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

**Pavel Brazdil, Purificação Silvano, Fátima Silva,
Shamsuddeen Hassan, Fátima Oliveira,
João Cordeiro, António Leal**

Workshop SALLD-1@ LDK 2021

1 Sept 2021

Research Areas, People & Institutions

AI, ML, Text Mining
Sentiment Analysis

- Pavel Brazdil
- Shamsuddeen Hassan (PhD student)
- João Cordeiro

Linguistics
Sentiment Analysis

- Purificação Silvano
- Fátima Silva
- Fátima Oliveira
- António Leal

Acknowledgements: 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.

Overview

1. Introduction (1)

Our interests:

Sentiment analysis in specific domains and for specific languages
e.g. economics and finance;
texts in European Portuguese.

Approaches that use **(semi-)automatic methods** for
the construction of the **domain-sensitive (DS) sentiment lexicon**

Our aim:

Conceive solutions relatively **easy to explain to the user**
(i.e. why certain polarity/rating is attributed to the given text)

Lexicon-based approaches are of interest
as performance is often not the only important criterion

1. Introduction (2)

Sentiment analysis (SA) is an important area (product reviews, policy reviews, etc.)

The aim is to predict the sentiment value for a given text (to see what the user thinks about a given product or policy)

The sentiment values can be

- Categorical (positive, neutral, negative), or
- Numeric (e.g. on scale -3 to 3).

We prefer this option, as it enables to explore more information

1. Introduction (3)

Many diverse approaches to SA exist:

- **Manual** annotation of texts (by experts, crowdsourcing), (requires manual effort)
- **Lexicon-based** approaches,
- **Machine learning (ML)** approaches,
  - Classical ML approaches,
    (e.g. decision trees, random forests, etc.),
  - Deep neural networks (e.g. CNN).

1. Introduction (4)

Atteweldt et al. (2021) compared various approaches and showed:

DeepNN obtained best performance
although they are behind the manual approach.

However, is not easy to see why certain predictions were made.

Lexicon-based approaches offer better explainability,
but tend to have rather poor performance in specific domains
(e.g., economics and finance, bioinformatics etc.),
when a general-purpose lexicon is used.

Hence, we need a domain-specific (DS) lexicon (or extension).

1. Introduction (5)

Strategies for developing DS lexicon:

- Manual - we exclude this, as this requires a lot of manual effort
- Automatic – followed here

Automatic method for developing DS lexicon for unigrams

Our approach is based on
Almatarneh et al. (2017) and Muhammad et al. (2020).

It requires labelled texts (with numeric ratings).

More details on the method are given in Section 2.

1. Introduction (6)

Lexicon that includes single words only (unigrams) is insufficient to cover all situations (e.g. phrase “is not good”)

=> **We need to be able to deal with (some) short phrases**.

Alternative solutions:

- **Enumerating** various short phrases and storing their sentiment value in the lexicon.

There can be too many!

- **Deriving the sentiment value** of some short phrases using rules (e.g. “*crescimento alto*” (*high growth*))
  This is applicable to short phrases satisfying certain “patterns”

Followed here

- **Hybrid approach** combining the two approaches above

1. Introduction (7)

Short phrases considered:

We focus on certain type of short phrases that involve intensification, downtoning/attenuation, reversal/inversion.

They are represented with the help of ***modifier patterns,*** which are:

- Applicable to various domains;
- Constitute generally useful linguistic knowledge;
- Acquired manually, but **exploit DS lexicon** (acquired automatically);
- The sentiment value of short phrases is obtained in an automatic way.

2. Automatic Generation of DS Sentiment Lexicon

The methodology includes:

1. Corpus preparation and annotation
2. Preprocessing
3. Generating the distributions of occurrence of words
4. Generating the sentiment values
5. Combining domain-specific lexicon with general purpose lexicon

2.1 Corpus preparation and annotation (1)

Four linguists prepared 23 (1+22) texts from articles on **finance and economy** from different Portuguese online newpapers.

Each text contained between 2 to 30 sentences (or their fractions) (the total was 408).

At least two annotators annotated each sentence (or phrase) with the sentiment value on the **scale of -3 to 3**.

Examples:

Doc	No	Frase	SVal
1	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
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
1	3	O **saldo positivo** das nossas trocas **compensa largamente** o financiamento das atividades do país.	2

2.1 Corpus preparation and annotation (2)

Number of occurrences of different ratings in the corpus:

Rating	-3	-2	-1	1	2	3	Total
Nº of occurrences	21	100	147	81	52	7	408

Somewhat unbalanced data:
More negative ratings than positive ones.
Later we show how we deal with this.

A part of this data was used to construct the lexicon;
another part to generate the predictions and evaluate them.
(More details later).

2.2 Preprocessing

The data was read-in with *Quanteda* package of R.

Further processing was done with *udpipe* package:
POS tagging, lemmatization

Not all lexical categories include sentiment bearing words (Martin and White, 2005; Taboada et al., 2011; Liu, 2012, etc.)

