & Linguistic Linked Data - September 1, 2021 - Zaragoza, Spain
pbrazdil@inescporto.pt (P. Brazdil); msilvano@letras.up.pt (P. Silvano); mhenri@letras.up.pt (F. Silva); shmuhammad.csc@buk.edu.ng (S. Muhammad); moliv@letras.up.pt (F. Oliveira); jpcc@ubi.pt (J. Cordeiro); jleal@letras.up.pt (A. Leal)
http://www.liaad.up.pt/area/pbrazdil/pavel-brazdil (P. Brazdil); https://www.purisilvano.pt (P. Silvano); https://www.shmuhammad.com (S. Muhammad); https://www.di.ubi.pt/~jpaulo/ (J. Cordeiro)
0000-0002-4720-0486 (P. Brazdil); 0000-0001-8057-5338 (P. Silvano); 0000-0003-2360-5136 (F. Silva); 0000-0001-7708-0799 (S. Muhammad); 0000-0003-2110-1049 (F. Oliveira); 0000-0003-0466-1618 (J. Cordeiro); 0000-0002-6198-2496 (A. Leal)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

a certain polarity was attributed to the given text. So, how can we generate a domain-specific lexicon for EP in this setting?

Developing a domain-sensitive sentiment lexicon manually was excluded from consideration, as it requires a great deal of manual effort. Each domain-specific word, multiword or short phrase needs to be identified manually and added to the existing lexicon.

So, a question arises whether this process can be automated. Almatarneh et al. [1] have shown a way to do this, provided a set of labelled texts is available, as in supervised learning, where the labels represent the sentiment values attributed by experts to these texts. Muhammad et al. [2] improved this approach by processing all tokens in the labelled texts and, for each one, by constructing a distribution of ratings. This distribution was used to infer the sentiment value of the given token. Tokens with values different from zero are added to the sentiment lexicon. This approach is used as the basis for the method used here.

The quality of the induced lexicon depends on the number of labelled texts available. As labelling is done by human experts, the amount of labelled data is normally limited. So, because of this limitation, we restrict the automatic approach to single words (unigrams). Phrases, such as "crescimento alto" (high growth), are acquired in another way, with recourse to modifier patterns that include contextual shifters or modifiers of the sentiment value of a particular word or expression. The modifiers can be of different types, depending on whether the operation involved is intensification, downtoning/attenuation or reversal [3]. All these patterns are identified manually from the labelled texts. In addition to this, we also consider other similar variants (synonyms or paraphrases) that may occur in similar situations. The sentiment value of the pattern is obtained in an automatic way, by applying rules, as discussed in Section 5.

Many of these patterns are not specific to a particular domain, but rather are quite general and applicable across domains and also languages. So, these patterns constitute a linguistic knowledge, which is generally useful and could be shared among different groups, as they are not specific to a particular domain.

Therefore, the aim of this paper is to describe an approach to the construction and use of domain-specific lexicon that uses both automatic and manual processes. We show the results of this approach on texts from the domain of economics and finance.

The rest of this article is organized as follows. The next section discusses related work. Section 3 describes the method of automatic generation of the sentiment lexicon. Section 4 discusses the experimental results obtained with this approach. Section 5 is dedicated to modifier patterns. It starts by explaining that they include a modifier and a focal element and then presents different types of modifiers (e.g., intensifiers, etc.) and the rules that are used to derive the sentiment value of specific patterns. Section 5.5 presents the results obtained with this approach. Section 6 includes the conclusions.

2. Related Work

The area of sentiment analysis has attracted a lot of attention in the past, and consequently, many approaches exist. One recent paper of Atteweldt et al. [4] compares manual annotation, crowd-coding, dictionary approaches and machine learning (ML). This work is relevant to the research discussed here, as it focuses on a similar subject area covered in this paper, namely

economics (more precisely economic headlines) in a language with limited language resources (Dutch), which in this respect can be compared to Portuguese. The conclusion of the authors is that off-the-shelf dictionaries do not perform well on new tasks, when compared to machine learning (ML) methods, particularly deep learning methods, such as, Convolutional Neural Networks (CNN). Although deep learning techniques often provide well performing solutions, it is not easy to see why a certain prediction was made and what it is based on. This motivated us to explore automated techniques in the process of lexicon construction.

Lexicon-based approaches can be divided into basically different groups, depending on whether the lexicon construction is manual, semi-automatic or automatic. Manual approach was used by Forte et al. [5], who analyzed users' commentaries in Portuguese concerning certain products and services. They used an existing sentiment lexicon as the basis for further domain-specific extensions generated manually, which led to a marked improvement of predictive performance. Silva et al. [6] followed a similar approach in a study oriented towards texts in Portuguese in the area of economics and finance. The downside of such approaches is that the domain-specific extensions require a great deal of manual effort. The following example illustrates this.

