An Emotional Word Analyzer for Portuguese*

Maria Inês Maia[1] and José Paulo Leal[2]

1 **CRACS & INESC-Porto LA, Faculty of Sciences, University of Porto, Porto, Portugal**
up201101593@fc.up.pt

2 **CRACS & INESC-Porto LA, Faculty of Sciences, University of Porto, Porto, Portugal**
zp@dcc.fc.up.pt

Abstract

The analysis of sentiments, emotions and opinions in texts is increasingly important in the current digital world. The existing lexicons with emotional annotations for the Portuguese language are oriented to polarities, classifying words as positive, negative or neutral. To identify the emotional load intended by the author it is necessary also to categorize the emotions expressed by individual words.

EmoSpell is an extension of a morphological analyzer with semantic annotations of the emotional value of words. It uses Jspell as the morphological analyzer and a new dictionary with emotional annotations. This dictionary incorporates the lexical base EMOTAIX.PT, which classifies words based on three different levels of emotions – global, specific and intermediate.

This paper describes the generation of the EmoSpell dictionary using three sources, the Jspell Portuguese dictionary and the lexical bases EMOTAIX.PT and SentiLex-PT. Also, this paper details the web application and web service that exploit this dictionary. It presents also a validation of the proposed approach using a corpus of student texts with different emotional loads. The validation compares the analyses provided by EmoSpell with the mentioned emotional lexical bases on the ability to recognize emotional words and extract the dominant emotion from a text.

1998 ACM Subject Classification H.3.1 [Content Analysis and Indexing] Dictionaries

Keywords and phrases Sentiment Analysis, Opinion Mining, Emotion API

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.17

1 Introduction

Sentiment Analysis, also known as Opinion Mining, can be classified as the identification of opinions, emotions and evaluation in texts [22] toward topics, events, entities or individuals. A frequent approach to perform this kind of analysis is to use dictionaries containing words annotated with semantic or polarity orientation. These dictionaries can be created manually or by using seed words, thus expanding automatically the list of words [20].

This paper describes the implementation of EmoSpell[1], an emotional word Analyzer for the Portuguese language. The motivation for EmoSpell comes from the research on the

* This work is financed by BIAL, trough project M-BW, BIAL 312/16, the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020 Programme, by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project POCI-01-0145-FEDER-006961, and by FourEyes, a Research Line within project TEC4Growth – Pervasive Intelligence.

[1] Available at `http://daar.up.pt:8080/EmoSpell/`.

cognitive processes in writing. This kind of research is based on a model that views writing as the interplay of several cognitive processes [14]. Examples of these processes are: planning the text, selecting grammar constructions, checking word spelling. If one of more of these processes take that too much time then writing pauses will occur.

The goal of the research presented in this paper is to integrate EmoSpell with HandSpy[2], a web based collaborative environment for researchers working on cognitive processes in writing [12]. This system, currently in its version 2.0, is being used by 13 research teams from 10 different countries. HandSpy is a tool to detect pauses in the writing productions collected with smartpens and investigate what may have cause them. EmoSpell integrated in HandSpy will enable researchers to relate pauses in writing with the grammatical categories and emotional load of the words where these pauses occurred.

EmoSpell contains a dictionary with semantic annotations regarding the emotional value of words. It is based on a morphological analyzer named Jspell [1] and EMOTAIX.PT [7], a lexical base that catalogs a set of Portuguese words based on their emotional load. SentiLex-PT was used to compare the polarities and extend the emotional classification of words.

Jspell uses a dictionary of lemmas with a set of morphological rules. It avoids the enumeration of all words in a language by creating rules that associate them to their lemmas. Moreover, this dictionary of lemmas can be extended to contain new categories, either syntactic or semantic. Hence, when the lemmas are annotated with their emotional load as semantic categories, their flexed forms also inherit these properties. Also, it is possible to reverse the emotional category of a word, when a negation prefix is added, such as “in”, “des”, “de”. For example, for the word, “feliz” (“happy”) and “infeliz” (unhappy), it will be annotated that this emotional classification will be the opposite.

EMOTAIX.PT is a lexical base with 3992 collected words. These words are classified according to their valence, that is, positive, negative and neutral and semantic nature, that consists in a hierarchical classification of emotional categories. There are other Portuguese lexicons with sentimental annotations such as SentiLex-PT [18]. This one in particular, provides only word polarity, by classifying them as positive, negative or neutral. Therefore, EmoSpell will allow a more detailed analysis of sentiments for the Portuguese language.

EmoSpell will be useful for other purposes than the research in cognitive processes in writing and its planed integration with HandSpy. To reach wide range of potential users, EmoSpell has two kinds of interfaces: a web GUI and an API. The former is intended for end users and was instrumental for the validation of the proposed approach. The latter was designed for the integration with applications, in particular HandSpy. To increase its interoperability, the API build on open XML standards for natural language processing and emotional analysis, namely the *Text Encoding Initiative* (TEI)[3] and *Emotion Markup Language* (EmotionML) [4].