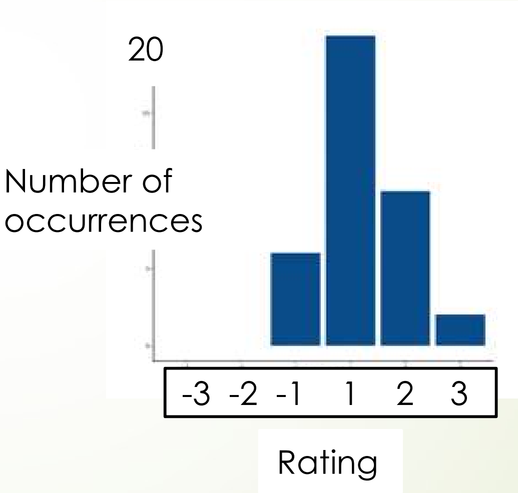
So, we focus on four categories of unigrams only: ***nouns***, ***verbs***, ***adjectives***, ***adverbs***.

2.3 Generating the distributions of occurrences

Consider each word (token) in the labelled sentences in the training data.

Examine the number of occurrences for the given ratings

Ex. word (token) “*bom*” (good):



The occurrences are transformed into probabilities (with Laplace smoothing).

2.4 Generating the Sentiment Value

The sentiment value is derived from
the distribution of probabilities and ratings -3, -2, -1, 1, 2, 3

$$SV_{ti} = \sum P''_{ti} * W$$

Examples of some entries induced:

Ecolex	
Word	**Sent.Val.**
investimento (investment)	**1.000**
respeito (respect)	**1.000**
importante (important)	**0.955**
crescer (grow)	**0.931**
bom (good)	**0.909**

Comparison with one existing sentiment lexicon:

	Sentilex-PT
Word/Idiom	**Sent.Val.**
investimento	-
faltar ao respeito	-1
importante	-
crescer	-
bom	1

2.5 Combining DS+GP Lexicons and Evaluation

Combination of lexicons:

- Use preferentially the DS lexicon (Ecolex)
- Add entries from the GP lexicon (Sentilex-PT)
  (those that do not appear in Ecolex)

2.5 Combining DS+GP Lexicons and Evaluation

Evaluation using 5-fold cross-validation

Repeat 5 times:

- Use 4 partitions (folds) as the “train data” to construct the lexicon.
- Balance the training data by adding some duplicates of positively rated sentences (or their fractions).

  This way we obtained approx. 318 cases + 100 duplicates
- Use 1 partition (fold) to as test data (≈ 81 cases).

3 Experimental Results

The accuracy of default system (random prediction) is 50% for balance data.

	Accuracy (%)					
	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**

The combined lexicon (Ecolex+Sentilex) is better than the constituents (about 10% improvement of accuracy)!

4 Short phrases and modifier patterns

We focus on certain short phrases only that include:

- **intensification** ex. *melhorar muito* (improve greatly)
- **downtoning/attenuation** ex. *melhorar pouco* (improve a little)
- **reversal/inversion** ex. *não é bom* (is not good)

modifier M

focal element F

Modifier and focal element form part of ***modifier patterns***.

These can be transformed into rules and
used to derive the sentiment value of the phrase.

4.1 Intensification (1)

Intensification includes **modifier M** that **increases** the sentiment value of the focal element F

Example of a *modifier pattern*:

$$F_V^{+} + X + M_{ADV}^{I}$$

Focal element must be a **verb (V)** (e.g. *melhorar* (improve))

in-between tokens (ex. 0)

Modifier (e.g. *muito* (very much)) must be an **adverb (ADV)** and acts as **intensifier (I)**

4.1 Intensification (2)

More details

$$F_V^{+} + X + M_{ADV}^{I}$$

No need to specify the word (just its class, e.g. "v")

Admissible adverbs are given.
The list was elaborated manually.
Currently it has 26 elements:

- *muito* (very much),
- *bastante* (rather a lot),
- *mais* (more),
- *poderosamente* (with a great power),
- *muitíssimo* (very, very much),
- *incrivelmente* (incredibly),
- *extraordinariamente* (extraordinarily),
- …

4.1 Intensification (3)

Determining the sentiment value of modifier patterns

$$\mathbf{SV(F_V^{+} + X + M_{ADV}^{I})} = C_{MI} * \mathbf{SV(F_V^{+})}$$

Sentiment value of the modifier pattern

Constant (e.g. 2)

Sentiment value of the focal element (retrieved from the lexicon)

Ex. SV("melhorar", "muito") = 2 * SV("melhorar")

= 2 * 0.6 = 1.2

4.1 Intensification (4)

Modifier pattern	**F** (domain-specific)	**M**
$F_V^+ + X + M_{ADV}^I$	compensar, melhorar, crescer	muito, tanto, largamente, como nunca
$F_N + X + M_{ADJ}^I$	saldo, números de economia	positivo, lisonjeiros
$F_N + X + M_{ADJ}^I$	portugueses	produtivos
$F_V^+ + X + M_{QUANT}^I$	crescer	mais que X
$F_V^- + X + M_{QUANT}^I$	perder	mais que X
$F_N^+ + X + M_V^I$	crescimento, exportações	suplantou, evoluíram
Modifier pattern	**M**	**F**
$M_{ADV}^I + X + F_{ADJ}^+$	muito, bastante, mais	interessante, lisonjeiro, 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	economia
$M_N^I + X + F_{NP}^+$	o acelerar	crescimento económico