> *A dívida ingerível, o crescimento anémico, os impostos altos, o investimento nulo, a regulação ineficaz, a separação dos portugueses entre protegidos e excluídos exigem mais do que um simples "virar de página" da austeridade. (The unmanageable debt, anaemic growth, high taxes, zero investment, inefficient regulation, the separation of the Portuguese between the protected and the excluded demand more than a simple "turning of the page" on austerity.)*

The lexicon SentiEcon [7] is relevant to our work, as it is a domain-specific lexicon for sentiment analysis applications in English. It contains 6,470 entries (single words and multiwords) annotated with semantic orientation and intensity. This sentiment lexicon is intended for use in the financial/economic domain in conjunction with a general sentiment lexicon. Also, a similar strategy on Portuguese was followed in the work of Silva et al. [6].

Some semi-automatic approaches start with a relatively small lexicon that contains certain important seed words that can be provided manually, particularly when trying to construct a domain-specific lexicon. The process of label propagation is used to transfer the values to other domain-specific terms, identified as synonyms. Various methods exist for that, including, for instance, synsets of Wordnet or domain-sensitive embeddings [8, 9].

Various authors investigated automatic approaches for lexicon construction. In the work of Almatarneh et al. [1], the words were classified into three categories - negative, neutral and positive, depending on the rating. Sentiment values are inferred from the frequencies in the negative and positive category. This approach has the disadvantage of considering all ratings in the negative (or positive) category as equivalent. In other words, it does not exploit the valence within the positive/negative category. The approach described in this paper improves this shortcoming.

One recent work related to our work is oriented towards sentiment classification of economic text in Portuguese [10]. Their data involves 400 manually annotated sentences on economics extracted from Portuguese newspapers. Although they have performed experiments with different ML approaches, the best performance was obtained with rules generated manually.

3. Automatic generation of sentiment lexicon with single words

The methodology adopted involves the following steps:

1. Corpus annotation
2. Carrying out pre-processing
3. Generating the distributions of probabilities of occurrence of words
4. Using the distribution of probabilities to generate the sentiment values
5. Rescaling the sentiment values for a specific task

More details on some of the above steps follow.

3.1. Corpus annotation

The group of four linguists prepared 23 texts on the topic of finance and economy from different online newspapers. Each text contained a certain number of sentences (or phrases) varying from 2 to 30. The total number of sentences was 408. At least two annotators annotated each sentence (or phrase) with the sentiment value on the scale of -3 to 3. In cases when the value diverged, the other two annotators were called upon to establish the agreement. The following table shows examples of some of the sentences/phrases used.

S/N	Sentences	Polar
1	Primeiro porque quem está habituado a lidar com a exportação de serviços sabe que a falta de qualificação dos portugueses é uma falsa questão.	1
2	Porque não só o trabalho dos portugueses se vende como nunca, como também diversas mega empresas europeias estão a mudar para cá os seus serviços mais sofisticados.	2
3	O saldo positivo das nossas trocas compensa largamente o financiamento das atividades do país.	2
4	Os números da economia portuguesa (trabalhados a partir de dados do Pordata), na parte em que interessa, são bastante lisonjeiros.	2
5	Este último número é muito interessante pelo facto de ser próximo da inflação na zona Euro, o que significa que, em termos de valor, as importações pouco se alteraram no número global.	1

Table 1
Examples of some of the sentences and rating.

The distribution of ratings and counts is highly skewed towards negative values. The corresponding distribution is shown in Table 2.

Rating	-3	-2	-1	1	2	3	Total
Number of occurrences	21	100	147	81	52	7	408

Table 2
Distribution of ratings for our data

3.2. Preprocessing

The data is read-in using R language. This process generates a data frame. The column containing the text is passed to `udpipe` package to extract lemmas. Further processing is done with `tm` package of R. Previous studies have shown that not all lexical categories include sentiment bearing words (e.g., [11, 12, 13]). Our experiments have indicated that the most useful categories to consider are nouns, verbs, adjectives, and adverbs. Therefore, we focused on these four categories only when considering unigrams.

3.3. Generating the distributions of probabilities of occurrence of words for ratings

In this process, we consider each term that appears in the given texts (e.g., term ti = "good") and examine its number of occurrences for different ratings in vector $W = (-3, -2, -1, 0, +1, +2, +3)$. So, let $N_{ti,rj}$ represent the number of occurrences of term ti in the texts with rating rj. Hence, the estimate probability of occurrence of term ti in text with rating rj could be estimated by ratio

$$p_{ti,rj} = \frac{N_{ti,rj}}{N_{ti,*}} \tag{1}$$

where $N_{ti,*}$ represents the total number of occurrences of term ti across all ratings. Let P_{ti} represent the vector of estimates of probabilities of occurrence associated with different ratings.