The remainder of this article is organized as follows. Section 2 reviews sentiment analysis APIs that detect emotion on texts. EmoSpell was created based on the Jspell and EMOTAIX.PT, also using the Sentilex-PT. The analyzer and the two lexical bases are detailed in Section 3. The design choices and implementation of the EmoSpell dictionary, as well as the system created to validate this analyzer, are presented in Section 4. Section 5 presents a validation of the proposed approach and the final section summarizes the contribution of this research.

[2] Available at `http://daar.up.pt/index.php/pt/handspy`.
[3] `http://www.tei-c.org/index.xml`

2 State of the Art

Opinion analysis has always been an important topic, either for individuals who want to know the general opinion regarding a certain products of their interest, or for businesses and organizations that want to know their customers' opinions regarding their products and services.[11]

The availability of information from *Microblogging Web-Sites* users turns this into the most famous source of data in this area. Application-oriented research papers referring to *Microblogging Web-Sites*, such as *Twitter*, are based on consumers' opinions regarding products or services, which is valuable information for companies [6]. It is also useful to gauge the political opinion of the population on a certain subject [13].

There are two main approaches to Sentiment Analysis: the classifier-based and the lexicon-based approaches. The former uses machine learning techniques, considering this analysis as a problem of text categorization. The latter uses sentiment lexicons to analyze the text and perceive the sentiments that it contains.

A sentiment lexicon can be constructed either manually or automatically. In the manual creation, there are no algorithms. It is based on the enumeration of a list of words and the annotation of the sentiment classification, using a dictionary or a corpus as resource. This annotation is made by humans so, besides the possibility of human errors, the creation time and lexicon size are also not as expected. The automatic creation of sentiment lexicon uses a set of seed words and rules or methods to expand these words.

Sentiment lexicons are created to analyze the emotional context of texts. They assign an emotional classification to words, so that when they are found in a text they provide information regarding their emotional load. This emotional load can be represented by polarities (positive, negative and neutral), values representing the emotion strength or emotional categories.

Through words, an individual can express his emotions and feelings towards a certain subject. For example, if a sentence contains the words "*like*", "*approved*" and "*good*", we understand that the position of the author towards the matter is positive. On the other hand, if the words are "*hated*", "*disapproved*" and "*bad*", the position is negative.

However, understanding the exact meaning intended by author is a challenge. Words can be used with different meanings, thus conveying different emotions. It is not as simple as to classify some words as positive and others as negative.

When writing a text or a review, it is possible that the writer uses words of an opposite emotional category opposite to the context of the text. When classifying singles words with specific emotional classification, the phrase that incorporates those words can express a different sentiment. The simplest example is the use of a positive word in a negative sentence – "*It was not great*" – despite being a negative opinion, the annotated lexicon translates this as a sentence with a positive word "*great*" [21].

Also, authors are not express their emotions and feeling straightforwardly. For example, the use of sarcasm in sentences is a problem in the emotional analysis of texts. In the sentence "*I Absolutely adore it when the bus is late*", it is used the positive word "*adore*", but sarcastically, to highlight a negative situation [17].

There are lexicons available for several languages. For Portuguese, the already mentioned SentiLex-PT is explained in more detail in the next section, since it was used to create EmoSpell. Examples of these lexicons are the following:

- **Linguistic Inquiry and Word Count (LIWC)** – uses a dictionary of words and word stems, each filed into one or more sub-dictionaries. It classifies words in psychological relevant categories. [15] A similar tool is **EMOTAIX**, but for the French language. [16]

- **SentiWordNet** – it is an extension of **WordNet**. WordNet uses a dictionary containing nouns, adjectives, verbs and adverbs that can be called *"synsets"*. [9] SentiWordNet added three sentiment values to each "*synset*", that is, a "*synset*" will have a positive, negative and neutral score. [8]
- **Sentiment Orientation CALculator (SO-CAL)** – is a system that contains a dictionary of annotated words with semantic orientation. SO-CAL have two assumptions – the first is that the individual words have a "prior polarity", which is a semantic orientation independent of context. The second is that this orientation can be identified with a numerical value, the strength. [20]

The lexicons are usually made available as *Sentiment Analysis API*, frequently in complement to other natural language processing features. Examples of these APIs are the following.

- **TweetSentiments** – returns the sentiment of Tweets using the supervised learning algorithm Support Vector Machine (SVM). It has two online APIs that analyze Tweets from Twitter API calls, returned by a Twitter search query[4].
- **ML Analizer** – provides several text analyzes, including feelings, text classification, language detection, locations extractor, adult content analyzer and article summarization[5].
- **WebKnox Text-Processing** – natural language processing of texts such as determining the feeling, identification of the language, the classification of the quality of writing, the auto-correction of a text, the extraction of data and locations and the tagging of a text with part-of-speech tags[6].
- **Skyttle** – provides services to extract patterns from text such as sentiment terms, constituent terms (meaningful expressions) and entities such as names of people, place and things. Supported languages are English, French, German and Russian[7].
- **Sentiment Analysis Spanish** – Sentiment analysis for tweets in Spanish[8].
- **nlpTools** – text classification and sentiment analysis for Natural language. It is an API focused on online news media[9].
- **Yactraq Speech2Topics** – converts audiovisual content into topic metadata. This conversion is done through speech recognition and natural language processing[10].