4.2 Downtoning/Attenuation (1)

Downtoning/Attenuation Includes **modifier M** that **decreases** the sentiment value of the focal element F

Example of a *modifier pattern*:

$$F_V^{+} + X + M_{ADV}^{A}$$

Focal element must be a **verb** (e.g. *melhorar*)

Modifier (e.g. *pouco*) must be an **adverb** and acts as **attenuator**

Determining the sentiment value:

$$SV(F_V^{+} + X + M_{ADV}^{A}) = C_{MA} * SV(F_V^{+})$$

Constant (e.g. 0.5)

4.2 Downtoning/Attenuation (2)

Our study includes various modifier patterns for downtoning/attenuation, e.g.:

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

4.3 Reversal/Inversion (1)

Reversal/Inversion includes **modifier M** that **inverts** the sentiment value of the focal element F

Example of a *modifier pattern*:

$$F_{ADJ}^{+} + X + M_{ADV}^{R}$$

Focal element must be a **adj** (e.g. *good*)

Modifier (e.g. *not*) must be an **adverb** and acts as **reversal**

Determining the sentiment value:

$$SV(F_{ADJ}^{+} + X + M_{ADV}^{R}) = - C_{MR} * SV(F_{ADJ}^{+})$$

Constant (e.g. 1)

4.3 Reversal/Inversion (2)

Our study includes various modifier patterns for reversal/inversion, e.g.:

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

5 Applying Modifier Patterns and Results (1)

So far, we have conducted experiments with some modifier patterns only.

So far, we have noted a modest improvement.

We plan to complete this study shortly.

We expect that we can gain further 5-10% on accuracy.

6 Conclusions - Note on *Linguistic Linked Data*

The topic is of interests to us, because we are:

- interested in collaborating with others on different SA approaches,
- willing to share/exchange data.

Currently, our data is stored in CSV format, so it can be interchanged.

This applies both to:

- labelled texts with sentiment score,
- modifier patterns (these are transformed into rules by our program).

Transforming this data into other formats could be done.

6 Conclusions – Contributions of this Work

We have shown a method for **automatic construction of a domain-specific (DS) lexicon**
(requires modest number of labelled examples)

It is advantageous to use **labelled data with numeric values**, rather than with just categorical labels.

It is useful to use a **combined DS + GP lexicon,**
i.e., combine the induced DS lexicon
with an existing general-purpose (GP) lexicon.

Experiments show that this is a promising line to follow.
We have obtained about **10% increase in accuracy**
when compared with the existing GP lexicon.

6 Conclusions – Contributions of this Work

We have designed a method (based on previous work) for **representing and applying modifier patterns** (for intensification, downtoning/attenuation, reversal/inversion).

We have shown how these can be combined with the lexicon-based approach.

Experiments are in progress and we expect that we can obtain substantial gains in accuracy (5-10%).

Additional advantage – explainability:
The predictions of sentiment values can be justified to the user (i.e., we can show how they were derived).

References (selection)

Almatarneh, S., Gamallo, P. (2017). Automatic construction of domain-specific sentiment lexicons for polarity classification. In: Advances in Intelligent Systems and Computing, pp. 175–182.

van Atteweld, W., van der Velden, M. A. C. G., Bokes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms, Communication Methods and Measure

Forte, A. C., Brazdil, P. (2016). Determining the Level of Clients' Dissatisfaction from their Commentaries, Proc. PROPOR 2016, Tomar, Springer, Portugal, 74-85

Liu, (2012). Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies. California: Morgan & Claypool Publishers.

Martin, J.R., White, P.R.R. (2005). The Language of Evaluation. New York: Palgrave.

Muhammad, S. H., Brazdil, P., Jorge, A. (2020). Incremental Approach for Automatic Generation of Domain-Specific Sentiment Lexicon. In: Advances in Information Retrieval, LNCS, vol. 12036. Proc. ECIR 2020, 619-623.

References (selection)

Silva, F., Silvano, P., Leal, A., Oliveira, F., Brazdil, P., Cordeiro, C., Oliveira, D. (2018). Análise de sentimento em artigos de opinião, Linguística: Revista de estudos linguísticos da Universidade do Porto, 13, 79-114.

Trnavac, R., Das, D., Taboada, M. (2016). Discourse relations and evaluation, Corpora 11 (2), 169-190.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Association for Computational Linguistics. 37(2): 267-307

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

Pavel Brazdil, Purificação Silvano, Fátima Silva, Shamsuddeen Hassan, Fátima Oliveira, João Cordeiro, António Leal

Workshop SALLD-1@ LDK 2021

1 Sept 2021