Words that are relatively rare could have many values $N_{ti,rj}$ equal to 0. As we know, estimates carried out on small samples are rather unreliable. To avoid this problem, we use a kind of smoothing technique, which is based on the notion of Laplace smoothing. To calculate the estimate of probability of occurrence of term ti in text with rating rj, we use

$$p'_{ti,rj} = \frac{N_{ti,rj} + C_1}{N_{ti,*} + C_2} \tag{2}$$

where C_1 and C_2 are constants. In this study we use $C_1 = 1$ and $C_2 = 2$. Let P'_{ti} represent the vector of individual values. To guarantee that all estimates sum to 1, we divide the individual values by the total

$$p''_{ti,rj} = \frac{p'_{ti,rj}}{\sum p'_{ti,rj}} \tag{3}$$

To calculate the sentiment value (SV_{ti}) of a particular term ti, we use

$$SV_{ti} = \sum p''_{ti} * W \tag{4}$$

which represents a sum of the inner product of the two vectors (p''_{ti} and W). For **example**, let us see how our method can be applied to calculate the sentiment value of the term "bom" (good) on the basis of one preliminary study. Table 3 shows the following entities:

- $N_{bom,rj}$, representing the number of occurrences of the word bom in the given sentences/phrases with a particular rating,
- ratio $p_{bom,rj}$,

- adjusted value $p'_{bom,rj}$ using Laplace correction, which increases small values of $p'_{bom,rj}$ and decreases large values,
- adjusted estimate of probability $p''_{bom,rj}$,
- product $p''_{bom,rj} * W$ used in the calculation of the sentiment value.

Rating rj	$N_{bom,rj}$	$p_{bom,rj}$		$p'_{bom,rj}$		$p''_{bom,rj}$	$p''_{bom,rj} * W$
-3	0	0/38	0.000	1/40	0.025	0.023	-0.068
-2	0	0/38	0.000	1/40	0.025	0.023	-0.045
-1	6	6/38	0.158	7/40	0.175	0.159	-0.159
+1	20	20/38	0.526	21/40	0.525	0.477	0.477
+2	10	10/38	0.263	11/40	0.275	0.250	0.500
+3	2	2/38	0.053	3/40	0.075	0.068	0.205
Total	38		1.000		1.100	1.000	**0.909**

Table 3
Distribution of occurrences of "bom" across different ratings.

Figure 1 accompanies Table 3 and shows the distribution of the value $N_{bom,rj}$. As can be seen, the distribution is highly skewed towards positive ratings.

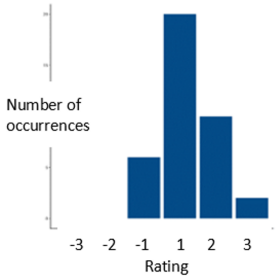

Figure 1: Distribution of the value $N_{bom,rj}$ for the word "bom".

3.4. Using the distribution of probabilities to generate the sentiment values

The estimates of probabilities can be used to calculate the sentiment value of the word "bom" (good). As was shown before, this is done by calculating the inner product of the vector of estimates of probabilities P''_{bom} and the vector of weights W. Figure 2 shows the result of this operation. The final sentiment value of the word "bom" is calculated as the sum of the individual contributions. The final sentiment value of this word is 0.907. A part of the lexicon induced in our preliminary study from the data (Ecolex) is shown in Table 4. Table 5 shows the distribution for different lexical classes.

The words that have a clear positive sentiment are highlighted in bold (e.g., *excelente*, *bom*). Some of the words in this sample are related to economics and finance (e.g., *investimento*, *crescer*), while others seem to be rather spurious entries (e.g., *opinião*, *chave*, *exigir*). These can arise because the training data is relatively small.

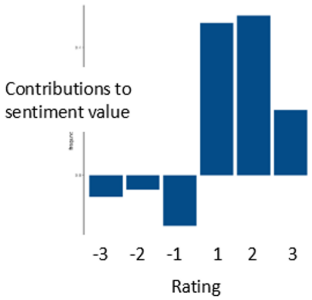

Figure 2: Contributions to the sentiment value of the word "bom".

Ecolex					Sentilex-PT	
Word	**Sent.Val.**	**Count**	**Class**		**Word/Idiom**	**Sent.Val.**
merecer	1.200	6	V		merecer o que tem	-1
excelente (excellent)	**1.167**	8	ADJ		excelente	1
opinião (opinion)	1.143	10	N		opinião	-
chave (key)	1.000	6	ADJ		chave	-
chave (key)	1.000	6	N		chave	-
exigir (require)	1.000	10	V		exigir	-
investimento (investment)	**1.000**	11	N		investimento	-
respeito (respect)	**1.000**	6	N		faltar ao respeito	-1
importante (important)	**0.955**	18	ADJ		importante	-
crescer (grow)	**0.931**	25	V		crescer	-
bom (good)	**0.909**	40	ADJ		bom	1

Table 4
Examples of some lexicon entries in Ecolex and Sentilex-PT.