3 Background

EmoSpell is based on the extension of a dictionary of the morphological analyzer Jspell using the lexical base of emotional words, EMOTAIX.PT. Jspell has features that improve the efficiency of classifying words by providing a set of rules to generate words from radicals. The development of EmoSpell used also a Sentiment Lexicon called SentiLex-PT, mostly to compare the polarities of EMOTAIX.PT and to identify relevant missing words in EMOTAIX.PT. This section details these three different tools used in the development of EmoSpell.

---

[4] Available from `https://www.programmableweb.com/api/tweetsentiments`.
[5] Available from `https://market.mashape.com/mlanalyzer/ml-analyzer`.
[6] Available from `http://webknox.com/`.
[7] Available from `http://www.skyttle.com/`.
[8] Available from `https://market.mashape.com/molinodeideas/sentiment-analysis-spanish#!documentation`.
[9] Available from `http://nlptools.atrilla.net/web/`.
[10] Available from `http://yactraq.com/`.

3.1 Jspell

The morphological analyzer Jspell is an extension of the Ispell spell checker. Although not itself a morphological analyzer, Ispell already includes the possibility of definition and usage of elementary morphological rules [1].

Since natural language applications needed to be able to handle grammatical and semantic information of words, it is important to have a lexical classifier able to provide information on a given word. This information can be based in its origin and grammatical category. The lexical is fundamental in the parsing of these types of applications.

For that purpose, Jspell uses a dictionary, which is a list of words that are classified based on a set of formation rules. These rules functionally bring an important advantage for analyzers, which simplifies the exhaustive enumeration of all the dictionary words, thus creating a list with the lemmas of words and morphological rules. Consequently, this will associate flags to each word in the list that contains the rules that may be applied to the word [1].

The dictionary structure consists basically of entries, each one containing [19]:

- a **lemma**, which is a word from where you can get other words, by derivation or inflection and that cannot be obtained by any other lemma;
- **morphological description**, that is, a list of morphological properties that are key-value pairs of grammatical classification of lemmas that may contain macros to simplify, which will be explained below;
- **derivation rules** which is a set of identifiers of derivation or inflection rules (flags) that are defined in a separate file called affix rules.

Therefore, a typical entry in the dictionary is a line of the form

word/classification/flags[/comment]

The Portuguese dictionary of Jspell contains about 400 000 entries and the rules associated to it. Since dictionaries are in text format they can be easily modified. Thus, it is fairly simple to expand and to create new dictionaries with this analyzer.

To conclude, Jspell can be used in ways, ranging from an interactive web application with menus and options, to a library, It includes also a line interpreter where the user can write a word and receive the corresponding information. This interpreter can be used other programs interacting with Jspell trough pipes. Besides that, Jspell can be used as a standard library (dll/so/dylib) which can be an advantage in efficiency manners.

3.2 SentiLex-PT

SentiLex-PT, as the name suggests, is a sentiment lexicon which was designed to analyze the sentiment and opinion about human entities in texts written in Portuguese[11]. It contains 6.321 adjectival lemmas and 25.406 inflected forms. The lexicon entries correspond to human predicates – adjectives, nouns, verbs and idiomatic expressions. In a sentence, to classify a word based on its polarity, the target of sentiment is verified in order to identify whether it has a subject or complement function. For example, in the case of the word “fat”, as a modifier of a name of human nature, (e.g. “fat guy”), it has a negative polarity, but it has the opposite polarity when combined with a name such as salary (e.g. “fat salary”). [18]

[11] Available from `http://dmir.inesc-id.pt/project/SentiLex-PT_02`.

SentiLex-PT includes two dictionaries: *SentiLex-lem*, which describes the lemmas and the *SentiLex-flex* which is the dictionary that corresponds to the inflected forms. In the Sentilex-lem [5], each line includes:

- **Lemma;**
- **Grammar Category** (adjective, noun, verb or idiom);
- **Sentiment Attributes** (polarity and target of polarity which corresponds to a human subject, being N0 the subject and N1 the complement).

For instance, one entry of SentiLex-Lem is:

```
enganar.PoS=V;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN
```

In *SentiLex-flex*, the entries are associated to the lemma and in addition to the information contained in the *SentiLex-lem*, in this format it also describes information about inflection like gender and number.

Although it was an important step in the sentiment analysis of Portuguese texts, the SentiLex-PT classifies each word through polarity – the word can be negative, positive or neutral. So, as will be detailed in the description of EMOTAIX.PT, the last dictionary will allow a much more detailed sentiment analysis, because each word will not only be classified based on three polarities but by categories of sentiment as well.