Class	Count
ADJ	106
ADV	36
N (noun)	246
V (verb)	114

Table 5
Frequencies of lexicon entries for different lexical classes.

Table 4 shows the sentiment values for the same words in Sentilex-PT, an off-the-shelf lexicon [14, 15]. This lexicon includes not only words, but also various idiomatic expressions and phrases (e.g., *faltar ao respeito*). At first sight, this seems to be an advantage. However, our solution based on modifier patterns, discussed in Section 4, can, in principle, generate many phrases including "*faltar ao respeito*", and hence is more general. Sentilex-PT does not include some words that clearly have positive sentiment value in economics and finance that our system introduced (e.g., *investimento*, *crescer*).

3.5. Rescaling the sentiment values for a specific task

The sentiment values calculated as shown are applied to the given sentences to generate predictions. Here, we use the usual method that sums up all the sentiment values of words that appear in the lexicon and in the sentence. However, large texts may include many positive (negative) terms and hence the sum may exceed the limits of -3 and 3. This was confirmed, as the maximum and minimum predictions on the training data were 22.60 and -9.20. To prevent this, we carry out rescaling of the values. This is done by generating the predictions for the training data and identifying the 15 and 85 percentiles, which represent an adjusted minimum and maximum value (the true minimum and maximum could be “outliers", i.e., sentences that are either too short or too long). The 85 and 15 percentiles of predictions on the training data were 4.90 and -1.07.

The lexicon values are stretched (or squashed) so that the adjusted minimum and maximum value would coincide with -3 and +3. All values smaller than -3 are substituted by -3, and, similarly, all values larger than 3 are substituted by 3.

3.6. Combining domain-specific lexicon with an existing general lexicon

As the number of items in the domain-specific training dataset is usually limited, so is the size of such lexicon and the quality of the ratings induced. This is why some authors have maintained a position that the domain-specific lexicon should complement a given general purpose lexicon [6, 7]. In this work, we also consider this alternative and use Ecolex as the domain-specific lexicon and Sentilex-pt, as the general purpose lexicon. The combination of the two lexicons is identified as Ecolex+Sentilex. If the same word is used in both lexicons, preference is given to the sentiment value in the domain-specific lexicon (i.e., Ecolex).

4. Experimental set up, evaluation and results

4.1. Evaluation set up

Evaluation methodology

In this work we adopt a 5-fold cross validation. The existing data is divided into five partitions (folds). The data of one fold (i.e., approximately 408/5 cases rounded up to an integer value) is used as the test data. The remaining four folds are used as “training data” that is used to construct the lexicon (Ecolex). As there are five different ways we can select a fold for testing (there are five folds to choose from), we obtain five different pairs of training and test datasets.

As this training data is unbalanced, i.e., contains more cases with positive rating than the negative ones, the training data is modified before it is used to generate the lexicon. Some positively rated sentences (or their fractions) are duplicated. This way, the training data includes typically 324 cases plus 100 duplicates.

Evaluation measures

Evaluation is done using different types of measures, depending on whether we use our system to generate numeric or categorical predictions. When the system is used as a regression system and hence generates numeric predictions, the *mean absolute error (*MAE*)* is used in the evaluation. It is calculated using $MAE = |\widehat{SV}_{dj} - SV_{dj}|$, where $\widehat{SV}_{dj}$ represents the predicted

sentiment value on text dj and SV_{dj} the true value. When the system is used as a classifier and hence generates categorical predictions, the measures that are commonly used are accuracy, precision, recall, F1 and macroF1 [16]. In this paper we report the results solely in terms of accuracy.

4.2. Evaluation of SA system to generate predictions

Before discussing the results obtained by our system, we describe a simple baseline called default system, which predicts the most frequent class for each case. This value can be compared with the correct value, which in turn allows us to calculate mean accuracy.

The results of our SA system with different sentiment lexicons are shown in Table 6. The columns Fold 1 to Fold 5 show the performance across different folds. The mean value is shown in the last column. As we can see, the accuracy of our system that uses the combined lexicon (Ecolex+Sentilex) is better than the SA system that uses just Sentilex. The improvement is significant - about 10%.

	Accurracy (%)					
Lexicon	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean**
Ecolex	60.37	54.21	65.71	47.06	55.66	56.60
Sentilex-PT	62.26	39.25	65.71	**70.58**	**58.49**	59.60
Ecolex+Sentilex	**76.41**	**77.57**	**73.33**	66.67	55.66	**69.93**

Table 6
Accuracy results with different lexicons.

5. Generation of modifier patterns

5.1. Modifier patterns

As many authors have shown, certain words or phrases can act as *modifiers or contextual valence shifters* [17] of the sentiment value of the words (or expressions) that appear in the same syntactic structure. Typically, it is assumed that these shifters can be adverbs, negative function words, modal verbs or connectors, which modify nouns, verbs, adjectives or adverbs. However, Schulder et al. [18] argue that content words like verbs, nouns and adjectives can act as polarity shifters.

A modifier is a word or expression that changes the polarity of another word or expression. This can be done by the means of the operations of *intensification, downtoning/attenuation or reversal*. In this work, modifiers are represented by symbol "M". The words (or expressions) whose sentiment value is affected by the modifiers are referred to as focal elements and are represented by symbol F.

Whenever we need to indicate that the focal element has positive (or negative) sentiment value, we will use F^+ (or F^-). The combination of the modifier and the corresponding focal element is referred to as a "*modifier pattern*". So, the pattern can be represented by "$F + M$" (or by "$M + F$"), depending on whether the modifier appears before (or after) the focal element. In this expression, the symbol "+" can be seen as a shorthand for "*is followed by*".

The phrase "*melhorar muito*" (*improve greatly*) is an example of the pattern "$F + M$", as F=*melhorar* and M=*muito*. The modifier in this example is an intensifier. Some patterns also include other elements between F and M. In such cases we will use the pattern "$F + X + M$", where X represents one or more words in between F and M. An example of such pattern is "*crescimento parece bom*", where F^+ = "*crescimento*", X = "*parece*", M^I = "*bom*". More examples of patterns are given further on.

5.2. Different types of modifier patterns

The modifiers can perform one of the following operations: intensification, downtoning, and reversal, following the terminology of Trnavac et al. [3]. All three types of operations are discussed below.

Intensification Typically, intensification will increase the magnitude of the sentiment value of the focal element, as the name of the operation suggests. Let us reformulate this using the terminology introduced above: the sentiment value (SV) of the modifier pattern that includes an intensifier will normally be greater (i.e., more positive) than the SV of the focal element if its sentiment value is positive, as in "*melhorar muito*" (*improve greatly*). This can be captured shortly by $SV(F^+ + M^I) > SV(F^+)$, where M^I represents the given intensifier (e.g., "*muito*"). One way of computing the value of SV(F^++M^I) is by introducing a constant C greater than 1 (e.g., 2) and calculating the product $C * SV(F^+)$.

Similarly, if the SV of the focal work is negative, the sentiment value (SV) of this pattern will be further down on the negative scale, as in "piorar muito" (deteriorate greatly). This can be captured shortly by $SV(F^- + M^I) < SV(F^-)$. So $SV(F^- + M^I)$ can be calculated as the product $C * SV(F^-)$.

Both cases above can be unified by introducing symbol "$*$" to represent either "$+$" or "$-$". So, the rule can be expressed as $SV(F^* + M^I) = C * SV(F^*)$.

Attenuation/downtoning This operation works in the opposite way of intensification. Normally, it decreases the absolute value of the magnitude of the sentiment value of the focal element, as in "*melhorar pouco*" (*improve slightly*). One way of computing the value of $SV(F^+ + M^A)$ is by introducing a constant C smaller than 1 (e.g., 0.5) and calculating the product $C * SV(F^+)$, where MA represents the given modifier that is the downtoner. A similar method can be used for focal elements whose value is negative. In this case $SV(F^- + M^A) = C * SV(F^-)$, where C is smaller than 1 (e.g., 0.5).

Reversal/inversion The result of this operation depends on whether the focal element has a positive or negative sentiment polarity, i.e., whether we are dealing with F^+ or F^-. If we are dealing with F^+, reversal inverts the polarity of the sentiment value of the focal element, like, for instance, in "*não é bom*" (*is not good*). In this example, the modifier pattern is $M^R + X + F^+$, where M^R represents the reversal/inversion operation. So, in this case, M^R = "não", X = "é", F^+ = "bom". Here, we use "X" to represent some intermediate words between the modifier and the focal element. We use this pattern also in cases when X is empty. One way of computing the value of this pattern is by inverting the sentiment value of the focal element, that is, by $SV(M^R + X + F^+) = -SV(F^+)$. Most people would agree that the SV of "*não é bom*" (*is not good*) is rather negative.

The use of inversion operation applied to a focal element with negative sentiment value is

more complicated. Consider, for instance, the expression "*não é mau*" (*is not bad*). Most people would not consider this equivalent to "*bom*" (*good*), but rather near a neutral value. This can be represented by $SV(M^R + X + F^-) = 0$.