3.3 EMOTAIX.PT

Tools have been created in order to classify words according to their emotional load and to automate the vocabulary analysis process used in the writing of texts.

In 2001, Pennebaker, Francis and Booth developed the *Linguistic Inquiry and Word Count* (LIWC) [15], that divides English words in psychological relevant categories. In 2009, *Piolat and Bannour* created a similar tool – EMOTAIX [16] – for the French lexicon, which were explained in more detail in the previous section. Due to the relevance of these tools and the absence of a lexical database with emotional words in European Portuguese, Sara Costa created an EMOTAIX adaptation for this lexicon/language – the EMOTAIX.PT. [7]

EMOTAIX.PT is based on a database of 3992 words, classified according to their valence (positive/negative) and semantic nature. Therefore, there is a main division of words in positive and negative. Negative emotions are divided in three broad categories: "*Malevolência*" (Malevolence), "*Mal-estar*" (Malaise) and "*Ansiedade*" (Anxiety). These categories are further divided in basic categories. In addition to categories with positive and negative valence, EMOTAIX.PT is still constituted by three additional categories of free valence: *Surprise*, *Indifference* and *Non-Specified*. For a better understanding, the figure 1 below represents graphically the different levels of organization:

With this, the EMOTAIX.PT, consists in 2*25 basic categories (center) organized in three hierarchical levels, on each side of a hedonic axis (positive and negative valence), that is, for a given category, if we draw a diagonal line, we can obtain the opposite category at the end of that same line.

4 Design and Implementation

The core of EmoSpell is an extension of the Jspell dictionary with the emotional annotations, accessible via two interfaces for word and text analysis. This section explains the dictionary generation procedure, presents the design of the generator, illustrates the kind of analysis provided by EmoSpell, and describes the main features of its user and application interfaces.

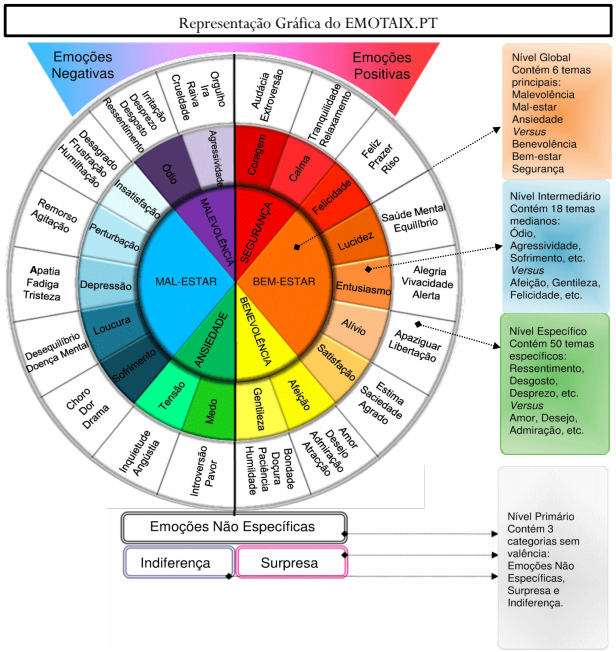

Figure 1 EMOTAIX.PT – Levels of organization (source: [7], in Portuguese).

4.1 Dictionary generation procedure

The EmoSpell dictionary was developed using a Java program that imports three different dictionaries – Jspell, EMOTAIX.PT and SentiLex-PT – and stores them in memory using a common format. These sources are processed to generate the EmoSpell dictionary.

Since the EmoSpell dictionary is created from multiple sources, inconsistencies should be expected. The words that occur in multiples sources where verified for common features. In particular, EMOTAIX.PT and SentiLex-PT were checked against each other looking for inversions in polarity. Also, words in SentiLex-PT missing in EMOTAIX.PT are reported for future inclusion in this lexical base.

The EmoSpell dictionary extends the Jspell dictionary with a new type of classification of words for emotional value. This classification comes from the EMOTAIX.PT that classifies a word in three emotional category levels: global, intermediate and specific.

To add this emotional classification,there were created new 52 macros in the new Jspell dictionary, representing all the possible emotional classifications. A macro entry example is: `"#E26/EmoGlobal=bem estar,EmoIntermediate=entusiasmo,EmoSpecific=alegria`" which corresponds to the classification "*well being, enthusiasm, joy*".

This Java program can be represented by the UML activity diagram of the figure 2.

All EMOTAIX.PT words were compared with Jspell words, both the lemmas and the flexed words. If an emotional word from EMOTAIX.PT is a lemma in the Jspell dictionary then it is only necessary to add the emotional classification for this entry. Otherwise, there are several cases to consider.