5.3. Definitions of specific modifier patterns

Many different modifier patterns typically appear in texts. Therefore, our goal was to analyze the given texts and define a set of such patterns considered useful. These patterns were organized in three different tables, depending on the type of modifier involved. Table 7 shows some patterns that involve intensifiers, Table 8 includes downtoners/attenuators, and Table 9 includes reversal/inversion. Also, each table consists of two parts, depending on the order of the F and M elements in the text.

All patterns also show the class of the focal element (i.e., if it is a noun, verb or adjective). So, for instance, F_V (or F_{ADJ}, F_N) is used to represent a focal element that is a verb (or adjective, noun). The class of the modifier is represented in a similar way.

Modifier pattern	**F** (domain-specific)	**X**	**M**
$F_V^+ + X + M_{ADV}^I$	compensar, melhorar, crescer		muito, tanto, largamente
$F_N^* + X + M_{ADJ}^I$	saldo, números de economia		positivo, lisonjeiros
$F_N^* + X + M_{ADJ}^I$	portugueses	são	produtivos
$F_V^+ + X + M_{QUANT}^I$	crescer		mais que V
$F_V^- + X + M_{QUANT}^I$	perder		mais que V
$F_N^+ + X + M_V^I$	crescimento, exportações		suplantou, evoluíram
Modifier pattern	**M**	**X**	**F**
$M_{ADV}^I + X + F_{ADJ}^+$	muito, bastante, mais		interessante, sofisticado
$M_{ADV}^I + X + F_{ADJ}^-$	muito, bastante, mais		negativo
$M_{ADJ}^I + X + F_N^+$	maior		credibilidade
$M_V^I + X + F_N^+$	aumentaram		qualificação
$M_V^I + X + F_N^+$	aumentará	a dimensão de	economia
$M_N^I + X + F_{NP}^+$	o acelerar	de	crescimento económico

Table 7
Examples of modifier patterns that include intensification.

5.4. Deriving the sentiment values of short phrases

Let us now examine how we can use the modifier patterns to derive new sentiment values. This process requires that the sentiment lexicon is used to derive the value of single words (unigrams), as discussed in Section 3. Then all patterns are processed one by one. The sentiment value of a given pattern is derived from the sentiment value of the focal element using a rule

Modifier pattern	**F** (domain-specific)	**X**	**M**
$F_N^+ + X + M_{ADJ}^A$	saldo		negativo
$F_N^- + X + M_{ADJ}^A$	déficit		controlado
$F_N^+ + X + M_{PREP}^A$	crescimento		apenas V
$F_N^- + X + M_{PREP}^A$	desemprego		abaixo de V
Modifier pattern	**F** (domain-specific)	**X**	**M**
$M_{ADV}^A + X + F_V$	pouco	se	alterou
$M_{PREP}^A + X + F_{ADJ}^-$	apesar	de	negativo
$M_N^A + X + F_N^+$	falta	de	qualificação

Table 8
Examples of modifier patterns that include downtoning/attenuation.

Modifier pattern	**M**	**X**	**F**
$M_{ADV}^R + X + F_N^+$	não	tem	solução
$M_{ADV}^R + X + F_{ADJ}^-$	não, nem	é	desajustado, mau, problemático
$M_V^R + X + F_{NP}^-$	inverta		ciclo negativo
Modifier pattern	**M**	**X**	**F**
$F_N^- + X + M_{ADJ}^R$	falta de N^+	parece	disparatado
$F_N^* + X + M_{NP}^R$	F_N	é	falsa questão

Table 9
Examples of modifier patterns that include reversal/inversion.

that has the following form

$$SV(Pattern) \leftarrow C * SV(F) .$$

Let us now focus on one of the patterns to see how this is done. Let us consider, for instance, $F_V^+ + X + M_{ADV}^I$ discussed above. In this case, the rule has the form

$$SV(F_V^+ + X + M_{ADV}^I) \leftarrow C * SV(F_V^+) ,$$

where C is the given constant, whose value (i.e., 2) can be retrieved a table accompanying this pattern. This rule determines the sentiment value of the combination of the focal element F_V^+ and modifier M_{ADV}^I is determined from the sentiment value of the focal element and multiplying it by constant C.

The process of applying a given modifier pattern involves the following steps:

1. Extract the list of modifier words from the pattern definition (e.g., muito, bastante, mais etc.),

2. Identify where these modifiers appear in a given text and check whether they are of the required class (here ADV). Suppose, we have identified the modifier "*muito*" (*very much*) in the text.
3. Determine which way to look for the focal element F. This is determined by the order in which "M" and "F" appear. Considering our example pattern, we should be looking for a focal element at position -1 with respect to the position of the modifier.
4. Test whether the focal element is of the required class (here V). If our text to process included the phrase "*crescer muito*" (*improve greatly*), we would identify the focal element "*crescer*" (*grow*). As indeed, this token is a verb (V), we can proceed.
5. For each focal element found, retrieve its sentiment value from the lexicon and multiply it by the constant C. Suppose, we find that the sentiment value of "*crescer*" is 0.84. So, the sentiment value of "*crescer muito*" is 1.68.
6. The sentiment value obtained is used as the sentiment value of the short phrase covered by the respective pattern. This value is used instead of the original sentiment value of the focal element.

5.5. Results

We have carried out experiments with combination of intensifier patterns and Ecolex using only one-fold (Fold 1 train and test data). No significant improvement was obtained by addition of the modifier patterns.

5.6. Analysis of the current approach

We have analysed both the entries in our lexicon Ecolex and some sentences in the test data whose sentiment value was not predicted correctly. In our view the problems can be attributed mainly to the following two different causes:

- relatively low quality of the induced lexicon,
- patterns are based on a proximity criterion (rather than dependency links).

If some of the lexicon values are wrong, applying modifier patterns will not improve the outcome. For instance, the phrase "*bastante negativo*" (*rather bad*) was attributed an incorrect sentiment value, because the value of F="*negativo*" in the induced lexicon is wrong (0.33). So the application of the intensifier pattern $M_{ADVI} + X + F^{-}_{ADJ}$ magnifies this error. Further on in this section, we discuss two different approaches that can be used to obtain a lexicon that could be relied on.

As we have explained in Section 5.4, the proposed method retrieves a given pattern and then searches through the target text to identify the modifier words. This approach has the following limitation. Let us assume that we encounter the phrase "*é muito bom*". We would expect that the result is positive. Indeed, the sentiment value of F="*bom*" is 0.8, the invocation of the pattern $M^{I}_{ADV} + X + F^{+}_{ADJ}$, where $M =$ "*muito*" gives 1.6. However, in this case another pattern applies, namely $F^{+}_{V} + X + M^{I}_{ADV}$, where $F =$ "*ser*" (*is*) (representing "*é*" in lemmatized form) and $M =$ "muito". This introduces wrongly a sentiment value of "*é muito*" (*is very much*) and this way corrupts the final result. This pattern fires, as it uses a proximity criterion to

identify the focus word. Our future plan involves the usage of dependency links resulting from dependency parsing [19] to identify the focal element. This would prevent the latter pattern from firing.

It could be argued that the word "is" should not be considered altogether because it is a stopword. However, if our aim is to use modifier patterns, care is needed regarding which stopwords are in fact dropped. The words "*não*" (*no*), "*mais*" (*more*), "*muito*" (*many*) are considered stopwords, yet they play an important role in our patterns.

5.7. Some thoughts on how to induce a better quality sentiment lexicon

Consider the issue of how to obtain a sentiment lexicon that has better quality than the one obtained by the current approach. There are two line that can be followed:

- improve the quantity and quality of the training data,
- improve the automatic method of inducing the lexicon.

As the method of inducing the sentiment lexicon can be compared to other ML approaches, the quality of the result depends on the quantity of labelled sentences. As we have mentioned in Section 3.1, the available data consists of 408 labelled sentences (or phrases), which is a relatively modest number.

Many sentences include 15 words or more. In one sentence, for instance, there are 29 tokens and 13 of them are potential sentiment-bearing words (i.e., N, V, ADJ, and ADV). So, the system has a rather hard task of working out which sentiment value to attribute to each of these tokens. It is clear that if the sentence was separated into various parts and labelled, the training data would have a better quality.

The method of inducing the sentiment lexicon could be improved too, by reusing the ideas from the areas of expectation maximization (EM) [20] and Bayesian optimization [21]. The induction would be done in several iterations. The first iteration would follow the approach described in Section 3. Then, the SA system would be applied with the induced lexicon to the training data to calculate the errors. This process would take into account the patterns too. The errors would be used to generate modified labels of the training sentences and then the sentiment values of all the words are recalculated. This process would be repeated several times or until the values stabilize.

Our future work will explore the observations above. Our aim is to demonstrate the usefulness of the approach outlined in this paper.

6. Conclusions

We have described an approach to the construction of a sentiment analysis system, which was developed on the basis of texts on economics and finance in European Portuguese. This system includes a domain-specific sentiment lexicon, modifier patterns and rules.

The acquisition of the domain-specific sentiment lexicon is discussed in Section 3. As we have shown, this is done in an automatic manner based on the distribution of ratings for different words that occur in the training data (labelled texts, i.e., sentences or phrases). We have shown

that this approach can identify some words that have clear positive or negative sentiment in the area of economics and finance in Portuguese. Given that the training data was relatively small (300+ labelled sentences or phrases), one cannot expect the predictive performance to be very high. Still, we have shown that it was better than the default prediction.