If the word being analyzed is not a lemma, this means that either the word does not exist for Jspell or it is flexed from a lemma. In the latter, the generated words inherit

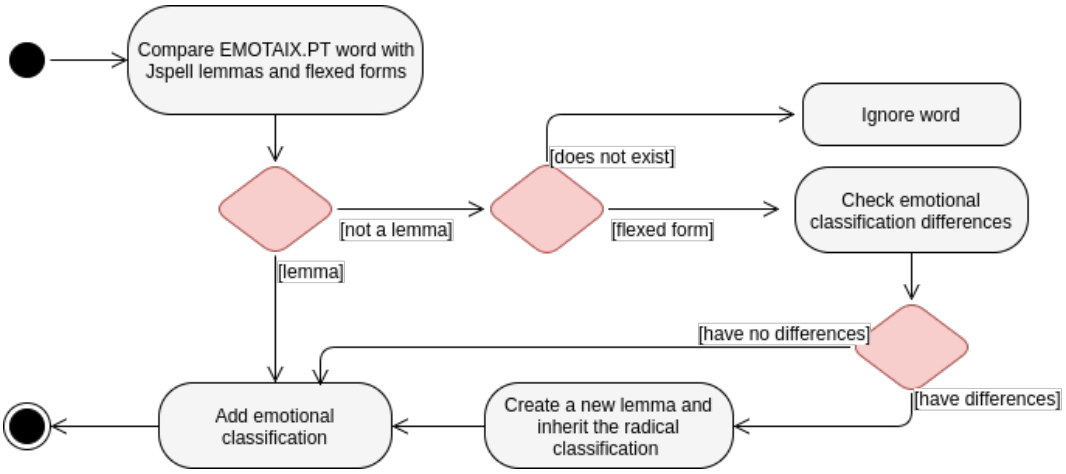

Figure 2 Dictionary Generation – UML Activity Diagram.

the classifications from their radicals. However, if the emotional classification from the EMOTAIX.PT is different between the radical and the flexed forms, then these differences have to be added. A new lemma is created in the Jspell dictionary with the EMOTAIX.PT classification. For example, in the case of the radical "*terror*" and the flexed form "*terrorismo*" ("terrorism"), it was necessary to create a new lemma for the generated word. The emotional classification of the radical is "dread", but for the generated word "*terrorism*" is "cruelty".

For these flexed forms, instead of creating a new lemma, it would be possible to change the Jspell affix file, adding a new classification for the flexed forms. This would be the best approach if we only classified words as polarities. Since we have various classifications, this option would depend on a post-processing. So, when creating this new flexed form entry it is necessary to remove the flag from the copied lemma entry due to the various forms that it generates.

A particular case is when the emotional classification of the radical is the opposite of the generated word. For instance "*feliz*" ("*happy*") and "*infeliz*" ("*unhappy*"). In Jspell, "*feliz*" is the lemma, an entry in the dictionary, and "*infeliz*" the generated word. This is possible due to the existence of a rule that associates all the prefixes that can be added to a word to create their opposite. Being these two words associated, their classifications are the same, so, if we had the emotional classification to "*happy*", "*unhappy*" would also inherit this emotional value which is not true. This problem was solved by adding a new lemma for the opposite word in the dictionary. The new lemma has the categories of its opposite in the original Jspell dictionary and the emotional categories for that word in EMOTAIX.PT.

Another particular case are the irregular verbs. The Jspell dictionary contains entries with the each verbal forms since, by definition, it is not possible to flex them from the radical. If the word from EMOTAIX.PT is a verbal form of an irregular verb, it is necessary to also add to the other forms the emotional classification.

Given that the file to be used in the Jspell is a dictionary of lemmas, after building the new analyzer with the generated file and the affix file of morphological rules, the flexed forms of the annotated words also inherit the emotional classification.

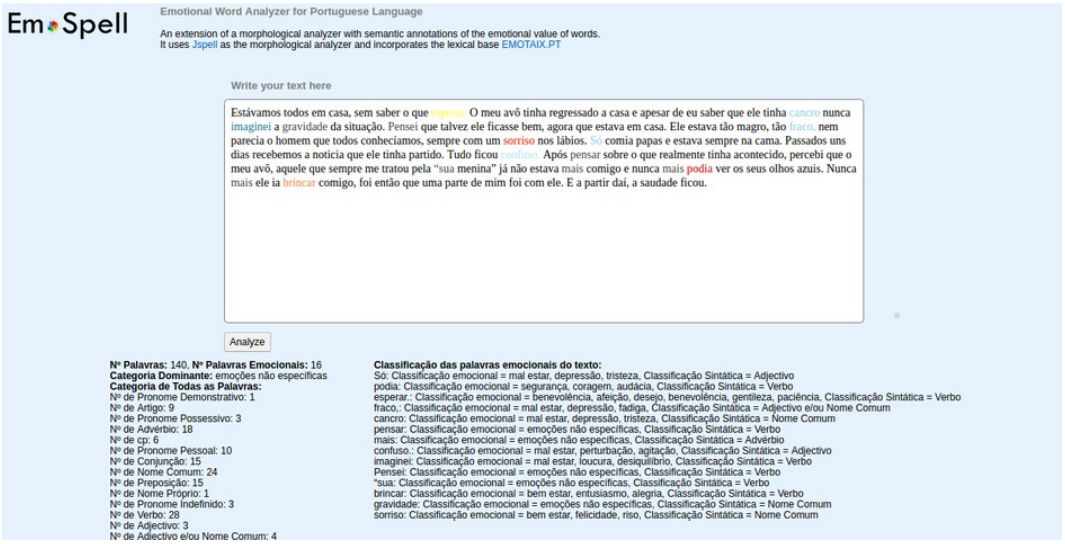

Figure 3 GUI Screenshot – Analyzed text with classification information.