Further improvements can be obtained with recourse to modifier patterns discussed in Section 5. These enable us to generate the sentiment values of certain phrases captured by these patterns. All modifier patterns were acquired manually. Each pattern includes a modifier and a focal element. As we have shown, the modifiers can be of different types, depending on whether the operation is intensification, downtoning or reversal. The focal element is restricted to some particular word class (e.g., a noun) that appears in the acquired lexicon. So, this work builds on the work discussed in the previous sections.

This approach combines an automatic process (lexicon construction) with human knowledge in the form of modifier patterns and specific modifiers. This human knowledge can come from specialists in the field of economics, linguistics, among others. This work was our first experiment with this particular conceptualization, representation and coding and represents more a proof of concept, rather than the final solution. We are planning to develop this work further and expect substantial improvements in performance.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020 of INESC TEC and project UIDB/00022/2020 of CLUP.

References

[1] S. Almatarneh, P. Gamallo, Automatic construction of domain-specific sentiment lexicons for polarity classification, in: International Conference on Practical Applications of Agents and Multi-Agent Systems, Springer, 2017, pp. 175–182. URL: https://doi.org/10.1007/978-3-319-61578-3_17.

[2] S. H. Muhammad, P. Brazdil, A. Jorge, Incremental approach for automatic generation of domain-specific sentiment lexicon, in: Advances in Information Retrieval, LNCS, volume 12036, Springer, 2020, pp. 619–623.

[3] R. Trnavac, D. Das, M. Taboada, Discourse relations and evaluation, Corpora 11 (2016) 169–190.

[4] W. van Atteveldt, M. A. van der Velden, M. Boukes, The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms, Communication Methods and Measures 15 (2021) 121–140.

[5] A. C. Forte, P. B. Brazdil, Determining the level of clients' dissatisfaction from their commentaries, in: International Conference on Computational Processing of the Portuguese Language, Springer, 2016, pp. 74–85.

[6] F. Silva, P. Silvano, A. Leal, F. Oliveira, P. Brazdil, J. Cordeiro, D. Oliveira, Análise de senti-

mento em artigos de opinião, Linguística: Revista de estudos linguísticos da Universidade do Porto 13 (2018) 79–114.
[7] A. Moreno-Ortiz, J. Fernández-Cruz, C. P. C. Hernández, Design and evaluation of sentiecon: A fine-grained economic/financial sentiment lexicon from a corpus of business news, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 5065–5072. URL: https://www.aclweb.org/anthology/2020.lrec-1.623.pdf.
[8] A. Muhammad, N. Wiratunga, R. Lothian, R. Glassey, Domain-based lexicon enhancement for sentiment analysis., in: Proceedings of the BCS SGAI Workshop on Social Media Analysis, Citeseer, 2013, pp. 7–18. URL: http://ceur-ws.org/Vol-1110/paper1.pdf.
[9] W. L. Hamilton, K. Clark, J. Leskovec, D. Jurafsky, Inducing domain-specific sentiment lexicons from unlabeled corpora, in: Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing, volume 2016, NIH Public Access, 2016, p. 595.
[10] C. Tavares, R. Ribeiro, F. Batista, Sentiment analysis of portuguese economic news, in: 10th Symposium on Languages, Applications and Technologies (SLATE 2021), Article 17, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021, pp. 17:1–17:13.
[11] J. R. Martin, P. R. White, The language of evaluation. basingstoke and new york, 2005.
[12] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.
[13] B. Liu, Sentiment analysis and opinion mining, Synthesis lectures on human language technologies 5 (2012) 1–167.
[14] M. J. Silva, P. Carvalho, L. Sarmento, Building a sentiment lexicon for social judgement mining, in: International Conference on Computational Processing of the Portuguese Language (PROPOR), Springer, 2012, pp. 218–228.
[15] P. Carvalho, M. J. Silva, Sentilex-pt: Principais características e potencialidades, Oslo Studies in Language 7 (2015) 425–438.
[16] C. J. v. Rijsbergen, Information retrieval (2nd edition), Butterworth-Heinemann, 1979.
[17] L. Polanyi, A. Zaenen, Contextual valence shifters, in: Computing attitude and affect in text: Theory and applications, Springer, 2006, pp. 1–10.
[18] M. Schulder, M. Wiegand, J. Ruppenhofer, B. Roth, Towards bootstrapping a polarity shifter lexicon using linguistic features, in: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017, pp. 624–633.
[19] D. Jurafsky, J. H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3ºedition draft), 2021. URL: https://web.stanford.edu/~jurafsky/slp3/14.pdf.
[20] T. Hastie, R. Tibshirani, J. Friedman, The EM Algorithm, in The Elements of Statistical Learning, Springer series in statistics, Springer, 2001, pp. 272–278.
[21] J. Mockus, V. Tiesis, A. Zilinskas, The application of bayesian methods for seeking the extremum, Towards global optimization 2 (1978) 117–129.