4.2 Text Analysis

The extended dictionary enables EmoSpell to analyze words and texts both with syntactic and emotional categories. This subsection presents an example of the output of analyzer with the extended dictionary, and explains the morphological and emotional description.

When a word is analyzed by EmoSpell, the syntactic information from Jspell is mixed with emotional classification from EMOTAIX.PT. Herewith, an example of this word analysis for the "*medo*" (fear) word, using the command line interface of Jspell.

```
medo

* medo 0 :lex(medo [CAT=nc,G=m,N=s,EmoGlobal=ansiedade,
EmoIntermediate=medo,EmoSpecific=pavor], [], [], [])
```

The first line echoes the user's input, the word to be analyzed. The following are EmoSpell's classification of the word. The first three classifications ("`CAT=nc,G=m,N=s`") are based on the morphological description of the word and, in this case, "`medo`" is a common name, male gender and singular name. The following are the emotional categories levels, "`EmoGlobal=ansiedade`" correspond to the global emotional category (anxiety), "`EmoIntermediate=medo`" is the intermediate level (fear) and "`EmoSpecific=pavor`" to the most specific level is "`pavor`" (dread).

4.3 Interfaces

Although accessible from the command line as a Jspell dictionary, as shown in the previous subsection, EmoSpell has two interfaces for word and text analysis: a GUI (*Graphical User Interface*) and an API (*Application Programming Interface*). The GUI offers direct interaction with end user via a web browser, and the API enables the integration with remote applications using web services.

The EmoSpell Web Application was developed with GWT (*Google Web Toolkit*). Client server communication relies on the GWT RPC (*Remote Procedure Calls*) framework. Figure 3 is a screen shot of the web application. It contains a rich text area where users can write the

text they want to have analyzed. The submitted text is classified as a "bag of words" doing the linguistic and emotional analysis of each word. Also, the text is annotated in place with the result of words by word analysis, followed by a summary of the text analysis.

In the analyzed text, emotional words are formatted with the colors of the categories assigned by EMOTAIX.PT, shown in Figure 1. This provides a simple and immediate understanding of the text emotional load.

The panel below the text displays the general information such as number of emotional words, the dominant category, the morphological and emotional classification of the emotional words.

The API exposes as a web service the functions of EmoSpell's server that are invoked by RPC from the web client. The API follows a RESTfull architectural model, with a single function for analyzing texts. The request is implemented as an HTTP POST method that receives the text an HTTP parameter. The response is a XML document with EmoSpell's analysis.

The document type of the response builds on EmotionML (*Emotion Markup Language*) [4] and TEI (*Text Encoding Initiative*) [10]. EmotionML is a W3C recommendation to represent emotions [3]. This markup language consists of a root document, with `<emotionml>` annotation that contain one or more `<emotion>` elements that represent the emotional classification and can have more elements like category, action-tendency and dimension. EmotionML fragments can also be embed in documents of other languages. TEI is a much more complex XML norm than EmotionML and only a small part of it is actually used by EmoSpell, namely feature structures (`<fs>`), features (`<s>`), segments (`<seg>`) and choices (`<choice>`), which are enough to represent the syntactic categories of words. These TEI elements can also be mixed with elements from other types.

The XML Schema definition (XSD) of EmoSpell's API responses combines elements from the EmotionML and TEI. It defines a minimal structure to bind elements from the imported types: a sequence of `<word>` elements and a `<summary>`. The former combines elements of TEI with the `<emotion>` element from EmotionML. The summary is an element of summarization to represent the analysis information of words.

In summary, EmoSpell provides the GUI interface for end users, making available the text and words analysis for any user, as well as the API interface, thus providing a service to other systems to implement and communicate with EmoSpell.

5 Validation

The validation of the proposed approach compared the results obtained on the analyzes of the same corpus with EmoSpell, EMOTAIX.PT and SentiLex-PT. The corpus used for validation consists of texts written by university students. Each student was asked to write 3 texts describing: a traumatic moment, a happy experience and their daily routine.

Table 1 compares the results obtained with EMOTAIX.PT and EmoSpell for the same texts. In this table, the columns with headers labeled **EX** and **ES** refer respectively to EMOTAIX.PT and EmoSpell results. It lists results of texts analysis from three participants; one positive, one negative and one neutral for each participant. For each text, it lists the number of emotional words given by EMOTAIX.PT and EmoSpell, distinguishing the positive, negative and neutral words. As it can be seen, in all the texts, EmoSpell can detect a larger number of emotional words.

For example, if we observe the positive text from the third participant, the line 3-Pos, we can verify that EMOTAIX.PT detects 10 emotional words and EmoSpell 17 words. Also, it

Table 1 Comparison between EMOTAIX.PT (EX) and EmoSpell (ES).

Participant – Texts	Words											
	Emotional			Positive			Negative			Neutral		
	EX	**ES**	**Δ%**	**EX**	**ES**	**Δ%**	**EX**	**ES**	**Δ%**	**EX**	**ES**	**Δ%**
1-Pos	10	15	50	4	6	50	1	2	100	5	7	40
1-Neg	11	18	60	1	5	400	6	8	30	4	5	25
1-Neut	2	3	50	2	3	50	0	0	0	0	0	0
2-Pos	10	14	40	4	5	25	4	5	25	2	4	100
2-Neg	11	16	40	2	4	100	3	5	60	6	7	20
2-Neut	3	9	200	1	2	100	0	5	0	2	2	0
3-Pos	10	17	70	5	12	140	0	0	0	5	5	0
3-Neg	12	18	50	4	7	75	7	7	0	1	4	300
3-Neut	1	3	200	0	1	∞	1	2	100	0	0	0
Average	7.8	12.6	84	2.6	5	104	2.4	3.8	35	2.8	3.8	54

detects that this is a positive text taking into account the large number of positive words and the absence of negative words.

One of the aims of this experiment was to verify if it is possible to categorize between positive, neutral and negative texts. From the results of 81 written texts obtained by the university students, it was verified that they write more emotional words in the negative and positive condition than if in the neutral condition. Also, it was observed that there were more positive words in the positive texts, and more negative words in the negative texts, as expected. In Table 1, it is also possible to observe that ability to synthesize the polarity of texts. Also, in general, all participants used at least one emotional word of the opposite category.

The existence of a large number of words of contrasting polarity, in particular in negative texts, has several explanations. One explanation for the high number positive word in the negative and neutral texts is the way EMOTAIX.PT categorizes words. Words such as "boyfriend" / "girlfriend" and "friend" are defined in EMOTAIX.PT as positive. When participants write about an experience they typically use these words to refer to the persons involved, without a emotional connotation.

Another explanations is the fact that the participants use contrasting words to intensify the emotional experience. One example is "*we were leaving a birthday party at 5am, we were all very happy and cheerful because the party had been incredible. My friend got in the car ... a few meters ahead the car capsized, they were moments of panic and deep fear*".

As mentioned before, EmoSpell was also validated against SentiLex-PT. Table 2 compares the results obtained with SentiLex-PT and EmoSpell for the same texts. In this table, the columns with headers labeled **S** and **ES** refer respectively to SentiLex-PT and EmoSpell.

The emotional words detection was based not just on the SentiLex-PT lemmas but also on their flexed forms, with the SentiLex-flex file. This increases emotional word detection, just as Jspell does for EmoSpell.

In contrast with the EmoSpell and EMOTAIX.PT comparison, SentiLex-PT can verify a large number of emotional words and in some cases SentiLex-PT detects more emotional words than EmoSpell. This was expected since a number of emotional words in SentiLex-PT were found to be missing in EMOTAIX.PT and reported as part of the generation process, as explained in subsection 4.1. These words are currently being categorized by the authors of EMOTAIX.PT and will be available on the next version of EmoSpell.

Table 2 Comparison between SentiLex-PT (EX) and EmoSpell (ES).

Participant – Texts	Words											
	Emotional			Positive			Negative			Neutral		
	S	ES	Δ%	S	ES	Δ%	S	ES	Δ%	S	ES	Δ%
1-Pos	13	15	15.4	5	6	20	5	2	−60	3	7	133
1-Neg	13	18	38.5	4	5	25	6	8	33.3	3	5	66.6
1-Neut	8	3	−62.5	5	3	−40	1	0	−100	2	0	−100
2-Pos	12	14	16.6	3	5	66.6	5	5	0	4	4	0
2-Neg	9	16	77.7	0	4	0	6	5	−16.6	3	7	133.3
2-Neut	7	9	28.6	2	2	0	5	5	0	0	2	0
3-Pos	15	17	13.3	9	12	33.3	3	0	−100	3	5	66.6
3-Neg	15	18	20	7	7	0	6	7	16.6	2	4	100
3-Neut	7	3	−57.1	1	1	0	2	2	0	4	0	−100
Average	11	12.6	21.16	4	5	11.6	4.3	3.8	−25.19	2.6	3.8	33.28

In any event, this comparison shows that EmoSpell detects, on average, as many emotional words as the SentiLex-PT. However, the main advantage of EmoSpell is that it provides more information on emotional words. Besides presenting a hierarchy of emotional classification, EmoSpell also detect the dominant emotional category of each text.

As part of the validation, the top emotional categories of the texts shown in Tables 1 and 2 were also synthesized. For example, the dominant categories of the three negative is "*benevolence – affection – love*", and "*non-specific emotions*". The three positives texts have the same dominant categories as the negative ones. This can be explained by the fact that the participants used several times words such as "friend", "girlfriend/boyfriend", "to feel/feeling" that are words of "*benevolence*" category. For the "*non-specific words*" it was also noticed that the participants usually write words to intensify their feelings like "think", "more", "larger", "great" and "hard", as already explained.

6 Conclusion

Emotional analyzers have several applications, in research, business and governance. For instance, the opinions expressed on social media regarding a particular product or subject provide important information for companies, organizations and government. Nevertheless, the motivation for this research work came from the application of sentiment analysis to the research on cognitive processes in writing.

Sentiment analysis requires an analyzer capable of detecting a wide range of emotional words in texts, classifying their emotional value and synthesizing the writers' emotional state. EmoSpell improves sentiment analysis for the Portuguese language by classifying words according to emotional categories, not just discovering their polarity or valence. To achieve it, EmoSpell uses EMOTAIX.PT, a lexical base structured in several hierarchical levels of emotional value.

The proposed system build on the lexical analyzer Jspell to enhance the recognition power of EMOTAIX.PT. The main contribution of this work is a procedure for generation of a new Jspell dictionary integrating EMOTAIX.PT. A secondary contribution is the two interfaces to this dictionary, an interactive web application and a text analysis web service. The former was used to validate the proposed approach and is available for experiments in emotional text analysis. The latter will be integrated in the project HandSpy – a web environment

for managing experiments on cognitive processes in writing – and is also available to other systems requiring a syntactic and emotional analyzer of Portuguese texts.

As future work, this emotional analyzer can be improved to better support sentiment analysis in Portuguese. As previously mentioned, there are challenges to overcome such as multiple word analysis and the differentiation of the polarity.

Multiple word analysis would be an important feature to this project. The problem with the use of positive words in a negative context, such as "*I don't like*", could be solved with this addition. Some lexicons already overcame this challenge [2]. One approach is to analyze the phrases not only word by word but in a multi-level way, calculating the sentence polarity by verifying the noun and verb in phrases and identifying their polarities.

As SentiLex-PT, the differentiation of the polarity target can accomplish a more accurate analysis. In EmoSpell, the division of the polarity target into subject and complement would allow a word analysis with different meaning words. That is, it would solve the problem that a word can have opposite polarity when combined with another word.

Acknowledgements. The authors wish to thank to Alberto Simões and José João Almeida, the authors of Jspell, for their help.

References

1 José João Almeida and Ulisses Pinto. Jspell – um módulo para análise léxica genérica de linguagem natural. In *X Encontro da Associaçao Portuguesa de Linguistica*, pages 1–15, 1994.

2 Amna Asmi and Tanko Ishaya. Negation identification and calculation in sentiment analysis. In *Second International Conference on Advances in Information Mining and Management*, pages 1–7, 2012.

3 Felix Burkhardt, Catherine Pelachaud, Björn W. Schuller, and Enrico Zovato. EmotionML. In Deborah A. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 65–80. Springer, 2017.

4 Felix Burkhardt and Marc Schröder. Emotion markup language (EmotionML) 1.0. W3c recommendation, World Wide Web Consortium, 2014.

5 Paula Carvalho and Mário J. Silva. SentiLex-PT: Principais características e potencialidades. *OSLa, Oslo Studies in Language*, 7(1):425–438, 2015.

6 Wilas Chamlertwat, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, and Choochart Haruechaiyasak. Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, 18(8):973–992, 2012.

7 Sara Filipa Oliveira Costa. Adaptação e teste de uma base lexical de palavras emocionais para o português europeu: (EMOTAIX.PT). Master's thesis, Universidade do Porto, 2012.

8 Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: a high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26, 2007.

9 Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

10 Nancy Ide and Jean Véronis. *Text encoding initiative: Background and contexts*, volume 29. Springer Science & Business Media, 1995.

11 Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Llanguage Technologies*, 5(1):1–167, 2012.

12 Carlos Monteiro and José Paulo Leal. Managing experiments on cognitive processes in writing with HandSpy. *Computer Science and Information Systems*, 10(4):1747–1773, 2013.

13 Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Web and Social Media*, volume 11, pages 122–129, 2010.

14 Thierry Olive, Rui Alexandre Alves, and São Luís Castro. Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, 21(5):758–785, 2009.

15 James W. Pennebaker, Martha E. Francis, and Roger J. Booth. *Linguistic inquiry and word count: LIWC2001*, 2001.

16 Annie Piolat and Rachid Bannour. EMOTAIX: un scénario de tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année psychologique*, 109:655–698, 2009.

17 Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 704–714, 2013.

18 Mário J. Silva, Paula Carvalho, and Luís Sarmento. Building a sentiment lexicon for social judgement mining. In *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 218–228, 2012.

19 Alberto Simões and José João Almeida. jspell.pm: um módulo de análise morfológica para uso em processamento de linguagem natural. In *Associação Portuguesa de Linguística*, pages 485–495, 2001.

20 Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

21 G. Vinodhini and R. M. Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292, 2012.

22 Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.